Journal of Machine Learning Research 24 (2023) 1-39 Submitted 5/23; Revised 10/23; Published 12/23

Set-valued Classification with Out-of-distribution Detection for Many Classes

Zhou Wang WANGZH@MATH.BINGHAMTON.EDU
Department of Mathematics and Statistics
Binghamton University, State University of New York
Binghamton, NY 13902, USA

Xingye Qiao XQIAO@BINGHAMTON.EDU
Department of Mathematics and Statistics
Binghamton University, State University of New York
Binghamton, NY 13902, USA

Editor: Xiaotong Shen

Abstract

Set-valued classification, a new classification paradigm that aims to identify all the plausible classes that an observation belongs to, improves over the traditional classification paradigms in multiple aspects. Existing set-valued classification methods do not consider the possibility that the test set may contain out-of-distribution data, that is, the emergence of a new class that never appeared in the training data. Moreover, they are computationally expensive when the number of classes is large. We propose a Generalized Prediction Set (GPS) approach to set-valued classification while considering the possibility of a new class in the test data. The proposed classifier uses kernel learning and empirical risk minimization to encourage a small expected size of the prediction set while guaranteeing that the class-specific accuracy is at least some value specified by the user. For high-dimensional data, further improvement is obtained through kernel feature selection. Unlike previous methods, the proposed method achieves a good balance between accuracy, efficiency, and out-of-distribution detection rate. Moreover, our method can be applied in parallel to all the classes to alleviate the computational burden. Both theoretical analysis and numerical experiments are conducted to illustrate the effectiveness of the proposed method.

Keywords: set-valued classification, out-of-distribution, kernel learning, empirical risk minimization, statistical learning theory

1. Introduction

The traditional multicategory classification paradigms have limitations in several aspects: 1) They return a single class label as the prediction for each data point without a confidence measure attached. This means that the user has no idea of the level of correctness for the decision made. 2) They are forced to always return a class prediction for any point, even when the chance of mistake is high, e.g., for data points near the classification boundary. In some high-stake fields like medicine, the military, or autonomous vehicles, these incorrect decisions caused by high uncertainty can lead to severe and irreversible consequences. In such cases, it may be preferable to make a "partial" prediction (to be defined later) or even abstain from making a prediction until there is more clarity or even human intervention. 3)

Standard single-valued classifiers are obtained by minimizing the overall misclassification rate, ignoring the varying importance of each class. This is problematic in such applications as medical diagnosis and triage, where the cost of misclassifying someone who needs immediate care into a non-urgent group is clearly much higher than the cost of misclassifying someone who does not require immediate attention to the urgent class. Additionally, this approach can be susceptible to imbalanced data, where the minor class may be misclassified as the major class without significantly deteriorating overall accuracy. 4) The conventional classification assumes that the future data points have the same distribution as the training data, which may not be practical in the open world. If a new class emerges, it becomes necessary to detect those out-of-distribution (OOD) or anomaly points. In public health, for example, confidently identifying the strain of a prevailing virus in a community is essential, but it is also important to detect new strains, such as a new COVID-19 variant.

To reach a certain confidence guarantee and reduce the risk of incorrect predictions, it is hence desirable to extend single-valued prediction to assigning multiple possible labels for an observation. Conformal Prediction (CP) (Vovk et al., 2005; Shafer and Vovk, 2008; Balasubramanian et al., 2014) is an increasingly popular framework that outputs a prediction set with a pre-specified confidence guarantee. Classification with Confidence (Lei, 2014; Wang and Qiao, 2018; Sadinle et al., 2019), another approach, considered the set-valued classification from an optimization perspective. In particular, the goal is to minimize the expected size of the prediction set while controlling class-specific error rates. However, these set-valued classification methods are not designed with the capacity of OOD detection.

As for whether to make predictions for those observations with high uncertainty, Herbei and Wegkamp (2006); Bartlett and Wegkamp (2008); Ramaswamy et al. (2015) and Charoenphakdee et al. (2021) proposed and developed Classification with a Reject Option (CRO) by training a classifier and a rejector at the same time. A rejector determines when to refuse to make a classification for those difficult points, by assigning all labels to those observations (this is coined "ambiguity rejection"). For multicategory classification with K classes, Zhang et al. (2018) proposed Classification with Reject and Refine Option, which allows less difficult, but still uncertain, observations to be assigned k labels ($1 < k < K$). Although CRO is more prudent than single-valued classification due to the reject or/and refine options, it lacks an explicit confidence guarantee and is not capable of OOD detection.

To improve on the practice of minimizing the overall misclassification rate that treats all classes equally for imbalanced data, Qiao and Liu (2009); Qiao et al. (2010) up-weighted the priority or minor classes, but these approaches still consider the overall, albeit weighted, misclassification rate. The area under the ROC curve has been used as an alternative performance metric to the overall misclassification rate; however, in some fields, an explicit accuracy measure is still required. The Neyman-Pearson classification (Rigollet and Tong, 2011; Tong et al., 2016) aimed to control the type I error (incorrectly assigning the positive label to a negative observation) while minimizing the type II error. However, this is still a single-valued classification, and it cannot control the error rate for both classes concurrently. Moreover, it is unclear how to control the error rates for multiple priority classes in the multicategory classification setting.

The task to detect anomalies or new classes that do not exist in the training data is related to out-of-distribution (OOD) detection (Yang et al., 2021) or open-set recognition (OSR), but they typically fall under the umbrella of the single-valued classification and suffer

the same issues mentioned above. Recent works on adapting set-valued classifiers to handle OOD detection (Hechtlinger et al., 2018; Guan and Tibshirani, 2022) often focus on using the conformal prediction framework: first, a score function is obtained; second, a cutoff is determined using conformal splits; third, new observations are classified by thresholding the score with the cutoff. Cautious Deep Learning (CDL) (Hechtlinger et al., 2018) used the covariates' density given class as the score. However, the acceptance region for each class is learned with no regards to any other class; in the sense of minimizing the prediction set size, this approach was shown to be suboptimal (Dümbgen et al., 2008). Balanced and Conformal Optimized Prediction Sets (BCOPS) (Guan and Tibshirani, 2022) considered the classification problem between a given class k and the entire test data and thresholded the resulting estimates of $\mathbb{P}(Y = k \mid \boldsymbol{X} = \boldsymbol{x})$. Both methods depend on probability or density estimation, which is a challenging task when the dimension is large (Wu et al., 2010; Zhang et al., 2013a). It is hence desirable to propose a method without estimating probability. Most importantly, score functions in the aforementioned works were estimated without the goal of ultimately minimizing the prediction set size in mind. For example, though thresholding the true score $\mathbb{P}(Y = k \mid \boldsymbol{X} = \boldsymbol{x})$ can guarantee the minimization of the prediction set size, empirically a finite-sample estimate may not share this property.

In this article, we propose the Generalized Prediction Set (GPS) method to simultaneously solve the above limitations of single-valued classification, and overcome the difficulties in current set-valued classification methods. We have made three contributions in this article. First, we propose a new large-margin set-valued classification method with the capacity of OOD detection without involving probability estimation. Our model is estimated by minimizing the empirical prediction set size penalized by a term that encourages OOD detection, subject to a bounded misclassification rate for each class. Second, using weighted kernel and regularization, we enable feature selection for our method in high-dimensional settings. Finally, we conduct a thorough theoretical analysis of the proposed method, showing the convergence rate of the gap between its true and empirical misclassification rates, the convergence rate of the excess risk, and its variable selection consistency. It is worth noting that, in contrast to methods that solve an optimization problem involving all the classes simultaneously, our proposed method can be conducted for each class separately, hence is well-positioned for parallel computing, allowing fast classification even when there are many classes. The code is publicly available at `https://github.com/Zhou198/GPS`.

2. Preliminaries

We first review the background of the set-valued classification and the out-of-distribution detection problems.

2.1 Set-valued Classification

Consider a multicategory classification setting with input space $\mathcal{X} = \mathbb{R}^p$ and labels $\mathcal{Y} = \{1, \cdots, K\}$. Let $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ follow an unknown distribution $\mathcal{P}$. One way to obtain set-valued classifiers is to conduct a series of hypothesis tests that determine if the test observation belongs to a given class k. The set of all observations that are not rejected as being from class k is called the acceptance region for class k, denoted as $\mathcal{C}_k \subset \mathcal{X}$. Given all the $\mathcal{C}_k$, $k \in [K]$, a set-valued classifier $\phi : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ can be defined as $\phi(\boldsymbol{x}) := \{k : \boldsymbol{x} \in \mathcal{C}_k\}$,

that is, all the classes whose acceptance regions contain $\boldsymbol{x}$. Note that set-valued classification is different from multi-label classification (Zhang and Zhou, 2007; Sadinle et al., 2019), where each observation in the former has only one true label while the observation in the latter has multiple ground true labels.

Typically, there are two competing metrics for a set-valued classifier, namely, accuracy and efficiency. The accuracy may be quantified by the (unconditional) misclassification rate $\mathbb{P}\left(Y \notin \phi\left(\boldsymbol{X}\right)\right)$ or class-specific misclassification rate $\mathbb{P}\left(Y \notin \phi\left(\boldsymbol{X}\right) \mid Y = k\right)$, with the latter being the type I error rate for the hypothesis test. The efficiency (See the definition in Appendix B.4) is related to the size of the prediction set $|\phi(\boldsymbol{x})| := \sum_{k=1}^{K} \mathbb{1}\{\boldsymbol{x} \in \mathcal{C}_k\}$. A set-valued classifier with a small prediction set on average is more efficient and informative but may be less accurate. On the other hand, a set-valued classifier with $|\phi(\boldsymbol{x})| \equiv K$ everywhere is always correct, but contains no useful information, and hence is inefficient. In practice, one may want to balance the two metrics and obtain a classifier with both high accuracy and high efficiency. Classification with Confidence (Lei, 2014; Sadinle et al., 2019; Wang and Qiao, 2018, 2022) obtained a prediction set by deliberately maximizing the expected efficiency while controlling class-specific error rates. Denis and Hebiri (2015, 2017) worked with its dual problem, minimizing error rates with the prediction set size controlled.

Conformal Prediction (CP) (Vovk et al., 2005; Shafer and Vovk, 2008; Lei et al., 2013), under the assumption of exchangeable data distribution, is a framework for constructing prediction sets with a controlled error rate γ. Specifically, given the training data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$ and the query $\boldsymbol{X}_{n+1}$ associated with an unknown label Y_{n+1}, CP produces a prediction set ϕ with $\mathbb{P}(Y_{n+1} \notin \hat{\phi}(\boldsymbol{X}_{n+1})) \leq \gamma$. For example, in regression problems, a conformal prediction set for the query $\boldsymbol{X}_{n+1}$ can be defined as $\hat{\phi}(\boldsymbol{X}_{n+1}) := \{y : \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbb{1}\{s_j(y) \leq s_{n+1}(y)\} \geq \gamma\}$, where $s_j(y) := -|\hat{f}_y(\boldsymbol{X}_j) - Y_j|$ is the score function, and $\hat{f}_y$ is the regression function trained using $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n+1}$ with $Y_{n+1} = y$. The computation is expensive in conformal prediction since one must retrain the model for each observation with its possible labels. Lei et al. (2013); Lei (2014); Lei et al. (2015) used the split-conformal method to produce a prediction set with low computational cost. It is worth noting that conformal predictions only guarantee accuracy, but their efficiency is not explicitly optimized; instead, the efficiency depends on the choice of score functions.

The Classification with Reject Options (CRO) literature typically considers only rejections due to ambiguity. Herbei and Wegkamp (2006) and Ramaswamy et al. (2015) used 0-d-1 loss to quantify the loss for different prediction errors. For example, each misclassification costs 1 and each rejection costs a pre-specified $d \in [0, (K-1)/K]$. The Bayes optimal rule (Chow, 1970) under the 0-d-1 loss predicts label k to $\boldsymbol{x}$ if $k = \operatorname{argmax}_{k'} \mathbb{P}(Y = k' \mid \boldsymbol{x})$ and $\mathbb{P}(Y = k \mid \boldsymbol{x}) > 1 - d$, or rejects to predict $\boldsymbol{x}$ otherwise (in this case, we may think of $|\phi(\boldsymbol{x})| = K$). Bartlett and Wegkamp (2008) used the bent hinge loss as a surrogate to the 0-d-1 loss and proved the Fisher consistency. Zhang et al. (2018) introduced a refine option to consider smaller prediction sets with $1 < |\phi(\boldsymbol{x})| < K$. The CRO framework does not explicitly control the accuracy or the efficiency.

2.2 Out-of-distribution (OOD) Detection

Out-of-distribution (OOD) detection aims to identify anomaly observations that do not belong to the same distributions as the existing observations. We use the terms OOD

and anomaly interchangeably. Commonly used anomaly detection methods include one-class SVM (OCSVM), deep one-class classification (Ruff et al., 2018), density level set estimation (Breunig et al., 2000; Chen et al., 2017), and positive-unlabeled learning (PU learning) (du Plessis et al., 2014).

Suppose we have a random sample $\{\boldsymbol{x}_i\}_{i=1}^m$ from $\mathcal{X}$. Let $\Phi : \mathcal{X} \to \mathcal{H}$ be a kernel map from the input space to the feature space. OCSVM (Schölkopf et al., 2000) aims to separate data features from the origin with a maximum margin. The OCSVM marks an observation $\boldsymbol{x}$ as an anomaly if the decision function $f(\boldsymbol{x}) := \boldsymbol{w}^\top \Phi(\boldsymbol{x}) - \rho$ yields $f(\boldsymbol{x}) < 0$, where $\boldsymbol{w}$ and ρ are obtained by solving $\min_{\boldsymbol{w},\rho} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m\nu}\sum_{i=1}^m \xi_i - \rho$, subject to $\boldsymbol{w}^\top \Phi(\boldsymbol{x}_i) \geq \rho - \xi_i, \xi_i \geq 0, i \in [m]$. Here the tuning parameter ν controls the number of observations treated as anomalies. Ruff et al. (2018, 2021) achieved anomaly detection using deep learning.

Steinwart et al. (2005) showed that anomaly detection can be achieved by solving a classification problem between all the normal, existing, classes combined and the new OOD class, assuming that the proportion of the OOD class is known. Motivated by this observation, they proposed to train a cost-sensitive SVM between these two classes. To the same token, by having prior information about the OOD class, Liu et al. (2018) proposed Open Category Detection with a theoretical guarantee to achieve a pre-specified OOD detection rate via estimating the corresponding distribution. However, in practice, labeled OOD class data may be not observed. du Plessis et al. (2014, 2015) proposed PU learning to classify between the entire training data and the entire test data.

OOD detection can also be done in conjunction with a standard classification task. For example, Jumutc and Suykens (2013) and Hanczar and Sebag (2014) conducted two separated one-class SVMs in order to achieve binary classification and OOD detection. Other related works are open-set recognition (OSR) (Bendale and Boult, 2015) and generalized out-of-distribution detection (Yang et al., 2021). However, the classification component of their procedures, if any, still falls into the scope of the single-valued, as opposed to set-valued, classification.

3. Methodology

We first formulate the proposed GPS method as an optimization problem, which is decoupled into several sub-problems. To solve each sub-problem, we use kernel learning to find a decision function based on the training data from each of the K classes and the test data. Finally, we extend the method to kernel feature selection, to improve its performance for high-dimensional data.

3.1 Overview of Methodology

Suppose our training sample and test sample are two i.i.d. samples from distribution $\mathcal{P}$ and distribution $\mathcal{Q}$, respectively. There are K existing normal classes in both $\mathcal{P}$ and $\mathcal{Q}$; in addition, there are potentially OOD classes in $\mathcal{Q}$. Except for the OOD classes, the two distributions are only different in their prior probabilities for the K classes. The below assumption indicates that we only allow label shifts between the training and test data.

Assumption 1 *For each $k \in [K]$, the conditional probability density of $\boldsymbol{X}$ given class k, $p_k(\boldsymbol{x}) = p(\boldsymbol{x} \mid Y = k)$, is the same between distribution $\mathcal{P}$ and distribution $\mathcal{Q}$.*

Remark 2 *Assumption 1 in the OOD detection and open-set recognition literature is motivated by the emergence of a novel class (du Plessis et al., 2015; Yang et al., 2021; Guan and Tibshirani, 2022; Katz-Samuels et al., 2022; Garg et al., 2022). For instance, in the field of autonomous driving, the visual characteristics of known entities such as pedestrians, vehicles, and traffic signs remain constant. However, the inclusion of novel elements like electric scooters or unanticipated obstructions shifts the overall object distribution within the environment. In cybersecurity, the prevailing network activity tied to current intrusion types remains unaltered, yet the advent of a new attack transforms the frequency distribution across known attacks. In the context of natural language processing, e.g., analyzing sentiment in Yelp reviews, the linguistic attributes conveying sentiment tones (positive or negative) remain steadfast, but the emergence of new emotional tones reshapes the prevalence of sentiments.*

There are some related works on domain adaption to relax the assumption of identical class-conditional density $p_k(\boldsymbol{x})$ between training and test distribution. For instance, Tachet des Combes et al. (2020) proposed a generalized label shift to enhance its scope. In particular, they assumed the identical class-conditional density holds in a feature space, namely, there exists a representation mapping $\tilde{g}$ such that $p_{\mathcal{P}}(\tilde{g}(\boldsymbol{X}) \mid Y = k) = p_{\mathcal{Q}}(\tilde{g}(\boldsymbol{X}) \mid Y = k)$. However, it relies on an extra assumption of the partition in its feature space. Wu et al. (2019); Kumar et al. (2020) focused on relaxed label shift setting by assuming there is a minor divergence between class-conditional distributions, but the practicality of ascertaining this divergence remains a challenge (Garg et al., 2023). Zhang et al. (2013b); Gong et al. (2016) studied the location-scale generalized target shift (LS-GeTarS): by assuming there exists an affine transformation for each dimension of $\boldsymbol{X}$ given Y between the source and target domain, one can use the kernel embedding method to match the distribution of transformed labeled data from the source domain and unlabeled data from the target domain.

To extend our future work under the scenario of location-scale generalized target shift for normal classes, one potential strategy is to apply our method on the test data and transformed training data returned from the LS-GeTarS method.

Let $\phi(\cdot)$ be a set-valued classifier. Our goal is to maximize both efficiency and accuracy of $\phi(\cdot)$ for future test data drawn from $\mathcal{Q}$. To this end, consider minimizing the prediction set size with the class-specific misclassification rates bounded:

$$\min_{\phi} \ \mathbb{E}_{\mathcal{Q}}\left[|\phi(\boldsymbol{X})|\right], \quad \text{s.t. } \mathbb{P}_{\mathcal{Q}}\left(Y \notin \phi(\boldsymbol{X}) \mid Y = k\right) \leq \gamma_k \text{ for } k \in [K], \tag{1}$$

where the upper bound γ_k of the class-specific misclassification rate defined in the context of set-valued classification can be arbitrarily small. The inherent feasibility of the constraint in Problem (1) remains intact regardless of the true Bayes error (in single-valued classification). Even if the Bayes error is high, the stringent imposition of γ_k compels the method to produce larger prediction sets encompassing the true class label. In particular, the trivial case is to return a prediction set including all class labels, leading to a zero misclassification rate in the set-valued paradigm. On the other hand, a higher Bayes error signifies that observations face difficulty in being accurately differentiated from other classes due to intrinsic class overlaps. This underscores the need for set-valued classification to report plausible class labels, mitigating the risk caused by overly confident single-valued predictions.

Mostly different from Classification with Confidence methods, we drop the common restriction $|\phi(\boldsymbol{X})| \geq 1$ everywhere in this paper. This means that it is possible for $\phi(\boldsymbol{X})$ to be empty for certain observations (i.e., $|\phi(\boldsymbol{X})| = 0$), implying that $\boldsymbol{X}$ is unlike any of the existing classes in the training data. Note that observations with $|\phi(\boldsymbol{X})| = 0$ and $|\phi(\boldsymbol{X})| = K$ correspond to OOD-rejected observations and ambiguity-rejected observations, respectively. They are rejected for different reasons. Observations that are easy to classify generally should have $|\phi(\boldsymbol{X})| = 1$; the decision with $1 < |\phi(\boldsymbol{X})| < K$ corresponds to the refine option in Zhang et al. (2018).

Since $\mathbb{E}_{\mathcal{Q}}[|\phi(\boldsymbol{X})|]$ does not depend on the class label Y, it may be assessed using the unlabeled test data from $\mathcal{Q}$. Moreover, since we assume that $p_k(\boldsymbol{x})$ is the same between both distributions, we have that $\mathbb{P}_{\mathcal{Q}}(Y \notin \phi(\boldsymbol{X}) \mid Y = k) = \mathbb{P}_{\mathcal{P}}(Y \notin \phi(\boldsymbol{X}) \mid Y = k)$. This allows us to make use of the labeled training data from $\mathcal{P}$ to assess the misclassification rate in the constraint.

Recall that the set-valued classifier $\phi(\cdot)$ is defined using all K acceptance regions, $\phi(\boldsymbol{x}) := \{k : \boldsymbol{x} \in \mathcal{C}_k\}$. We use a decision function $f_k : \mathcal{X} \to \mathbb{R}$ to define $\mathcal{C}_k$, e.g., $\mathcal{C}_k := \{\boldsymbol{x} : f_k(\boldsymbol{x}) \geq 0\}$. Define the size (probability measure) of $\mathcal{C}_k$ as $\mathcal{R}(f_k) := \mathbb{P}_{\mathcal{Q}}(f_k(\boldsymbol{X}) \geq 0)$. Under these notations, $\mathbb{E}_{\mathcal{Q}}[|\phi(\boldsymbol{X})|] = \sum_{k=1}^{K} \mathcal{R}(f_k)$. Therefore, the optimization (1) can be decoupled to K separate optimization problems: for each $k \in [K]$, we solve

$$\min_{f_k \in \mathcal{F}} \mathcal{R}(f_k), \quad \text{s.t. } \mathcal{R}^+(f_k) \leq \gamma_k, \tag{2}$$

where $\mathcal{R}^+(f_k) := \mathbb{P}_{\mathcal{Q}}(f_k(\boldsymbol{X}) < 0 \mid Y = k)$ and $\mathcal{F}$ is a function space for f_k. Throughout this article, we consider the case $\gamma_k = \gamma$ for all k for simplicity without loss of generality. Problem (2) is equivalent to the Neyman-Pearson classification (Scott and Nowak, 2005; Rigollet and Tong, 2011) where class k is considered as the null class and the test data is the alternative class. Guan and Tibshirani (2022) used the plug-in method to approximate the Bayes optimal rule (see Theorem 9) of Problem (2).

In practice, one aims to estimate f_k (and hence ϕ) based on labeled training data $\{(\boldsymbol{x}_i, y_i = k)\}_{i \in \mathcal{G}_k}$ along with unlabeled test data $\{\boldsymbol{x}_j\}_{j \in \mathcal{G}_{te}}$, where $\mathcal{G}_k$ (with size $n_k := |\mathcal{G}_k|$) and $\mathcal{G}_{te}$ (with size $m := |\mathcal{G}_{te}|$) are index sets for observations in class k of the training data and the unlabeled test data, respectively. The expectations $\mathcal{R}^+(f_k)$ and $\mathcal{R}(f_k)$ can be replaced by their empirical estimates from the training and test data respectively:

$$\min_{f_k \in \mathcal{F}} \frac{1}{m} \sum_{j \in \mathcal{G}_{te}} \mathbb{1}(f_k(\boldsymbol{x}_j) \geq 0), \quad \text{s.t. } \frac{1}{n_k} \sum_{i \in \mathcal{G}_k} \mathbb{1}(f_k(\boldsymbol{x}_i) < 0) \leq \gamma, \tag{3}$$

3.2 Surrogate Loss and Kernel Learning

It is challenging to solve Problem (3) due to the use of the indicator function in both the objective and the constraint. A common practice is to replace it with a convex surrogate loss function. Here we use hinge loss $\ell(u) = [1-u]_+ = \max(0, 1-u)$ to replace $\mathbb{1}\{u < 0\}$ in the constraint of (3); likewise, $\mathbb{1}\{u \geq 0\}$ in the objective is replaced by $\ell(-u)$. See Figure 1. In addition, we use penalty function $J(f_k)$ to control the complexity of decision functions so that Problem (3) becomes:

$$\min_{f_k \in \mathcal{F}} \frac{1}{m} \sum_{j \in \mathcal{G}_{te}} [1 + f_k(\boldsymbol{x}_j)]_+ + \lambda J(f_k), \quad \text{s.t. } \frac{1}{n_k} \sum_{i \in \mathcal{G}_k} [1 - f_k(\boldsymbol{x}_i)]_+ \leq \gamma. \tag{4}$$

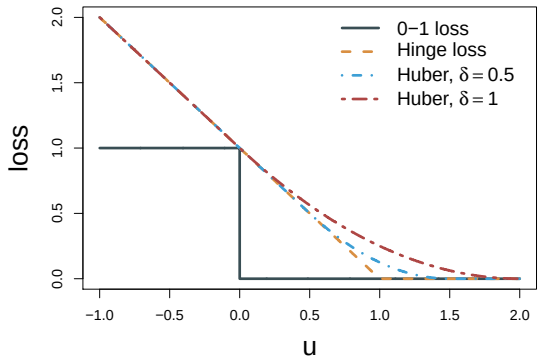

Figure 1: 0-1 loss, hinge loss, and two Huberized loss functions.

Here the decision function takes the form of $f_k(\boldsymbol{x}) := g_k(\boldsymbol{x}) - \rho_k$, where g_k belongs to a Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}_K$, associated with a kernel function $K(\cdot,\cdot)$. Throughout this article, we use the Gaussian kernel $K(\boldsymbol{x},\boldsymbol{x}') = \exp(-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\sigma^2)$ as used in OCSVM (Schölkopf et al., 2000). Moreover, following the common practice in the anomaly detection literature (Jumutc and Suykens, 2013; Schölkopf et al., 2018; Shilton et al., 2020), the penalty $J(f_k)$ is taken as $\frac{1}{2}\|g_k\|^2_{\mathcal{H}_K} - \rho_k$. Employing this penalty and the formulation of f_k in the optimization of (4) leads to an increased distance between the origin and the hyperplane $g_k(\boldsymbol{x}) = \rho_k$ (Schölkopf et al., 2000). The pursuit of this enlarged spatial distance results in a narrower acceptance region $\mathcal{C}_k$ dedicated to the class k in the feature space. This consequently helps to easily identify potential OOD points, leading to an improved OOD detection performance.

After introducing some slack variables and using the KKT conditions (see details of the derivations in Appendix A), Problem (4) yields

$$\hat{f}_k(\boldsymbol{x}) = \sum_{i\in\mathcal{G}_k} \hat{\alpha}_i K(\boldsymbol{x},\boldsymbol{x}_i) - \sum_{j\in\mathcal{G}_{te}} \hat{\beta}_j K(\boldsymbol{x},\boldsymbol{x}_j) - \hat{\rho}_k,$$

and hence the acceptance regions and the set-valued classifier can be obtained accordingly. The $\hat{\alpha}_i, \hat{\beta}_j$'s are the solutions to the below quadratic programming:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\theta}\ & \frac{1}{2}\left(\boldsymbol{\alpha}^\top \mathbf{G}_1 \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \mathbf{G}_2 \boldsymbol{\beta} - 2\boldsymbol{\alpha}^\top \mathbf{G}_3 \boldsymbol{\beta}\right) - \mathbf{1}_{n_k}^\top \boldsymbol{\alpha} - \mathbf{1}_m^\top \boldsymbol{\beta} + n_k \theta\gamma, \\ \text{s.t. } & \mathbf{0} \preceq \boldsymbol{\alpha} \preceq \theta \cdot \mathbf{1}_{n_k},\ \mathbf{0} \preceq \boldsymbol{\beta} \preceq C \cdot \mathbf{1}_m,\ \mathbf{1}_{n_k}^\top \boldsymbol{\alpha} - \mathbf{1}_m^\top \boldsymbol{\beta} = 1,\ \theta \geq 0, \end{aligned} \tag{5}$$

where $\mathbf{G}_1[i,i'] = K(\boldsymbol{x}_i,\boldsymbol{x}_{i'})$, $\mathbf{G}_2[j,j'] = K(\boldsymbol{x}_j,\boldsymbol{x}_{j'})$, $\mathbf{G}_3[i,j] = K(\boldsymbol{x}_i,\boldsymbol{x}_j)$, $\boldsymbol{\alpha} = (\ldots,\alpha_i,\ldots)^\top$, $\boldsymbol{\beta} = (\ldots,\beta_j,\ldots)^\top$, $i,i' \in \mathcal{G}_k, j,j' \in \mathcal{G}_{te}, C = (\lambda m)^{-1}$, and $\mathbf{1}_{n_k}$ denotes a vector with length n_k and all elements taking value 1. The offset $\hat{\rho}_k$ can be obtained by plugging $\hat{f}_k$ back to the Problem (4) after solving for $\hat{\alpha}_i$ and $\hat{\beta}_j$'s.

3.3 Kernel Feature Selection

For high-dimensional data, irrelevant or noisy features may degrade set-valued classifiers' performance in terms of efficiency, accuracy, and OOD detection. Feature or variable selection is necessary in these scenarios. For linear learning, sparse learning using sparsity penalties (Tibshirani, 1996; Zou and Hastie, 2005; Zhang, 2010) has been effective for feature

selection. For kernel learning, Allen (2013) and Chen et al. (2018) studied weighted kernel feature selection methods. The main idea of these methods is to compute the kernel matrix based on weighted features with a weight vector $\boldsymbol{d}$, and then impose a sparsity-inducing regularization for weight $\boldsymbol{d}$ in the objective function. Adopting this idea, our decision function f_k can be solved using the below optimization problem that enables kernel feature selection:

$$\begin{aligned} \min_{\boldsymbol{d},\boldsymbol{\alpha},\rho_k} \ & \frac{1}{m}\sum_{j\in\mathcal{G}_{te}} \ell\left(-f_k(\boldsymbol{d}\circ\boldsymbol{x}_j)\right) + \lambda_1 J(f_k(\boldsymbol{d}\circ\cdot)) + \lambda_2\|\boldsymbol{d}\|_1, \\ \text{s.t. } & \frac{1}{n_k}\sum_{i\in\mathcal{G}_k} \ell\left(f_k(\boldsymbol{d}\circ\boldsymbol{x}_i)\right) \le \gamma,\ \boldsymbol{0}\preceq\boldsymbol{d}\preceq\boldsymbol{1}, \end{aligned} \tag{6}$$

where $\circ$ stands for the Hadamard product. Our decision function is defined as $f_k(\boldsymbol{d}\circ\boldsymbol{x}) := g_k(\boldsymbol{d}\circ\boldsymbol{x}) - \rho_k$. The first term $g_k(\boldsymbol{d}\circ\cdot)$ comes from a RKHS associated with kernel function $K_{\boldsymbol{d}}(\cdot,\cdot)$. Here we define $K_{\boldsymbol{d}}(\boldsymbol{x}_i,\boldsymbol{x}_j) := K(\boldsymbol{d}\circ\boldsymbol{x}_i, \boldsymbol{d}\circ\boldsymbol{x}_j)$. By the Representer theorem (Kimeldorf and Wahba, 1971), for some α_i and ρ_k, the minimizer to (6) satisfies

$$\hat{f}_k(\boldsymbol{d}\circ\boldsymbol{x}) = \sum_{i=1}^{n_k+m} \alpha_i K_{\boldsymbol{d}}(\boldsymbol{x},\tilde{\boldsymbol{x}}_i) - \rho_k,$$

where $\tilde{\boldsymbol{x}}_i$ comes from the training data associated with label k when $i = 1,\ldots,n_k$ and from the unlabeled test data when $i = n_k+1,\ldots,n_k+m$. The model complexity function is taken as $J(f_k(\boldsymbol{d}\circ\cdot)) := \frac{1}{2}\sum_{i,j=1}^{n_k+m}\alpha_i\alpha_j K_{\boldsymbol{d}}(\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}}_j) - \rho_k$. Define the kernel matrix $\mathbf{K}_{\boldsymbol{d}}$ as $\mathbf{K}_{\boldsymbol{d}}[i,j] := K_{\boldsymbol{d}}(\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}}_j)$. Let $C_1 := (\lambda_1 m)^{-1}$ and $C_2 := \lambda_2/\lambda_1$. Then we rewrite (6) as

$$\begin{aligned} \min_{\boldsymbol{d},\boldsymbol{\alpha},\rho_k} \ & \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{K}_{\boldsymbol{d}}\boldsymbol{\alpha} - \rho_k + C_1\sum_{j\in\mathcal{G}_{te}} \ell\left(\rho_k - \mathbf{K}_{\boldsymbol{d}}[j,:]\boldsymbol{\alpha}\right) + C_2\|\boldsymbol{d}\|_1, \\ \text{s.t. } & \frac{1}{n_k}\sum_{i\in\mathcal{G}_k} \ell\left(\mathbf{K}_{\boldsymbol{d}}[i,:]\boldsymbol{\alpha} - \rho_k\right) \le \gamma,\ \boldsymbol{0}\preceq\boldsymbol{d}\preceq\boldsymbol{1}. \end{aligned} \tag{7}$$

Neither the objective nor the first constraint in (7) is convex with respect to $(\boldsymbol{d},\boldsymbol{\alpha},\rho_k)$ despite the convex surrogate loss function (which we chose to be the hinge loss). To resolve this issue, we use an iterative approach (the pseudocode is outlined in Appendix B.1) by alternatively fixing $\boldsymbol{d}$ while optimizing with respect to $(\boldsymbol{\alpha},\rho_k)$, which amounts to convex optimization, and fixing $(\boldsymbol{\alpha},\rho_k)$ while optimizing with respect to $\boldsymbol{d}$. The latter optimization is still not convex. But we can use a linear approximation of the kernel matrix with respect to $\boldsymbol{d}$ to make it convex (Zou and Li, 2008; Lee et al., 2012). In particular, we approximate the kernel matrix by expanding it at $\boldsymbol{d}'$:

$$\mathbf{K}_{\boldsymbol{d}}[i,j] \approx \mathbf{K}_{\boldsymbol{d}'}[i,j] + \nabla\mathbf{K}_{\boldsymbol{d}'}[i,j]^\top(\boldsymbol{d}-\boldsymbol{d}').$$

Define an $(n_k+m)\times(n_k+m)$ matrix $\mathbf{A}_{\boldsymbol{d}'}$ with $\mathbf{A}_{\boldsymbol{d}'}[i,j] := \mathbf{K}_{\boldsymbol{d}'}[i,j] - \nabla\mathbf{K}_{\boldsymbol{d}'}[i,j]^\top\boldsymbol{d}'$ and a $p\times(n_k+m)$ matrix $\mathbf{B}_{\boldsymbol{\alpha}}$ with $\mathbf{B}_{\boldsymbol{\alpha}}[:,i] := \sum_{j=1}^{n_k+m}\alpha_j\nabla\mathbf{K}_{\boldsymbol{d}'}[i,j]$, where p is the dimension of the data. These allow to approximate (7) with $(\boldsymbol{\alpha},\rho_k)$ fixed:

$$\begin{aligned} \min_{\boldsymbol{d}} \ & \frac{1}{2}\boldsymbol{d}^\top\mathbf{B}_{\boldsymbol{\alpha}}\boldsymbol{\alpha} + C_1\sum_{j\in\mathcal{G}_{te}} \ell\left(\rho_k - \mathbf{A}_{\boldsymbol{d}'}[j,:]\boldsymbol{\alpha} - \mathbf{B}_{\boldsymbol{\alpha}}[:,j]^\top\boldsymbol{d}\right) + C_2\|\boldsymbol{d}\|_1, \\ \text{s.t. } & \frac{1}{n_k}\sum_{i\in\mathcal{G}_k} \ell\left(\boldsymbol{A}_{\mathbf{d}'}[i,:]\boldsymbol{\alpha} + \mathbf{B}_{\boldsymbol{\alpha}}[:,i]^\top\boldsymbol{d} - \rho_k\right) \le \gamma,\ \boldsymbol{0}\preceq\boldsymbol{d}\preceq\boldsymbol{1}. \end{aligned} \tag{8}$$

The above optimization is convex with respect to $\boldsymbol{d}$. After we have obtained the decision function $\hat{f}_k(\boldsymbol{d}\circ\cdot)$ for $k\in[K]$, a set-valued classifier can be constructed as $\hat{\phi}(\boldsymbol{x})=\{k\in[K]: \hat{f}_k(\boldsymbol{d}\circ\boldsymbol{x})\geq 0\}$. If $|\hat{\phi}(\boldsymbol{x})|=0$ for some $\boldsymbol{x}$, then $\boldsymbol{x}$ is determined as an OOD point.

4. Statistical Learning Theory

In this section, we study the theoretical properties of our proposed classifier. We will focus on the kernel learning setting. Without loss of generality, we consider the decision function for class 1. For simplicity, we abuse the notation slightly by letting $f(\boldsymbol{x})=f_1(\boldsymbol{d}\circ\boldsymbol{x})$, omitting the weight $\boldsymbol{d}$.

Let f be an element from the hypothesis space defined as $\mathcal{F}_{s,s'}=\{f: f(\boldsymbol{x})=g(\boldsymbol{x})-\rho, g\in \mathcal{H}_{K_{\boldsymbol{d}}}, J(f)\leq s^2, \|\boldsymbol{d}\|_1\leq s', \boldsymbol{0}\preceq\boldsymbol{d}\preceq\boldsymbol{1}\}$. Denote a subspace of it that contains decision functions with bounded class 1 error rate as $\mathcal{F}^+_{s,s'}(\gamma)=\left\{f\in\mathcal{F}_{s,s'}: \mathbb{E}_{\mathcal{Q}}\left[\ell(f(\boldsymbol{X}))\mid Y=1\right]\leq\gamma\right\}$, and denote its empirical counterpart as $\widehat{\mathcal{F}}^+_{s,s'}(\gamma)=\{f\in\mathcal{F}_{s,s'}: \frac{1}{n_1}\sum_{i\in\mathcal{G}_1}\ell(f(\boldsymbol{x}_i))\leq\gamma\}$. These allow us to consider an optimization problem by moving the penalties $J(f)$ and $\|\boldsymbol{d}\|_1$ to the constraints. Specifically, we consider

$$\operatorname*{argmin}_{f\in\widehat{\mathcal{F}}^+_{s,s'}(\gamma)}\frac{1}{m}\sum_{j\in\mathcal{G}_{te}}\ell\left(-f(\boldsymbol{x}_j)\right). \tag{9}$$

Denote $\mathbb{P}\left(f(\boldsymbol{X})\geq 0\mid Y=1\right)$ and $\mathbb{E}\left[\ell(f(\boldsymbol{X}))\mid Y=1\right]$ as risk functions of class 1 under the 0-1 loss and the ℓ loss, respectively. Theorem 3 shows one can bound the former by controlling the empirical counterpart of the latter.

Theorem 3 *Assume $\kappa=\sup_{\boldsymbol{x}\in\mathcal{X}}\sqrt{K_{\boldsymbol{d}}(\boldsymbol{x},\boldsymbol{x})}$, and the loss function ℓ in (9) has a subderivative bounded by $c:=\sup_u|\ell'(u)|$. Let $\hat{f}$ be a solution to (9). With probability at least $1-\zeta$ over the training sample (incl. $\mathcal{G}_1$ and $\mathcal{G}_{te}$), we have*

$$\mathbb{E}_{\mathcal{Q}}\left[\ell(\hat{f}(\boldsymbol{X}))\mid Y=1\right]\leq\frac{1}{n_1}\sum_{i\in\mathcal{G}_1}\ell(\hat{f}(\boldsymbol{x}_i))+r_{n_1}(\zeta,s,s'), \tag{10}$$

where $r_{n_1}(\zeta,s,s')=\frac{(\sqrt{2}s+2)c\kappa}{\sqrt{n_1}}\left(2+3\sqrt{2\log(2/\zeta)}\right)$.

For the Gaussian kernel employed throughout this article, $\kappa=1$. It is noteworthy to highlight that the Gaussian kernel's shift-invariant property allows the effect of s' on $\boldsymbol{d}$ to be absorbed into κ due to the fact of $K_{\boldsymbol{d}}(\boldsymbol{x},\boldsymbol{x})=\exp(-\|\boldsymbol{d}\circ\boldsymbol{x}-\boldsymbol{d}\circ\boldsymbol{x}\|^2/\sigma^2)=1$ regardless the restriction $\|\boldsymbol{d}\|_1\leq s'$. The shift-invariant property of the Gaussian kernel also extends to other potential kernels, such as the Laplace kernel and the Cauchy kernel.

Theorem 3 applies to any convex loss function ℓ bounded from below by the 0-1 loss with a Lipschitz constant c satisfying $|\ell(u_1)-\ell(u_2)|\leq c|u_1-u_2|$ for any u_1 and u_2. In particular, $c=1$ for the hinge loss, the Huberized squared hinge loss (see Appendix B.2), and the logistic loss; the exponential loss has a Lipschitz constant only when the input space is bounded.

Bounding the empirical ℓ-risk $\frac{1}{n_1}\sum_{i=1}^{n_1}\ell(\hat{f}(\boldsymbol{x}_i))$ by γ may still lead to $\mathbb{E}_{\mathcal{Q}}\big[\ell(\hat{f}(\boldsymbol{X}))\mid Y=1\big]$ exceeding γ. Hence, to better control the true misclassification rate, one can strengthen the constraint by bounding $\frac{1}{n_1}\sum_{i=1}^{n_1}\ell(\hat{f}(\boldsymbol{x}_i))$ by $\gamma-\varepsilon$ with $\varepsilon=r_{n_1}(\zeta,s,s')$.

Let the ℓ-ambiguity be $\mathcal{R}_\ell(f) := \mathbb{E}_{\mathcal{Q}}[\ell(-f(\boldsymbol{X}))]$. Theorem 4 shows how the sample size and hypothesis space affect the convergence of the estimation error $\mathcal{R}_\ell(\hat{f}) - \inf_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \mathcal{R}_\ell(f)$.

Theorem 4 *Under the assumption in Theorem 3 with Huberized squared hinge loss, let* $\varepsilon = r_{n_1}(\zeta, s, s')$ *and*

$$\hat{f} = \operatorname*{argmin}_{f\in\widehat{\mathcal{F}}^+_{s,s'}(\gamma-\varepsilon)} \frac{1}{m}\sum_{j\in\mathcal{G}_{te}} \ell\left(-f(\boldsymbol{x}_j)\right).$$

With probability at least $1-2\zeta$, *we have*
(1) $\mathbb{E}_{\mathcal{Q}}\left[\ell(\hat{f}(\boldsymbol{X}))\,|\,Y=1\right] \le \gamma$; *(2)* $\mathcal{R}_\ell(\hat{f}) - \inf_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \mathcal{R}_\ell(f) \le 2r_m(\zeta,s,s') + \dfrac{(4+\delta)r_{n_1}(\zeta,s,s')}{\gamma - 2r_{n_1}(\zeta,s,s')}$,
where m *is the size of the sample from the distribution* $\mathcal{Q}$.

In order to ensure an estimation $\hat{f} \in \mathcal{F}^+_{s,s'}(\gamma)$, by Theorem 3, we restrict the hypothesis space as $\widehat{\mathcal{F}}^+_{s,s'}(\gamma-\varepsilon)$. In this setting, the estimation error converges at a rate of $O(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n_1}})$. This indicates it is possible for the empirical ℓ-ambiguity to converge to its minimum in a given hypothesis space using our method. Note that Theorem 4 also applies to the hinge loss (where $\delta = 0$).

Proposition 5 allows to bound the excess ambiguity by the excess ℓ-ambiguity.

Proposition 5 *(Rigollet and Tong, 2011) Let* $\mathcal{R}(\cdot)$ *and* $\mathcal{R}^+(\cdot)$ *be defined using the 0-1 loss as in (2). Given any function* $\tilde{f}$, *the following inequality holds*

$$\mathcal{R}(\tilde{f}) - \inf_{\mathcal{R}^+(f)\le\gamma} \mathcal{R}(f) \le \mathcal{R}_\ell(\tilde{f}) - \inf_{\mathcal{R}^+(f)\le\gamma} \mathcal{R}_\ell(f).$$

Proposition 5 shows that we can control the excess ambiguity by controlling the excess ℓ-ambiguity using a good estimate $\hat{f}$.

Let $\boldsymbol{d}^* = (d_t^*)$ be the weight in $f^* \in \operatorname{arginf}_{f\in\mathcal{F}^+_{\infty,p}(\gamma)} \mathcal{R}_\ell(f)$. The important and unimportant features are referred as those $\boldsymbol{x}_{\cdot,t}$ with $d_t^* > 0$ and $d_t^* = 0$, respectively. Theorem 6 shows feature selection consistency in terms of the sign of weight $\widehat{\boldsymbol{d}}$ under some conditions.

Theorem 6 *Consider the hypothesis space as a Gaussian kernel RKHS and the input space* $\mathcal{X}$ *is bounded. Let a Lipschitz continuous loss function* $\ell(u)$ *be differentiable, and* $\widehat{\boldsymbol{d}} = (\hat{d}_t)$ *be the solution to (9). Assume* f^* *comes from the RKHS and* $\left.\frac{\partial \mathbb{E}_{\mathcal{Q}}[\ell(-f^*(\boldsymbol{X}))]}{\partial d_t}\right|_{d_t=0,\ d_{t'}=d^*_{t'},\ \forall t'\neq t}$ *are negative and non-negative for those important and unimportant features* $\boldsymbol{x}_{\cdot,t}$, *respectively, then*

$$\mathbb{P}\left[\operatorname{sign}(\hat{d}_t) = \operatorname{sign}(d_t^*)\right] \longrightarrow 1,\ t\in[p].$$

Under the assumption for the partial derivatives, the optimization procedure in (9) will lead to a solution where the weight $\hat{d}_t > 0$ for important features and $= 0$ for unimportant features, respectively, for a large enough sample. Similar assumptions were used in Fan and Peng (2004) and Fan and Lv (2010).

5. Numerical Studies

For the GPS methods, we use `cvxopt` and `scipy` in Python to solve the convex optimization problems involved. For competing set-valued classification methods, e.g., one-class SVM with split-conformal (Lei et al., 2013) (OCSVM), one-versus-rest SVM with split-conformal, CDL (Hechtlinger et al., 2018), and BCOPS-RF (Guan and Tibshirani, 2022) involving Random Forest, we use their implementations in the `scikit-learn` library. We report the empirical class-specific accuracy, the efficiency, and the OOD detection rate. Furthermore, we report AUC-Det (the area under the curve of OOD detection rate v.s. accuracy) and AUC-Eff (the area under the curve of efficiency v.s. accuracy) to assess the overall performance of set-valued classifiers across various accuracies. See their definitions in Appendix B.4. We report the average of these metrics over 200 replications on a completely new set of test data. The implementation details are in Appendix B.5.

It is difficult to conduct an apple-to-apple comparison based on the three metrics: accuracy, efficiency and OOD detection rate, as a higher accuracy for normal classes is often associated with a lower efficiency and a lower OOD detection rate. Following Wang and Qiao (2018, 2022); Guan and Tibshirani (2022), we make use of the split-conformal method (Lei et al., 2013; Lei, 2014; Lei et al., 2015) to make the accuracy of all methods roughly the same, so that methods may be compared in terms of the efficiency and the detection rate. Specifically, let the decision rule $\hat{f}_k$ learned from each method be the conformal score function. Given any class k, we randomly split the data into the estimation set and the calibration set. We use the first set to train the classifier and the score function, and the second part to conduct the calibration and parameter tuning. A threshold $\hat{\tau}_k$ is taken as $(\gamma \times 100)$-th percentile of the scores among labeled data in the calibration set, given the pre-specified significance level γ. Finally, the prediction set of a given $\boldsymbol{x}$ is $\{k \in [K] : \hat{f}_k(\boldsymbol{x}) \geq \hat{\tau}_k\}$. Moreover, when $\hat{f}_k(\boldsymbol{x}) < \hat{\tau}_k$ for all $k \in [K]$, $\boldsymbol{x}$ is determined to be an OOD point. This decision rule applied on the unlabeled data in the calibration set helps us to select an optimal tuning parameter with the smallest prediction set size. The above split conformal calibration step is applied to all methods, ensuring that class-specific accuracy is $1 - \gamma$ on expectation.

5.1 Simulations

In this section, we conduct comparisons on two synthetic data sets.

Example 1: We generate data from four multivariate normal classes ($k = 1, \ldots, 4$) where $\boldsymbol{X} \mid Y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$, $\boldsymbol{\mu}_k = r_k(\cos\theta_k, \sin\theta_k)^\top$, $r_k \sim \text{Uniform}(0, 6)$, $\theta_k \sim \text{Uniform}(0, 2\pi)$ and $\Sigma_k^{1/2} = \text{diag}(\sigma_k, \sigma_k) + \varepsilon_k$, where $\sigma_k \sim \text{Uniform}(0.8, 1.2)$ and $\varepsilon_k \sim \text{Uniform}(-0.5, 0.5)$. The OOD class is a mixture of four uniform distributions (with equal weights) on four rectangle regions as shown in the left panel of Figure 2. After generating the above two-dimensional data, we augment them with 8 independent noise variables normally distributed with mean 0 and standard deviation 0.1.

In the left panel of Figure 2, the proposed GPS method is applied to only the first two dimensions to return the colored acceptance regions. The contours display the boundaries of acceptance regions for 4 classes. We can see that the OOD class is successfully ruled out from those acceptance regions, and points falling into the intersections of acceptance regions are observations difficult to be classified. For the right panel, we show the scatter plot of all the test data points using the first two principal components. Here the prediction sets

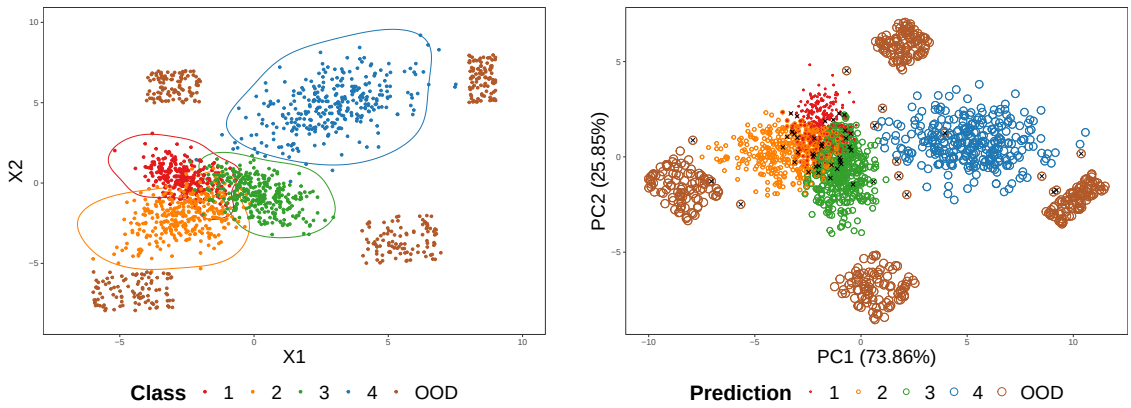

Figure 2: GPS classification on Example 1. Left panel: The scatter plot for the first two dimensions. Colored contours are boundaries of acceptance regions for the four normal classes with $\gamma = 5\%$. Right Panel: The scatter plot of the first two principal components for the test data, where the color of the circles indicates the predicted class. The radius of the circles differs for different classes to allow the visualization of those points which are classified into multiple classes. Points whose true labels do not belong to the prediction sets are labeled as black crosses. Points with empty prediction sets are marked as brown, indicating that they are determined to be OOD points.

returned by GPS (trained on all dimensions) are visualized by circles with different radii. Circles centering at the same observation but with different radii show that the prediction set size for that observation is more than 1. This means this observation is in an overlap area and is difficult to be confidently predicted using a single label. Brown circles denote those observations with $|\hat{\phi}(\boldsymbol{x})| = 0$, and hence are deemed as OOD points. Those with black crosses are those cases with $y \notin \hat{\phi}(\boldsymbol{x})$. From the right panel, we can see that this type of decision is more likely to appear on the tail of class distributions.

We set the nominal error rate $\gamma = 5\%$ in this simulation. It is noteworthy that both OCSVM and SVM were not initially designed for set-valued classification, but they were

		OCSVM	SVM	CDL	BCOPS-RF	GPS	GPSKFS
Accuracy	Class 1	96.5±0.085	96.2±0.089	96.3±0.116	96.1±0.097	95.5±0.113	95.5±0.12
	Class 2	95.4±0.105	96.9±0.076	95.5±0.118	96.7±0.083	95.4±0.084	95.5±0.091
	Class 3	96.4±0.079	95.7±0.086	95.4±0.108	95.6±0.112	95.3±0.097	95.5±0.103
	Class 4	96.1±0.087	96.5±0.088	95.3±0.116	96.3±0.086	95.1±0.118	95.3±0.119
Detection Rate		**99.9**±0.025	0±0	97.7±0.141	76.4±0.86	**99.9**±0.029	99.4±0.244
Efficiency		81.3±0.117	68.6±0.596	79.4±0.118	88.6±0.086	**90.8**±0.074	90.6±0.113
AUC-Det		**98.2**±0.051	7.0±0.49	95.4±0.176	74.5±0.519	96.7±0.132	94.2±0.285
AUC-Eff		80.1±0.119	72.7±0.265	78.0±0.108	87.3±0.08	**89.0**±0.055	88.4±0.123

Table 1: Average performance metrics for Example 1

adapted to it using the conformal prediction framework. Based on the results shown in Table 1, OCSVM outperforms other methods in the OOD detection performance due to its originally designed anomaly detection capability, but suffers in terms of the efficiency; SVM has the worst detection rate and the worst efficiency at the current accuracy in this example. CDL and BCOPS-RF, though designed to detect OOD data points, have a less competitive performance than OCSVM. However, they are generally more efficient than OCSVM and SVM. The proposed GPS methods discover more than 99% of the OOD points, performing on par with OCSVM and outperforming all other methods. Unlike the previous four methods, GPS and GPSKFS deliberately consider efficiency maximization, yielding the most efficient/informative prediction sets. Regarding the performance across a range of accuracy levels, GPS and GPSKFS have competitive AUCs on both detection and efficiency, indicating a better balance between the two metrics.

Example 2: This example is similar to Example 3 in Wang and Qiao (2018). We first generate radius-angle pairs (R, θ), where $\theta \sim \text{Uniform}(0, 2\pi)$. $R \mid Y = 1 \sim \text{Uniform}(0, 5)$, $R \mid Y = 2 \sim \text{Uniform}(4, 9)$ and $R \mid Y = 3 \sim \text{Uniform}(8, 13)$. For the OOD class in the test data, its radius $R \sim \text{Uniform}(15, 20)$. Then we define a 2-dimensional data vector $(R \cdot \cos\theta, R \cdot \sin\theta)$. Finally, we add 98 independent standard normal noise variables.

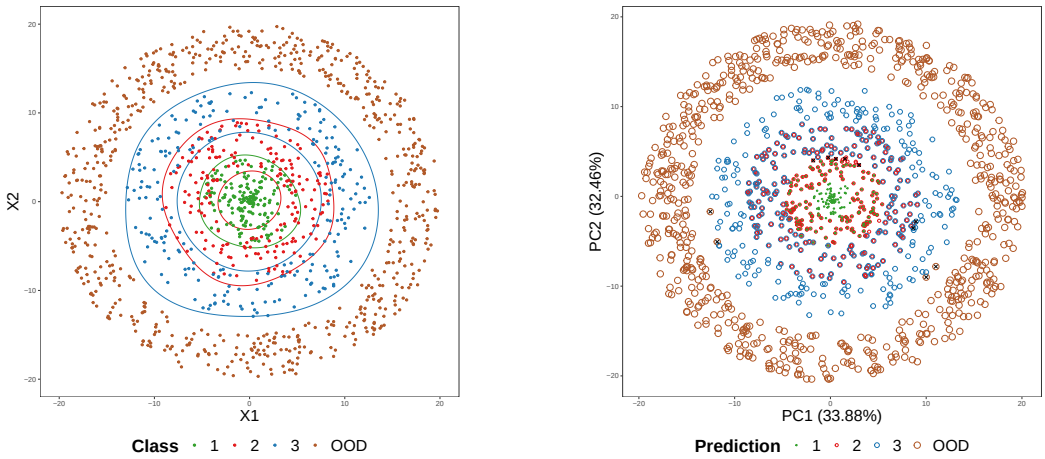

Figure 3: GPS classification on Example 2. The left panel shows the true class labels and the right panel shows the predicted classes.

For this example, we set the nominal error rate $\gamma = 1\%$. Similar to the data visualization in Example 1, we display the boundaries of the acceptance regions for Example 2 in the left panel of Figure 3. In this example, it is evident from Table 2 that OCSVM continues to demonstrate an impressive OOD detection performance, at the cost of poor efficiency, while SVM exhibits the lowest detection rate despite having higher efficiency. CDL and BCOPS-RF display poor detection rates and unacceptable efficiencies. Contrastively, when the performances are evaluated at the prescribed accuracy or across a range of accuracies, both GPS and GPSKFS demonstrate significantly higher OOD detection rates and higher efficiencies compared to all other methods (except that the efficiency of GPS is not as good as SVM, which has zero detection rate.) The GPSKFS further improves the performance

		OCSVM	SVM	CDL	BCOPS-RF	GPS	GPSKFS
Accuracy	Class 1	97.6±0.096	99±0.074	98.7±0.081	98.7±0.093	98.1±0.093	98±0.098
	Class 2	99.3±0.035	98.9±0.071	99.1±0.057	99.4±0.045	99.2±0.047	99.3±0.061
	Class 3	98.9±0.055	98.7±0.069	98.9±0.066	99.4±0.054	98.4±0.071	99±0.059
Detection Rate		98.8±0.044	0±0	33.4±0.289	25.4±0.702	99.4±0.056	**100**±0
Efficiency		36.3±0.128	53.5±0.113	17.2±0.134	26±0.36	42±0.123	**84.2**±0.173
AUC-Det		93.0±0.083	0.0±0.0	31.3±0.306	31.6±0.598	93.6±0.11	**95.1**±0.0
AUC-Eff		33.4±0.13	50.6±0.1	16.4±0.148	25.7±0.323	39.0±0.113	**79.9**±0.125

Table 2: Average performance metrics for Example 2

of the regular GPS method by virtue of its kernel feature selection capability. Comparing the performance of GPSKFS in Example 1 and Example 2, it appears that its performance is more effective when there are more noise features—98 in the current example.

As mentioned earlier, in the current setting, three performance metrics (accuracy, efficiency, and detection rate) are of interest. Here accuracy is controllable by the user, and a higher accuracy is often associated with lower efficiency and detection rate. These trade-offs are demonstrated in Figure 4, where the top row shows ROC-type curves of detection rate against accuracy for all methods in different examples, and the bottom row shows those curves of efficiency against accuracy. The first five rows of Tables 1 and 2 are corresponding to a particular prescribed accuracy for each example respectively, while Figure 4 provides a fuller picture across a range of accuracy levels. The AUC-Det and AUC-Eff are calculated as integrals of the corresponding curve from $1-2\gamma$ to 1, which is a neighborhood near $1-\gamma$ that is practically relevant.

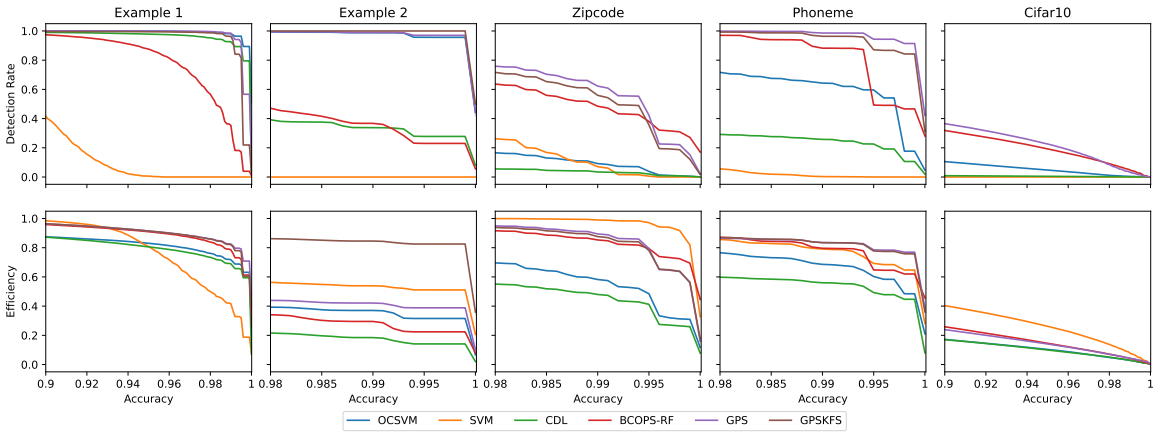

Figure 4: OOD detection rate and efficiency under varied accuracy.

5.2 Real Data Analysis

In this section, we conduct comparisons by considering three real data sets: Zipcode, Phoneme, and Cifar10.

Zipcode: This is a hand-written Zipcode data set consisting of 7291 training points with 256 features as well as 2007 test points. We first merge them together and treat labels 0, 6, 8, and 9 as the normal classes and the remaining as the OOD class. To generate our training data, we randomly sample from each of the normal classes with subsample sizes 550, 580, 495, and 574, respectively. The remaining data points from the normal classes, and all points in the OOD class, will form the test data. We set the nominal error rate as $\gamma = 1\%$ in this example.

Phoneme: This data set with 4509 points and 256 features is formed by selecting five phonemes for classification based on digitized speech. The phonemes are transcribed as follows: "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". We treat class "sh" as the OOD in test data and sample the other four classes with sizes around 500 for training data. Here the nominal error rate is also specified as 1% for each class.

Cifar10: This colored image data consists of 10 classes with 50000 training points and 10000 test points, where each image has dimension $32 \times 32 \times 3$. In this data set, we choose animals "bird", "cat", "deer", "dog", "frog", "horse" as 6 normal classes and the remaining 4 transportations ("airplaine", "automobile", "ship", and "truck") as the OOD class. To reduce the computation burden, we sample around 800 points from each normal class in the training data. Here the nominal error rate is set as $\gamma = 5\%$ for each normal class.

		OCSVM	SVM	CDL	BCOPS-RF	GPS	GPSKFS
Accuracy	Digit 0	99.1±0.032	99.4±0.027	99.1±0.036	99.5±0.026	99.1±0.039	98.5±0.044
	Digit 6	99±0.044	98.7±0.051	99.2±0.054	98.6±0.054	98.8±0.055	98.2±0.058
	Digit 8	98.2±0.078	98.8±0.045	98.7±0.066	98.5±0.07	98.2±0.078	97.2±0.077
	Digit 9	99.7±0.022	99.1±0.027	99.4±0.024	99.3±0.039	99±0.04	98.6±0.048
Detection Rate		8.2±0.254	5.6±0.329	3.3±0.074	46.9±0.524	66.2±0.706	**73**±0.468
Efficiency		52.5±0.473	**99**±0.036	43.5±0.348	84.2±0.336	90.9±0.329	94.1±0.229
AUC-Det		8.0±0.101	8.1±0.206	3.1±0.061	44.0±0.287	**51.1**±0.407	46.4±0.548
AUC-Eff		50.6±0.137	**92.1**±0.056	41.4±0.247	78.8±0.203	79.2±0.265	78.2±0.361

Table 3: Average performance metrics on Zipcode

Tables 3 to 5 provide an overview of the average performance of various methods on Zipcode, Phoneme, and Cifar10, respectively. Here we omit the GPSKFS on Cifar10 from Table 5 due to the limited computation resource. As these are all shallow models, their performances on Cifar10 are unsatisfactory, with lower OOD detection rates and efficiencies.

In these three data sets, the OCSVM method does not perform as well in OOD detection as it did in the previous simulations. SVM, in contrast, has a relatively competitive efficiency (as well as AUC-Eff) but at the expense of poor OOD detection. According to Figure 4, SVM has an inadequate balance on metrics, with the worst OOD detection rate when the class-specific accuracy is close to 1, despite dominating efficiencies. CDL performs poorly on both Zipcode and Cifar10 data sets, while BCOPS-RF is promising compared to other three competing methods. For the proposed GPS methods, they consistently return competitive, and most of the time, the best results. Figure 4, specifically columns 3 to 5, shows that

		OCSVM	SVM	CDL	BCOPS-RF	GPS	GPSKFS
Accuracy	Class "aa"	97.6±0.053	99.6±0.04	98±0.073	99.2±0.041	97.5±0.088	98.1±0.087
	Class "ao"	99.7±0.018	99.5±0.037	99.5±0.039	99.2±0.041	99±0.049	98.3±0.064
	Class "dcl"	98.9±0.032	98.8±0.057	99.4±0.046	99.2±0.062	98.7±0.062	98.3±0.068
	Class "iy"	99.5±0.034	99.4±0.042	99±0.05	99.4±0.035	99.4±0.04	98.5±0.064
Detection Rate		53±0.688	0±0.017	24.9±0.475	69.7±1.577	**98.6**±0.12	98.5±0.159
Efficiency		67.8±0.246	69.1±0.439	55.7±0.161	72±0.568	84±0.219	**85.9**±0.143
AUC-Det		55.1±0.402	1.0±0.099	22.3±0.349	75.3±0.755	**92.7**±0.118	89.6±0.437
AUC-Eff		62.8±0.153	72.9±0.233	51.5±0.173	73.1±0.256	**78.6**±0.135	78.4±0.187

Table 4: Average performance metrics on Phoneme

		OCSVM	SVM	CDL	BCOPS-RF	GPS
Accuracy	Bird	95.3±0.111	94.1±0.126	94.5±0.162	94.9±0.117	95.3±0.119
	Cat	95.9±0.071	96.4±0.115	95.4±0.1	96.6±0.098	96.2±0.118
	Deer	96.5±0.057	95.9±0.114	95.9±0.095	94.1±0.107	94.2±0.133
	Dog	94.1±0.105	96.1±0.098	94.6±0.134	96.9±0.092	96.4±0.077
	Frog	95.3±0.118	96.4±0.092	94.6±0.118	94.6±0.113	93.1±0.13
	Horse	95.7±0.086	95.4±0.098	95.8±0.101	94.4±0.129	94.1±0.121
Detection Rate		4±0.068	0±0	0.5±0.014	18.8±0.221	**22.1**±0.239
Efficiency		9.2±0.059	**23.3**±0.147	9.6±0.07	14.2±0.143	14.5±0.134
AUC-Det		4.7±0.031	0.0±0.0	0.5±0.008	17.9±0.172	**20.3**±0.12
AUC-Eff		9.7±0.014	**24.2**±0.054	9.5±0.025	14.0±0.08	13.4±0.054

Table 5: Average performance metrics on Cifar10

GPS and GPSKFS provide competitive OOD detection performance and higher efficiency over the range of accuracy. In summary, while it is not expected for a single method to outperform all others in all metrics and situations, the GPS methods exhibit satisfactory balanced results.

Furthermore, we conducted an experiment to evaluate the stability of the proposed method when the proportion of the OOD points among the test data increases (see results in Appendix B.6). The experimental results show that the efficiency does not change too much, and the OOD detection rate can be improved when there are increasing OOD observations in the test data.

5.3 Comparison with OSR Methods

There are a group of open-set recognition (OSR) methods that are capable of both OOD detection and normal class classification. Unlike CDL and BCOPS-RF compared in the previous sections, OSR methods generally only consider single-valued predictions for normal classes.

In this section, we compare the proposed GPS with a selection of deep learning-based OSR methods on the Cifar10 data set. Our results empirically demonstrate the inadequacy of single-valued classification as previously highlighted in Section 1. Specifically, we use the PyTorch library `pytorch-ood` (Kirchheim et al., 2022) to conduct open-set classification based on the three approaches proposed in Dhamija et al. (2018) and Liu et al. (2020), namely, EOS, Objectosphere, and EBL. We use the default values of parameters and employ the same neural network architecture, WideResNet (Zagoruyko and Komodakis, 2016), for these three methods. To ensure a fair comparison between the GPS methods and these deep learning-based models, we adopt the hybrid approach outlined in Ruff et al. (2019) by using the embeddings learned from the neural networks as input features for GPS. We refer to this variant of GPS as Hybrid GPS.

Note that different OSR model learns different embeddings, which in turn lead to different GPS results. Once the embeddings are obtained, we sample around 500 images from each normal class to train the GPS model for computational efficiency. For each of the OSR approaches, there is an option to set the threshold to adjust the strength of detection. We choose the threshold so that about 95% of the normal class observations are correctly marked as non-OOD. For the GPS methods, we set the nominal error rate to be $\gamma = 5\%$. Additionally, we include results from the regular shallow GPS (in Table 5, without deep representation learning) in Table 6 as a reference.

		Shallow GPS	EOS	Hybrid GPS	Objectosphere	Hybrid GPS	EBL	Hybrid GPS
Accuracy	Bird	95.3±0.119	83.6±1.937	95.4±0.273	76.3±2.516	94.1±0.491	89.5±0.801	95.9±0.356
	Cat	96.2±0.118	73.4±1.319	94.5±0.495	62.9±3.517	91.9±0.786	74.2±2.111	92.5±0.566
	Deer	94.2±0.133	88.4±1.56	96±0.496	77.5±3.159	94.8±0.603	88.9±0.912	96.7±0.466
	Dog	96.4±0.077	84.3±1.561	95.7±0.667	71.9±3.945	94.3±0.622	81.3±2.023	96±0.735
	Frog	93.1±0.13	91±1.091	95.9±0.577	82.5±4.05	93.3±0.71	92.2±0.599	96.5±0.438
	Horse	94.1±0.121	92.1±0.918	95.8±0.543	80.9±4.312	94.6±0.612	92.3±1.173	96.6±0.686
Detection Rate		22.1±0.239	**98.3**±0.194	**95.3**±0.344	94.8±1.573	93.7±1.787	48.1±1.107	52.7±4.973
Efficiency		14.5±0.134	**100**±0	**85.9**±3.684	100±0	75.8±5.992	100±0	77.8±6.161
AUC-Det		20.3±0.12	⁄	79.4±2.836	⁄	83.1±1.918	⁄	51.9±1.034
AUC-Eff		13.4±0.054	⁄	76.2±0.891	⁄	65.5±3.87	⁄	72.7±1.007

Table 6: Comparisons between OSR and (Hybrid) GPS on Cifar10

As shown in Table 6, the accuracy of each OSR method (EOS, Objectosphere, EBL) for each class is lower than those of the Hybrid GPS methods, while the latters are almost close to 95%. This is expected, as these OSR methods make single-label predictions (resulting in 100% efficiency), while the Hybrid GPS methods may yield more than one label for some observations, and afford a better class-specific accuracy. In terms of OOD detection, both OSR methods and their corresponding Hybrid GPS methods exhibit similar performance to each other.

The performance of Hybrid GPS highly depends on the embedding learned from the corresponding OSR model. It appears that the embeddings learned from EOS outperform those from Objectosphere and EBL, and no embedding at all (leading to the shallow GPS), due to the better performances from EOS-based Hybrid GPS method. Focusing on the EOS method and the EOS-based Hybrid GPS, we see a clear trade-off. EOS has 100% efficiency,

at the cost of suboptimal accuracy; Hybrid GPS improves the accuracy at the expense of efficiency. Their detection rates are somewhat similar and both are fairly high.

The detection rate and efficiency of the Hybrid GPS are significantly improved over the shallow GPS due to the representation learning. This suggests that exploring an end-to-end GPS integrated with neural networks is a promising direction for future research.

6. Conclusion

Motivated by the potential issues in the conventional single-valued classification, where absolute predictions are made without confidence guarantee, accuracy control for prioritized classes is not possible, and there is reliance on the assumption of no distribution shift, we proposed a set-valued classification method, i.e., GPS, to simultaneously address them. This method provides accuracy guarantees on each normal class and has OOD detection capability on label shift distribution data. In contrast to the existing set-valued classifier with OOD detection that may return less informative decisions, we explicitly minimized the prediction set size under a constraint of class-specific accuracy. Our experimental results demonstrate that GPS methods enjoy higher efficiency performance and higher OOD detection rates.

The GPS methods have a feature selection property in the kernel learning context. Additionally, we make use of the "divide-and-conquer" strategy to break down a large-scale problem into many sub-problems, involving only two classes of data, namely an existing class k and the test data which may include OOD classes. Because all sub-problems can be solved in an *embarrassingly parallel* way, the computational time can be greatly reduced compared to solving a large-scale optimization problem involving all classes (Zhang et al., 2018; Wang and Qiao, 2018, 2022).

There are many works targeting anomaly detection with different techniques. For example, one can use auxiliary data (Hendrycks et al., 2018; Neal et al., 2018), e.g., outlier exposure data or generating synthetic data, instead of the semi-supervised data as in our setting. However, unless the auxiliary data can mimic the unknown test data distribution, it is hard to obtain any optimality guarantee on the ambiguity risk. Most importantly, those works, including generalized OOD detection or OSR (Geng et al., 2020; Yang et al., 2021), belong to the single-valued classification framework with no accuracy guarantee. We believe that our proposed GPS methods enrich generalized OOD detection and OSR by introducing different decisions, i.e., ambiguous observations ($1 < |\phi(\boldsymbol{x})| < K$) and ambiguity-rejected observations ($|\phi(\boldsymbol{x})| = K$).

There are several areas of research that offer potential future investigation. One such area pertains to the derivation of theoretical guarantees for OOD detection. This task may rely on certain distribution assumptions regarding the OOD, as evidenced by prior studies (Liu et al., 2018; Fang et al., 2021). However, obtaining access to OOD information presents a significant challenge. Another direction is that, compared to network-based OSR, generalizing the use of GPS requires the development of a more scalable classifier capable of handling complex data.

Appendix A. Detailed Derivation of the Dual Problem to Problem (4)

For the linear kernel, the decision function takes the form of $f_k(\boldsymbol{x}) = \boldsymbol{w}_k^\top \boldsymbol{x} - \rho_k$ and hence the penalty $J(f_k) = \frac{1}{2}\|\boldsymbol{w}_k\|_2^2 - \rho_k$. With slackness variables $\eta_i := [1 - \boldsymbol{w}_k^\top \boldsymbol{x}_i + \rho_k]_+$ and $\xi_j := [1 + \boldsymbol{w}_k^\top \boldsymbol{x}_j - \rho_k]_+$, $C := (\lambda m)^{-1}$, problem (4) becomes

$$\begin{aligned} &\min_{\boldsymbol{w}_k, \rho_k, \{\eta_i\}, \{\xi_j\}} \ \frac{1}{2}\|\boldsymbol{w}_k\|_2^2 - \rho_k + C \sum_{j \in \mathcal{G}_{te}} \xi_j, \\ &\text{s.t. } \eta_i \geq 1 - \boldsymbol{w}_k^\top \boldsymbol{x}_i + \rho_k, \ \ \xi_j \geq 1 + \boldsymbol{w}_k^\top \boldsymbol{x}_j - \rho_k, \quad \sum_{i \in \mathcal{G}_k} \eta_i \leq n_k \gamma, \ \eta_i \geq 0, \ \xi_j \geq 0. \end{aligned} \tag{11}$$

By using the Lagrange Multiplier method, Problem (11) becomes

$$\begin{aligned} \mathcal{L} = &\frac{1}{2}\|\boldsymbol{w}_k\|_2^2 - \rho_k + C \sum_{j \in \mathcal{G}_{te}} \xi_j - \sum_{i \in \mathcal{G}_k} \alpha_i (\eta_i - 1 + \boldsymbol{w}_k^\top \boldsymbol{x}_i - \rho_k) - \sum_{j \in \mathcal{G}_{te}} \beta_j (\xi_j - 1 - \boldsymbol{w}_k^\top \boldsymbol{x}_j + \rho_k) \\ &+ \theta \left(\sum_{i \in \mathcal{G}_k} \eta_i - n_k \gamma \right) - \sum_{i \in \mathcal{G}_k} a_i \eta_i - \sum_{j \in \mathcal{G}_{te}} b_j \xi_j \\ = &\frac{1}{2}\|\boldsymbol{w}_k\|_2^2 + \sum_{j \in \mathcal{G}_{te}} (C - b_j - \beta_j) \xi_j + \sum_{i \in \mathcal{G}_k} (\theta - a_i - \alpha_i) \eta_i + \sum_{j \in \mathcal{G}_{te}} \beta_j \boldsymbol{w}_k^\top \boldsymbol{x}_j - \sum_{i \in \mathcal{G}_k} \alpha_i \boldsymbol{w}_k^\top \boldsymbol{x}_i \\ &+ \sum_{j \in \mathcal{G}_{te}} \beta_j + \sum_{i \in \mathcal{G}_k} \alpha_i + \rho_k \left(\sum_{i \in \mathcal{G}_k} \alpha_i - \sum_{j \in \mathcal{G}_{te}} \beta_j - 1 \right) - n_k \theta \gamma. \end{aligned}$$

Based on the stationary condition

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_k} = \boldsymbol{w}_k + \sum\limits_{j \in \mathcal{G}_{te}} \beta_j \boldsymbol{x}_j - \sum\limits_{i \in \mathcal{G}_k} \alpha_i \boldsymbol{x}_i = \boldsymbol{0} \\ \frac{\partial \mathcal{L}}{\partial \rho_k} = \sum\limits_{i \in \mathcal{G}_k} \alpha_i - \sum\limits_{j \in \mathcal{G}_{te}} \beta_j - 1 = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_j} = C - b_j - \beta_j = 0 \\ \frac{\partial \mathcal{L}}{\partial \eta_i} = \theta - a_i - \alpha_i = 0 \end{cases} \Longrightarrow \begin{cases} \boldsymbol{w}_k = \sum\limits_{i \in \mathcal{G}_k} \alpha_i \boldsymbol{x}_i - \sum\limits_{j \in \mathcal{G}_{te}} \beta_j \boldsymbol{x}_j \\ \sum\limits_{i \in \mathcal{G}_k} \alpha_i - \sum\limits_{j \in \mathcal{G}_{te}} \beta_j = 1 \\ C - b_j - \beta_j = 0 \\ \theta - a_i - \alpha_i = 0 \end{cases}.$$

and the KKT conditions, the dual problem of (11) is:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \theta} \ \frac{1}{2} \left(\boldsymbol{\alpha}^\top \mathbf{G}_1 \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \mathbf{G}_2 \boldsymbol{\beta} - 2 \boldsymbol{\alpha}^\top \mathbf{G}_3 \boldsymbol{\beta} \right) - \mathbf{1}_{n_k}^\top \boldsymbol{\alpha} - \mathbf{1}_m^\top \boldsymbol{\beta} + n_k \theta \gamma, \\ &\text{s.t. } \boldsymbol{0} \preceq \boldsymbol{\alpha} \preceq \theta \cdot \mathbf{1}_{n_k}, \ \boldsymbol{0} \preceq \boldsymbol{\beta} \preceq C \cdot \mathbf{1}_m, \ \mathbf{1}_{n_k}^\top \boldsymbol{\alpha} - \mathbf{1}_m^\top \boldsymbol{\beta} = 1, \ \theta \geq 0, \end{aligned}$$

where $\mathbf{G}_1[i, i'] := \boldsymbol{x}_i^\top \boldsymbol{x}_{i'}$, $\mathbf{G}_2[j, j'] := \boldsymbol{x}_j^\top \boldsymbol{x}_{j'}$, $\mathbf{G}_3[i, j] := \boldsymbol{x}_i^\top \boldsymbol{x}_j$, $\boldsymbol{\alpha} := (\ldots, \alpha_i, \ldots)^\top$, $\boldsymbol{\beta} := (\ldots, \beta_j, \ldots)^\top$, $i, i' \in \mathcal{G}_k, j, j' \in \mathcal{G}_{te}$. This is quadratic programming (QP) and can be solved with many off-the-shelf packages. For the non-linear kernel, e.g., the Gaussian kernel used in this article, we accordingly use $\mathbf{G}_1[i, i'] := K(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$, $\mathbf{G}_2[j, j'] := K(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$, $\mathbf{G}_3[i, j] := K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Appendix B. Details of the Algorithm and the Numerical Study

B.1 Outline of the Kernel Feature Selection Algorithm

Algorithm 1 Weighted Kernel Feature Selection

Initialization: $(\boldsymbol{\alpha}^{(0)}, \rho_k^{(0)}) = \mathbf{0}, \boldsymbol{d}^{(0)} = \mathbf{1}$
while $(\boldsymbol{\alpha}^{(t-1)}, \rho_k^{(t-1)}), \boldsymbol{d}^{(t-1)}$ not convergent **do**
  $(\boldsymbol{\alpha}^{(t)}, \rho_k^{(t)})$ = minimizer of (7) when fixing $\boldsymbol{d} = \boldsymbol{d}^{(t-1)}$
  while $\boldsymbol{d}^{\text{CV}}$ not convergent **do**
    $\boldsymbol{d}^{\text{CV}}$ = minimizer of (8) when fixing $(\boldsymbol{\alpha}, \rho_k) = (\boldsymbol{\alpha}^{(t)}, \rho_k^{(t)})$
    Compute a descent direction $\Delta \boldsymbol{d} = \boldsymbol{d}^{\text{CV}} - \boldsymbol{d}^{(t-1)}$. Conduct a line search to find ι such that $\boldsymbol{d}^{(t-1)} + \iota \Delta \boldsymbol{d}$ decreases the objective function in (7) when fixing $(\boldsymbol{\alpha}, \rho_k) = (\boldsymbol{\alpha}^{(t)}, \rho_k^{(t)})$
    $\boldsymbol{d}^{(t)} = \boldsymbol{d}^{(t-1)} + \iota \Delta \boldsymbol{d}$
  end while
  $\boldsymbol{d}^{(t)} = \boldsymbol{d}^{\text{CV}}$
end while

B.2 A Proposition about the Convergence of the Algorithm

Proposition 7 shows that we can obtain a local minimum of the objective in each iteration using Algorithm 1 under a certain loss function.

Proposition 7 (Proposition 1 in Allen (2013)) *If the convex loss function in (7) and (8) is continuously differentiable with respect to $(\boldsymbol{\alpha}, \rho_k)$, the kernel function is convex or concave and is continuously differentiable with respect to $\boldsymbol{d}$, then the solution obtained from Algorithm 1 converges to a local minimizer.*

The hinge loss is not continuously differentiable, as required by Proposition 7. One may substitute the hinge loss with a differentiable loss function such as the logistic loss, the squared hinge loss, or the Huberized squared hinge loss (Rosset and Zhu, 2007):

$$\ell(u) = \begin{cases} 1-u, & u \le 1-\delta \\ \frac{(1-u+\delta)^2}{4\delta}, & 1-\delta < u \le 1+\delta \\ 0, & u > 1+\delta \end{cases}$$

The parameter δ here is specified by the user. From Figure 1 we see that the Huberized squared hinge loss approximates the hinge loss with small δ. In practice, we directly work with the hinge loss and without line search since it empirically works well, and the optimization procedure is more efficient. Note that Problem (7) overall presents a broader challenge within the realm of non-convex optimization, and a local minimizer attained by Algorithm 1 might not be the global minimizer due to the complicated optimization landscape. To enhance the possibility of discovering an improved or even optimal solution, we suggest that one employ multiple initializations for training and subsequently select the model with the smallest objective function value.

B.3 Connection between Problem (6) and Problem (9)

Proposition 8 *Let $\check{f}$ parameterized by $(\check{\boldsymbol{d}}, \check{\boldsymbol{\alpha}}, \check{\rho})$ be a solution to Problem* (6) *for the given (λ_1, λ_2). Thus, $\check{f}$ is also a solution to Problem* (9) *when setting $s^2 = J(\check{f}(\check{\boldsymbol{d}} \circ \cdot)), s' = \|\check{\boldsymbol{d}}\|_1$.*

Proposition 8 shows that a solution to Problem (6) (or its re-parameterized variant, i.e., Problem (7)) can be obtained by solving Problem (9) under a certain set of values for (s^2, s'). This implies that the theoretical properties of Problem (6) may be obtained through studying the theoretical properties, e.g., Theorems 3 and 4, of Problem (9).

B.4 Empirical Metrics of Evaluation

In the context of set-valued classifiers, the prescribed class-specific accuracy drives both the OOD detection rate and the efficiency. Under a particular prescribed class-specific accuracy (such as $1-\gamma$), the sample accuracy for class k is

$$\frac{\sum_{j\in\mathcal{G}_{te}} \mathbb{1}\{Y_j = k \text{ and } Y_j \in \hat{\phi}(\boldsymbol{X}_j)\}}{\sum_{j\in\mathcal{G}_{te}} \mathbb{1}\{Y_j = k\}},\ k \in [K],$$

which is an unbiased estimate of the true accuracy, and should be close to $1-\gamma$. We also need to consider the OOD detection rate

$$\mathrm{Det}(1-\gamma) := \frac{\sum_{j\in\mathcal{G}_{te}} \mathbb{1}\{Y_j = \mathrm{OOD} \text{ and } |\hat{\phi}(\boldsymbol{X}_j)| = 0\}}{\sum_{j\in\mathcal{G}_{te}} \mathbb{1}\{Y_j = \mathrm{OOD}\}},$$

and the efficiency given non-OOD points

$$\mathrm{Eff}(1-\gamma) := 1 - \frac{1}{K-1}\left[\frac{\sum_{j\in\mathcal{G}_{te}} \mathbb{1}\{Y_j \neq \mathrm{OOD}\} \cdot |\hat{\phi}(\boldsymbol{X}_j)|}{\sum_{j\in\mathcal{G}_{te}} \mathbb{1}\{Y_j \neq \mathrm{OOD}\}} - 1\right]_+.$$

The closer to 1 the metric is, the better the classifier's corresponding performance is. Note that for the single-valued prediction paradigm, its efficiency is always 1.

To better understand the overall performance of set-valued classifiers, we propose two additional metrics called AUC-Det (the area under the curve of OOD detection rate v.s. accuracy) and AUC-Eff (the area under the curve of efficiency v.s. accuracy). By exactly aligning the class-specific accuracy on the test data at a grid of values from 0 to 1, we get the curve of Det and the curve of Eff, and hence the area under the curves (AUCs). Since a very low accuracy is rarely desirable in practice, we calculate the AUC in the neighborhood of $1-\gamma$, the prescribed accuracy, from $1-2\gamma$ to 1. More concretely, we have the below two metrics

$$\text{AUC-Det} := \frac{1}{2\gamma}\int_{1-2\gamma}^{1} \mathrm{Det}(t)dt, \quad \text{and} \quad \text{AUC-Eff} := \frac{1}{2\gamma}\int_{1-2\gamma}^{1} \mathrm{Eff}(t)dt.$$

The larger the AUC, the better the set-valued classifier on that corresponding performance. Note that the metric AUCs are still limited to some extent since they measure the overall OOD detection and efficiency by ignoring the specific requirement on accuracy.

B.5 Details of Tuning Parameter Selection

To choose the tuning parameters, the candidate hyper-parameters C_1, C_2 in GPSKFS is searched from grid $\{1, 2, 3\}$ and $10^\wedge\{\pm 2, \pm 1.5, \pm 1, \pm 0.5, 0\}$, respectively. The hyper-parameter C in GPS is searched from the grid $10^\wedge\{\pm 2, \pm 1.5, \pm 1, \pm 0.5, 0\}$. For the σ parameter in the Gaussian kernel $\exp(-\|\boldsymbol{d} \circ \boldsymbol{x} - \boldsymbol{d} \circ \boldsymbol{x}'\|^2/\sigma^2)$, we choose it from the $\{25, 37.5, 50, 62.5, 75\}$-th percentiles of all the pairwise Euclidean distances between the weighted training sample $\|\boldsymbol{d} \circ (\boldsymbol{x} - \boldsymbol{x}')\|_2$, where $\boldsymbol{d}$ is $\mathbf{1}$ in GPS, or is the currently estimated weight vector which can itself evolve in the iterations in GPSKFS.

For CDL, the bandwidth is searched from a grid $\{\hat{\sigma}_{(1)}, \hat{\sigma}_{(1)} + \frac{\hat{\sigma}_{(p)} - \hat{\sigma}_{(1)}}{p-1}, \ldots, \hat{\sigma}_{(p)}\} \times (\frac{4}{(p+2)n})^{1/(p+4)}$ based on Silverman's rule-of-thumb bandwidth estimator (Silverman, 2018), where $\hat{\sigma}_{(1)}$ and $\hat{\sigma}_{(p)}$ are the minimum and maximum standard deviation among all columns of data. We search parameter σ in the Gaussian kernel for both OCSVM and SVM in the same way as in GPS. For SVM, the parameter C is searched from the same grid as the one for C in GPS. The parameter ν in OCSVM is the upper bound of the overall proportion of points outside of any acceptance region, and hence is set as γ, which is its class-specific counterpart in our paper. For BCOPS-RF, the maximum depth of the tree is searched from $\{10, 20, \ldots, 90, 100\}$. Minimum samples to split an internal node, minimum samples at a leaf node, and the number of trees are searched from $\{2, 5, 10\}$, $\{2, 4, 6\}$, and $\{50, 150, 200\}$, respectively. All parameters are determined such that the prediction set size is minimized on the unlabeled data in the calibration set. We report involved metrics with their average and standard error after 200 replications for all the above methods.

In the Hybrid GPS methods described in Section 5.3, we adopt certain strategies to alleviate the computational burden. Specifically, we choose the 50-th percentile of pairwise distances on the training sample as the parameter σ, and we set the grid to be $10^\wedge\{\pm 1, 0\}$ for the parameter C. To ensure the robustness of our results, we conduct 10 replications and report the metrics for the three OSR methods as well as their corresponding Hybrid GPS methods.

B.6 Performances under Different Proportions of OOD Data

In this section, we study the stability of the proposed GPS on two real data sets for varied proportions (20%, 40%, 60%, and 80%) of OOD points in the test data. The class-specific accuracies are close to 99% in both real data sets and hence are omitted from Table 7. The efficiency is quite stable with respect to different OOD proportions. In contrast, the

OOD proportion	Zipcode		Phoneme	
	Detection Rate	Efficiency	Detection Rate	Efficiency
20%	33.8±0.928	95.1±0.183	96.8±0.331	84.3±0.22
40%	51.1±0.8	95.1±0.162	98.8±0.101	83.3±0.256
60%	62.1±0.68	93.9±0.195	99.7±0.033	82.3±0.317
80%	68.7±0.638	90.5±0.315	100±0.007	80.2±0.423

Table 7: Average performance metrics under increasing proportion of OOD points

detection rate can be improved if there are more points coming from the OOD class in the test data.

B.7 Simulations for Additional Label Shifts within Normal Classes

In the original Examples 1 and 2 presented in Section 5.1, the class priors for all K normal classes remained consistent between the training and test data, albeit experiencing a shift due to the presence of the OOD class in the test data. In this section, we introduce two additional simulation scenarios by allowing for label shifts even within the K normal classes. Specifically, the prior ratios for the normal classes in Table 8 are $0.226 : 0.258 : 0.255 : 0.262 \approx 1 : 1.14 : 1.13 : 1.16$ in the training data, but they are $0.130 : 0.261 : 0.130 : 0.217 \approx 1 : 2 : 1 : 1.5$ in the test data. In Table 9, the prior ratios for the normal classes in the training data are $0.272 : 0.345 : 0.383 \approx 1 : 1.26 : 1.4$, but they are $0.053 : 0.160 : 0.320 \approx 1 : 3 : 6$ in the test data. Overall, the proposed GPS methods demonstrate competitive performances, particularly in Example 2 (where many noise features are involved) shown in Table 9, even though there is a severe class imbalance issue.

		OCSVM	SVM	CDL	BCOPS-RF	GPS	GPSKFS
Accuracy	Class 1	94.2±0.192	95.2±0.135	94.5±0.193	94.8±0.181	94.4±0.115	94.6±0.131
	Class 2	95.5±0.155	95.8±0.134	95.3±0.179	96.1±0.144	95.3±0.148	95.2±0.143
	Class 3	94.6±0.17	96±0.172	95.7±0.141	95.4±0.175	95.6±0.165	95.8±0.162
	Class 4	96.2±0.091	95.6±0.124	95.8±0.11	96.4±0.136	95.2±0.153	95.1±0.164
Detection Rate		99.9±0.034	0±0	98.4±0.215	85.2±1.752	**100**±0.002	99.5±0.511
Efficiency		87±0.178	75.9±1.142	84.4±0.188	91.2±0.159	**92.8**±0.112	92.7±0.169
AUC-Det		**98.3**±0.066	2.1±0.45	96.2±0.275	76.5±0.654	94.9±0.337	91.5±0.612
AUC-Eff		84±0.189	72.9±0.426	82±0.172	88.4±0.101	**89.4**±0.106	88.4±0.185

Table 8: Example 1 with additional label shifts within normal classes

		OCSVM	SVM	CDL	BCOPS-RF	GPS	GPSKFS
Accuracy	Class 1	97.9±0.201	99.2±0.142	98.7±0.19	98.4±0.23	98.9±0.21	97.3±0.299
	Class 2	98.9±0.114	98.8±0.187	98.9±0.132	99.3±0.112	99±0.113	99.6±0.103
	Class 3	98.6±0.115	99.3±0.068	98.7±0.133	99.2±0.101	98.3±0.182	98.8±0.129
Detection Rate		98.9±0.088	0±0	32.3±0.586	20.3±1.135	99.6±0.1	**100**±0
Efficiency		52.9±0.236	53.3±0.182	22.3±0.312	37.8±1	59±0.234	**84.3**±0.266
AUC-Det		93.4±0.106	0±0	28.9±0.547	22.7±0.937	94.3±0.129	**95.1**±0
AUC-Eff		46.4±0.467	51.2±0.206	20.5±0.405	37.2±1.046	55.6±0.288	**80.6**±0.178

Table 9: Example 2 with additional label shifts within normal classes

Appendix C. Proofs of Theorems

Theorem 9 *Let $\tau_{k,\gamma}$ be the $\gamma\times100\%$ quantile of distribution $\frac{p_k(\boldsymbol{x})}{q(\boldsymbol{x})}$, where $q(\boldsymbol{x})$ is the density function of $\mathcal{Q}$. The Bayes optimal rule to Problem (2) is $f_k(\boldsymbol{x}) = \frac{p_k(\boldsymbol{x})}{q(\boldsymbol{x})} - \tau_{k,\gamma}$, and hence the optimal set-valued classifier to Problem (1) with $\gamma_k = \gamma$ is $\phi(\boldsymbol{x}) = \{k \in [K] : \frac{p_k(\boldsymbol{x})}{q(\boldsymbol{x})} \geq \tau_{k,\gamma}\}$.*

The above theorem can be derived from Lemma 2 in Sadinle et al. (2019). If we know the true density function $q(\boldsymbol{x})$ of target distribution $\mathcal{Q}$ and class-specific density functions $p_k(\boldsymbol{x}), k \in [K]$, we can set $\frac{p_k(\boldsymbol{x})}{q(\boldsymbol{x})}$ as the score function and $\gamma_{k,\gamma} \times 100\%$ quantile as the threshold to determine the acceptance region for class k.

Proof [Proof of Theorem 7] Denote the original objective function in Problem (7) as $\Psi(\boldsymbol{\alpha}, \rho_k, \boldsymbol{d})$. We first verify that $\Psi(\boldsymbol{\alpha}, \rho_k, \boldsymbol{d})$ is bounded below. Under the Gaussian kernel, the distance from the origin to the hyperplane in the feature space is $\frac{\rho_k}{\|g_k\|_{\mathcal{H}_{K_{\boldsymbol{d}}}}} \leq 1$, and hence $\rho_k \leq \|g_k\|_{\mathcal{H}_{K_{\boldsymbol{d}}}}$. Moreover, because the third and fourth terms in $\Psi(\boldsymbol{\alpha}, \rho_k, \boldsymbol{d})$ are non-negative, we have $\Psi(\boldsymbol{\alpha}, \rho_k, \boldsymbol{d}) \geq \frac{1}{2}\|g_k\|^2_{\mathcal{H}_{K_{\boldsymbol{d}}}} - \rho_k + 0 + 0 \geq \frac{1}{2}\|g_k\|^2_{\mathcal{H}_{K_{\boldsymbol{d}}}} - \|g_k\|_{\mathcal{H}_{K_{\boldsymbol{d}}}} \geq -\frac{1}{2}$.

Since the loss function is always bounded below; to prove that it converges to a stationary point, it suffices to prove the Algorithm 1 decreases $\Psi(\boldsymbol{\alpha}, \rho_k, \boldsymbol{d})$ in each step. It is easy to conclude $\Psi(\boldsymbol{\alpha}^{(t)}, \rho_k^{(t)}, \boldsymbol{d}^{(t-1)}) \leq \Psi(\boldsymbol{\alpha}^{(t-1)}, \rho_k^{(t-1)}, \boldsymbol{d}^{(t-1)})$ because updating for $(\boldsymbol{\alpha}, \rho_k)$ when fixing $\boldsymbol{d}^{(t-1)}$ is a convex optimization problem. Thus, it suffices to verify $\Psi(\boldsymbol{\alpha}^{(t-1)}, \rho_k^{(t-1)}, \boldsymbol{d}^{(t)}) \leq \Psi(\boldsymbol{\alpha}^{(t-1)}, \rho_k^{(t-1)}, \boldsymbol{d}^{(t-1)})$ when fixing $(\boldsymbol{\alpha}^{(t-1)}, \rho_k^{(t-1)})$ and updating for $\boldsymbol{d}$. We only focus on the case where $\frac{\partial \Psi}{\partial \boldsymbol{d}} \neq \mathbf{0}$ at $(\boldsymbol{\alpha}^{(t-1)}, \rho_k^{(t-1)}, \boldsymbol{d}^{(t-1)})$; otherwise we already arrive at a stationary point.

First of all, define

$$\mathbf{G}(\boldsymbol{d}) = [g_{i,j}(\boldsymbol{d})]_{i,j} := \begin{bmatrix} \mathbf{K}_{\boldsymbol{d}} & & & \\ & \boldsymbol{e}_1^\top \boldsymbol{d} & & \\ & & \ddots & \\ & & & \boldsymbol{e}_p^\top \boldsymbol{d} \end{bmatrix} \text{ and } \widetilde{\boldsymbol{\alpha}} := \begin{bmatrix} \frac{1}{\sqrt{2}}\boldsymbol{\alpha} \\ \sqrt{C_2}\mathbf{1}_p \end{bmatrix},$$

where $\mathbf{K}_{\boldsymbol{d}}$ is a $(n_k+m)\times(n_k+m)$ kernel matrix, $\boldsymbol{e}_l$ is a column vector with l-th element 1 but 0 elsewhere, and $\mathbf{1}_p$ is a p-dimensional column vector with all 1's. Given the above notations, the scalar $\mathbf{K}_{\boldsymbol{d}}[j,:]\boldsymbol{\alpha}$ for some j can be written as $\sum_i \tilde{\beta}_i g_{i,j}(\boldsymbol{d})$ for some $\tilde{\beta}_i$'s ($\widetilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}, \tilde{\beta}_i$ later will be replaced by the corresponding ones with a superscript of time $t-1$ when involved with the iterations). Then when fixing $(\boldsymbol{\alpha}^{(t-1)}, \rho_k^{(t-1)})$, we write the original objective function as a function of $\boldsymbol{d}$ only:

$$\Psi(\boldsymbol{d}) = \sum_i \sum_j \tilde{\alpha}_i^{(t-1)} \tilde{\alpha}_j^{(t-1)} g_{i,j}(\boldsymbol{d}) - \rho_k^{(t-1)} + C_1 \sum_j \ell(\rho_k^{(t-1)} - \sum_i \tilde{\beta}_i^{(t-1)} g_{i,j}(\boldsymbol{d})).$$

Since $C_1 > 0$ and $\ell(\cdot)$ is convex, the objective function is still convex with respect to $g_{i,j}(\boldsymbol{d})$. Without loss of generality and for the simplicity of notation, we can consider minimizing an objective function $\Psi(\boldsymbol{d}) = h\left(g(\boldsymbol{d})\right)$, where $h(\cdot)$ is a continuously differentiable and convex function, and $g(\boldsymbol{d})$ is continuously differentiable and convex or concave with respect to $\boldsymbol{d}$

(because of the assumption for kernel functions and the property of $\boldsymbol{e}_l^\top \boldsymbol{d}$). Moreover, denote $\widetilde{\Psi}_{\boldsymbol{d}^{(t-1)}}(\boldsymbol{d}) = h(g(\boldsymbol{d}^{(t-1)}) + \nabla g(\boldsymbol{d}^{(t-1)})^\top(\boldsymbol{d} - \boldsymbol{d}^{(t-1)}))$ as the approximated objective function where we linearize the kernel function at $\boldsymbol{d}^{(t-1)}$ to obtain a convex optimization Problem (8). For this sub-optimization problem, we always have $\widetilde{\Psi}_{\boldsymbol{d}^{(t-1)}}(\boldsymbol{d}^{(t)}) \leq \widetilde{\Psi}_{\boldsymbol{d}^{(t-1)}}(\boldsymbol{d}^{(t-1)})$.

Now we only need to verify $\Psi(\boldsymbol{d}^{(t)}) \leq \Psi(\boldsymbol{d}^{(t-1)})$ for those cases (Allen, 2013): (1) $h(\cdot)$ is deceasing or increasing when $g(\cdot)$ is convex, and (2) $h(\cdot)$ is deceasing or increasing when $g(\cdot)$ is concave.

When $g(\cdot)$ is convex, we have $g(\boldsymbol{d}) \geq g(\boldsymbol{d}^{(t-1)}) + \nabla g(\boldsymbol{d}^{(t-1)})^\top(\boldsymbol{d} - \boldsymbol{d}^{(t-1)})$. If $h(\cdot)$ is decreasing, then we have

$$\begin{aligned} & h(g(\boldsymbol{d})) \leq h(g(\boldsymbol{d}^{(t-1)}) + \nabla g(\boldsymbol{d}^{(t-1)})^\top(\boldsymbol{d} - \boldsymbol{d}^{(t-1)})) \\ \Rightarrow\ & h(g(\boldsymbol{d}^{(t)})) \leq h(g(\boldsymbol{d}^{(t-1)}) + \nabla g(\boldsymbol{d}^{(t-1)})^\top(\boldsymbol{d}^{(t)} - \boldsymbol{d}^{(t-1)})) \\ \Rightarrow\ & \Psi(\boldsymbol{d}^{(t)}) \leq \widetilde{\Psi}_{\boldsymbol{d}^{(t-1)}}(\boldsymbol{d}^{(t)}) \leq \widetilde{\Psi}_{\boldsymbol{d}^{(t-1)}}(\boldsymbol{d}^{(t-1)}) = \Psi(\boldsymbol{d}^{(t-1)}), \end{aligned}$$

which implies the original objective function decreases at this step although the solution $\boldsymbol{d}^{(t)}$ is obtained by solving Problem (8).

On the other hand, for any $0 \leq a \leq 1$, the convexity of g yields

$$g(a\boldsymbol{d} + (1-a)\boldsymbol{d}^{(t-1)}) \leq ag(\boldsymbol{d}) + (1-a)g(\boldsymbol{d}^{(t-1)}).$$

Note that h is convex. If $h(\cdot)$ is increasing, then we have

$$\begin{aligned} \Psi(a\boldsymbol{d} + (1-a)\boldsymbol{d}^{(t-1)}) = h(g(a\boldsymbol{d} + (1-a)\boldsymbol{d}^{(t-1)})) &\leq h(ag(\boldsymbol{d}) + (1-a)g(\boldsymbol{d}^{(t-1)})) \\ &\leq ah(g(\boldsymbol{d})) + (1-a)h(g(\boldsymbol{d}^{(t-1)})) \\ &= a\Psi(\boldsymbol{d}) + (1-a)\Psi(\boldsymbol{d}^{(t-1)}), \end{aligned}$$

which implies $\Psi(\boldsymbol{d})$ is convex at the neighborhood of $\boldsymbol{d}^{(t-1)}$, say $N(\boldsymbol{d}^{(t-1)})$.

Since $\Psi(\boldsymbol{d})$ is locally convex in $N(\boldsymbol{d}^{(t-1)})$, we can decrease it by taking a proper direction. So we take $\Delta\boldsymbol{d} = \boldsymbol{d}^{\text{CV}} - \boldsymbol{d}^{(t-1)}$ as a descent direction with a proper step size ι by the line search to decrease $\Psi(\boldsymbol{d})$, where ι is to make sure $\Psi(\boldsymbol{d})$ is decreased in the feasible region.

For the other two cases where $g(\cdot)$ is concave, similarly, we can verify $\Psi(\boldsymbol{d})$ also decreases when fixing $(\boldsymbol{\alpha}^{(t-1)}, \rho_k^{(t-1)})$. Therefore, the solution obtained from the algorithm converges to a local minimizer. ■

Proof [Proof of Proposition 8] Without loss of generality, we omit subscripts in f, ρ and their estimations.

Define $L(\boldsymbol{d}, \boldsymbol{\alpha}, \rho) = \frac{1}{m}\sum_{j \in \mathcal{G}_{te}} \ell\big(-f(\boldsymbol{d} \circ \boldsymbol{x}_j)\big)$ as the objective function when f is parameterized by $(\boldsymbol{d}, \boldsymbol{\alpha}, \rho)$. Let $\check{f}$ parameterized by $(\check{\boldsymbol{d}}, \check{\boldsymbol{\alpha}}, \check{\rho})$ be a solution to Problem (6) for the given (λ_1, λ_2). Let $\hat{f}$ parameterized by $(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}}, \hat{\rho})$ be a solution to Problem (9) when setting $s^2 = J(\check{f}(\check{\boldsymbol{d}} \circ \cdot))$ and $s' = \|\check{\boldsymbol{d}}\|_1$. Based on the minimum of Problem (6) and (9), we have

$$L(\check{\boldsymbol{d}}, \check{\boldsymbol{\alpha}}, \check{\rho}) + \lambda_1 J(\check{f}(\check{\boldsymbol{d}} \circ \cdot)) + \lambda_2\|\check{\boldsymbol{d}}\|_1 \leq L(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}}, \hat{\rho}) + \lambda_1 J(\hat{f}(\hat{\boldsymbol{d}} \circ \cdot)) + \lambda_2\|\hat{\boldsymbol{d}}\|_1 \quad (12)$$

$$L(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}}, \hat{\rho}) \leq L(\check{\boldsymbol{d}}, \check{\boldsymbol{\alpha}}, \check{\rho}) \quad (13)$$

$$\begin{aligned} (12) + (13) \quad \Longrightarrow \quad & \lambda_1 J(\check{f}(\check{\boldsymbol{d}} \circ \cdot)) + \lambda_2\|\check{\boldsymbol{d}}\|_1 \leq \lambda_1 J(\hat{f}(\hat{\boldsymbol{d}} \circ \cdot)) + \lambda_2\|\hat{\boldsymbol{d}}\|_1 \\ \text{by the constraints for } \hat{f}, \hat{\boldsymbol{d}} \quad & \leq \lambda_1 s^2 + \lambda_2 s', \end{aligned} \quad (14)$$

which implies all the constraints in Problem (9) must be active, i.e., $J(\hat{f}) = s^2, \|\hat{\boldsymbol{d}}\|_1 = s'$. This further implies $L(\check{\boldsymbol{d}}, \check{\boldsymbol{\alpha}}, \check{\rho}) = L(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\alpha}}, \hat{\rho})$ and hence the a solution to Problem (6) is a solution to (9) when setting $s^2 = J(\check{f}(\check{\boldsymbol{d}} \circ \cdot)), s' = \|\check{\boldsymbol{d}}\|_1$. ■

We will prove Theorem 3. As mentioned in Section 4, we abuse the notation slightly for the decision function by letting $f(\boldsymbol{x}) = f_1(\boldsymbol{d} \circ \boldsymbol{x})$, omitting the subscript 1 and weight $\boldsymbol{d}$. Before that, we need to introduce the below lemma regarding the boundedness of ρ and g.

Lemma 10 *Let $f(\cdot) = g(\cdot) - \rho \in \mathcal{F}_{s,s'}(s, s' \geq 0)$, where g belongs to the Gaussian kernel RKHS, $\mathcal{H}_{K_{\boldsymbol{d}}}$. We have, $\rho \leq \sqrt{2}s + 2$ and $\|g\|_{\mathcal{H}_{K_{\boldsymbol{d}}}} \leq \sqrt{2}s + 2$.*

Proof Under the Gaussian kernel, the distance from the hyper-plane to the origin is $\frac{\rho}{\|g\|_{\mathcal{H}_{K_{\boldsymbol{d}}}}} \leq 1$. Together with the hypothesis space complexity $\frac{1}{2}\|g\|^2_{\mathcal{H}_{K_{\boldsymbol{d}}}} - \rho \leq s^2$, we have $\rho \leq \sqrt{2s^2 + 1} + 1 \leq \sqrt{2}s + 2$ and hence $\|g\|_{\mathcal{H}_{K_{\boldsymbol{d}}}} \leq \sqrt{2(s^2 + \sqrt{2}s + 2)} \leq \sqrt{2}s + 2$. ■

Proof [Proof of Theorem 3] For simplicity, denote $\mathbb{E}_Q\left[\ell(f(\boldsymbol{X})) \mid Y = 1\right] = E_+\left[\ell(f(\boldsymbol{X}))\right]$ and hence $\mathbb{P}_Q[f(\boldsymbol{X}) < 0 \mid Y = 1] \leq E_+[\ell(f(\boldsymbol{X}))]$. Define $\psi(S) = \sup_{f \in \mathcal{F}^+_{s,s'}(\gamma)} E_+[\ell(f(\boldsymbol{X}))] - \frac{1}{n_1}\sum_{\boldsymbol{x}_i \in S} \ell(f(\boldsymbol{x}_i))$ and let S' be another sample from $\mathbb{P}_Q[\cdot \mid Y = 1]$ but only different from S on one observation $(\boldsymbol{x}', 1)$. Thus we have

$$
\begin{aligned}
&\left|\psi(S) - \psi(S')\right| \\
&= \left|\left(\sup_{f \in \mathcal{F}^+_{s,s'}(\gamma)} E_+[\ell(f(\boldsymbol{X}))] - \frac{1}{n_1}\sum_{\boldsymbol{x}_i \in S} \ell(f(\boldsymbol{x}_i))\right) - \left(\sup_{f \in \mathcal{F}^+_{s,s'}(\gamma)} E_+[\ell(f(\boldsymbol{X}))] - \frac{1}{n_1}\sum_{\boldsymbol{x}'_i \in S'} \ell(f(\boldsymbol{x}'_i))\right)\right| \\
&\leq \frac{1}{n_1} \sup_{f \in \mathcal{F}^+_{s,s'}(\gamma)} \left|\ell(f(\boldsymbol{x})) - \ell(f(\boldsymbol{x}'))\right| \\
&\leq \frac{c}{n_1} \sup_{\mathcal{F}^+_{s,s'}(\gamma)} \left|g(\boldsymbol{x}) - g(\boldsymbol{x}')\right| \leq \frac{2c}{n_1} \sup_{\mathcal{F}^+_{s,s'}(\gamma)} |\langle g, K_{\boldsymbol{d}}(\boldsymbol{x}, \cdot)\rangle| \leq \frac{2(\sqrt{2}s + 2)c\kappa}{n_1}.
\end{aligned}
$$

Together with McDiarmid inequality, with probability $1 - \zeta$, we have

$$
\psi(S) \leq \underset{S}{E_+}[\psi(S)] + (\sqrt{2}s + 2)c\kappa\sqrt{\frac{2\log\frac{1}{\zeta}}{n_1}},
$$

and hence

$$
E_+[\ell(f(\boldsymbol{X}))] \leq \frac{1}{n_1}\sum_{i=1}^{n_1} \ell(f(\boldsymbol{x}_i)) + \underset{S}{E_+}[\psi(S)] + (\sqrt{2}s + 2)c\kappa\sqrt{\frac{2\log\frac{1}{\zeta}}{n_1}},
$$

where

$$
\begin{aligned}
\underset{S}{E_+}[\psi(S)] &= \underset{S}{E_+}\left[\sup_{f\in\mathcal{F}^+_{s,s'}(\gamma)} E_+[\ell(f(\boldsymbol{X}))] - \frac{1}{n_1}\sum_{\boldsymbol{x}_i\in S}\ell(f(\boldsymbol{x}_i))\right] \\
&= \underset{S}{E_+}\left[\sup_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \underset{S'}{E_+}\left[\frac{1}{n_1}\sum_{\boldsymbol{x}'_i\in S'}\ell(f(\boldsymbol{x}'_i))\right] - \frac{1}{n_1}\sum_{\boldsymbol{x}_i\in S}\ell(f(\boldsymbol{x}_i))\right] \\
&\le \underset{S}{E_+}\underset{S'}{E_+}\left[\sup_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \frac{1}{n_1}\sum_{\boldsymbol{x}'_i\in S'}\ell(f(\boldsymbol{x}'_i)) - \frac{1}{n_1}\sum_{\boldsymbol{x}_i\in S}\ell(f(\boldsymbol{x}_i))\right] \\
&= \underset{S}{E_+}\underset{S'}{E_+}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\sup_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \frac{1}{n_1}\sum_{i=1}^{n_1}\sigma_i\left[\ell(f(\boldsymbol{x}'_i)) - \ell(f(\boldsymbol{x}_i))\right] \\
&\le \underset{S}{E_+}\underset{S'}{E_+}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\sup_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \frac{1}{n_1}\sum_{i=1}^{n_1}\sigma_i\ell(f(\boldsymbol{x}'_i)) + \underset{S}{E_+}\underset{S'}{E_+}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\sup_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \frac{1}{n_1}\sum_{i=1}^{n_1}-\sigma_i\ell(f(\boldsymbol{x}'_i)) \\
&= 2\,\underset{S}{E_+}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\sup_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \frac{1}{n_1}\sum_{\boldsymbol{x}_i\in S}\sigma_i\ell(f(\boldsymbol{x}_i)) \\
&= 2\mathfrak{R}_{n_1}(\ell\circ\mathcal{F}^+_{s,s'}(\gamma)),
\end{aligned}
$$

with $\mathfrak{R}_{n_1}$ defined as the Rademacher complexity. Applied again with McDiarmid inequality, with probability $1-\zeta$, we have

$$
\mathfrak{R}_{n_1}(\ell\circ\mathcal{F}^+_{s,s'}(\gamma)) \le \widehat{\mathfrak{R}}_{n_1}(\ell\circ\mathcal{F}^+_{s,s'}(\gamma)) + (\sqrt{2}s+2)c\kappa\sqrt{\frac{2\log\frac{1}{\zeta}}{n_1}}.
$$

According to Talagrand's lemma,

$$
\widehat{\mathfrak{R}}_{n_1}(\ell\circ\mathcal{F}^+_{s,s'}(\gamma)) \le c\cdot\widehat{\mathfrak{R}}_{n_1}(\mathcal{F}^+_{s,s'}(\gamma)).
$$

Additionally, by the definition of empirical Rademacher complexity, we have

$$
\begin{aligned}
\widehat{\mathfrak{R}}_{n_1}(\mathcal{F}^+_{s,s'}(\gamma)) &= \underset{\boldsymbol{\sigma}}{\mathbb{E}}\sup_{f\in\mathcal{F}^+_{s,s'}(\gamma)} \frac{1}{n_1}\sum_{\boldsymbol{x}_i\in S}\sigma_i f(\boldsymbol{x}_i) \\
&\le \underset{\boldsymbol{\sigma}}{\mathbb{E}}\sup_{\mathcal{F}^+_{s,s'}(\gamma)} \frac{1}{n_1}\sum_{\boldsymbol{x}_i\in S}\sigma_i\langle g, K_{\boldsymbol{d}}(\boldsymbol{x}_i,\cdot)\rangle + \underset{\boldsymbol{\sigma}}{\mathbb{E}}\sup_{\mathcal{F}^+_{s,s'}(\gamma)} \frac{1}{n_1}\sum_{\boldsymbol{x}_i\in S}-\sigma_i\rho \\
&\le \frac{(\sqrt{2}s+2)}{n_1}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left|\sum_{\boldsymbol{x}_i\in S}\sigma_i\sqrt{K_{\boldsymbol{d}}(\boldsymbol{x}_i,\boldsymbol{x}_i)}\right| \\
&\le \frac{(\sqrt{2}s+2)}{n_1}\left[\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left(\sum_{\boldsymbol{x}_i\in S}\sigma_i\sqrt{K_{\boldsymbol{d}}(\boldsymbol{x}_i,\boldsymbol{x}_i)}\right)^2\right]^{\frac{1}{2}} \\
&\le \frac{(\sqrt{2}s+2)}{n_1}\left(n_1\kappa^2\right)^{\frac{1}{2}} = \frac{(\sqrt{2}s+2)\kappa}{\sqrt{n_1}},
\end{aligned}
$$

combining above results, with probability $1-2\zeta$, we have

$$\mathbb{E}_Q\left[\ell(f(\boldsymbol{X}))\mid Y=1\right] \le \frac{1}{n_1}\sum_{i=1}^{n_1}\ell(f(\boldsymbol{x}_i)) + \underbrace{\frac{2(\sqrt{2}s+2)c\kappa}{\sqrt{n_1}} + 3(\sqrt{2}s+2)c\kappa\sqrt{\frac{2\log\frac{1}{\zeta}}{n_1}}}_{r_{n_1}(\zeta,s,s')}. \qquad \blacksquare$$

The shift-invariant property of the Gaussian kernel leads the term $r_{n_1}(\zeta,s,s')$ in Theorem 3 to overlook the dependence between the convergence rate and the parameter s'. To explicitly show the effect of s' on the convergence rate, we resort to the covering number to derive a new bound as shown in the below theorem.

Theorem 11 *Assume the input $\boldsymbol{x}\in\mathbb{R}^p$ is bounded, i.e., $\|\boldsymbol{x}\|_2\le c_0$, and the loss function ℓ has a sub-derivative bounded by c. For the Gaussian RKHS (i.e., $\kappa=1$) and the same ζ,s,s' given in Theorem 3, the term $r_{n_1}(\zeta,s,s')$ in (10) can be replaced by*

$$\begin{aligned}&3(\sqrt{2}s+2)c\sqrt{\frac{2\log\frac{2}{\zeta}}{n_1}} + \frac{8\sqrt{2}s+16}{epn_1}\\ &+192\left((s+\sqrt{2})s'c_0\sqrt{\log(2p^2)}+2p+4p\sqrt{\sqrt{2}s+2}\right)\frac{\log(4epn_1)}{\sqrt{n_1}}.\end{aligned}$$

Proof Let $L_2(\mathbb{P}_n) := \sqrt{\frac{1}{n}\sum_{i=1}^n(f(\boldsymbol{x}_i)-f'(\boldsymbol{x}_i))^2}$ be a metric to measure the distance between two functions f,f' on the given data set with sample size n.

Define the hypothesis class $\mathcal{F}_1=\{\boldsymbol{x}\mapsto\mathbf{D}\boldsymbol{x}:\mathbf{D}=\mathrm{Diag}(\boldsymbol{d})\in\mathbb{R}^{p\times p},\|\boldsymbol{d}\|_1\le s',\mathbf{0}\preceq\boldsymbol{d}\preceq\mathbf{1}\}$, under which we have $\mathbf{D}\boldsymbol{x}=\boldsymbol{d}\circ\boldsymbol{x}$. Denote $\mathcal{N}(\varepsilon_1,\mathcal{F}_1,L_2(\mathbb{P}_n))$ as the covering number for the hypothesis class $\mathcal{F}_1$ under the metric $L_2(\mathbb{P}_n)$, where ε_1 is the radius of each ball in the covering. By Theorem 5.18 in Ma (2022), we have the metric entropy

$$\log\mathcal{N}(\varepsilon_1,\mathcal{F}_1,L_2(\mathbb{P}_n)) \le \frac{(s')^2c_0^2}{\varepsilon_1^2}\log(2p^2).$$

Define another hypothesis class $\mathcal{F}_2=\{f:f(\boldsymbol{x})=g(\boldsymbol{x})-\rho,\ g\in\mathcal{H}_K,J(f)\le s^2\}$, where $\mathcal{H}_K$ is the Gaussian RKHS as mentioned in Section 3.2. By Theorem 20 in Ying and Zhou (2007), we have the metric entropy

$$\log\mathcal{N}(\varepsilon_2,\mathcal{F}_2,L_2(\mathbb{P}_n)) \le p\log 2+\left(p+\left(\frac{16p}{\varepsilon_2}+2\right)p\log\frac{2epn}{\varepsilon_2}\right)\log\frac{4p^2n}{\varepsilon_2} \le \frac{128p^2}{\varepsilon_2}\log^2\frac{2epn}{\varepsilon_2}.$$

For the composite hypothesis class $\mathcal{F}_2\circ\mathcal{F}_1$, by Lemma 5.23 in Ma (2022), we have the corresponding metric entropy

$$\log\mathcal{N}(\varepsilon_2+c_2\varepsilon_1,\mathcal{F}_2\circ\mathcal{F}_1,L_2(\mathbb{P}_n)) \le \log\mathcal{N}(\varepsilon_2,\mathcal{F}_2,L_2(\mathbb{P}_n))+\log\mathcal{N}(\varepsilon_1,\mathcal{F}_1,L_2(\mathbb{P}_n)),$$

where c_2 is the Lipschitz constant of $f \in \mathcal{F}_2$. Particularly, for the Gaussian RKHS,

$$\begin{aligned}
|f(\boldsymbol{x}) - f(\boldsymbol{x}')| &= |g(\boldsymbol{x}) - g(\boldsymbol{x}')| \\
&= |\langle g, K(\boldsymbol{x}, \cdot)\rangle - \langle g, K(\boldsymbol{x}', \cdot)\rangle| \\
&\leq \|g\|_{\mathcal{H}_K} \cdot \|K(\boldsymbol{x}, \cdot) - K(\boldsymbol{x}', \cdot)\|_{\mathcal{H}_K} \\
&\leq (\sqrt{2}s + 2) \cdot 2(1 - K(\boldsymbol{x}, \boldsymbol{x}')) \\
&\leq (2\sqrt{2}s + 4) \frac{d \exp(-\frac{z^2}{\sigma^2})}{dz}\Big|_{z=\tilde{c}} \cdot \|\boldsymbol{x} - \boldsymbol{x}'\|_2, \text{ where } \tilde{c} \in (0, \|\boldsymbol{x} - \boldsymbol{x}'\|_2) \\
&\leq (4s + 4\sqrt{2}) \cdot \|\boldsymbol{x} - \boldsymbol{x}'\|_2
\end{aligned}$$

implies $c_2 = 4s + 4\sqrt{2}$. By defining $\varepsilon_{12} := \varepsilon_2 + c_2\varepsilon_1$ and setting $\varepsilon_2 = \frac{\varepsilon_{12}}{2}, \varepsilon_1 = \frac{\varepsilon_{12}}{2c_2}$, we have

$$\log \mathcal{N}(\varepsilon_{12}, \mathcal{F}_2 \circ \mathcal{F}_1, L_2(\mathbb{P}_n)) \leq \log \mathcal{N}\left(\frac{\varepsilon_{12}}{2}, \mathcal{F}_2, L_2(\mathbb{P}_n)\right) + \log \mathcal{N}\left(\frac{\varepsilon_{12}}{2c_2}, \mathcal{F}_1, L_2(\mathbb{P}_n)\right). \quad (15)$$

By Localized Dudley's Theorem, together with (15), for any $0 < \epsilon_0 < 1$, we have

$$\begin{aligned}
&\widehat{\mathfrak{R}}_{n_1}(\mathcal{F}_2 \circ \mathcal{F}_1) \\
&\leq 4\epsilon_0 + 12 \int_{\epsilon_0}^{\sqrt{2}s+2} \sqrt{\frac{\log \mathcal{N}(\varepsilon_{12}, \mathcal{F}_2 \circ \mathcal{F}_1, L_2(\mathbb{P}_{n_1}))}{n_1}} d_{\varepsilon_{12}} \\
&\leq 4\epsilon_0 + 12 \int_{\epsilon_0}^{\sqrt{2}s+2} \sqrt{\frac{\log \mathcal{N}\big(\varepsilon_{12}/2, \mathcal{F}_2, L_2(\mathbb{P}_{n_1})\big)}{n_1}} d_{\varepsilon_{12}} \\
&\qquad + 12 \int_{\epsilon_0}^{\sqrt{2}s+2} \sqrt{\frac{\log \mathcal{N}\big(\varepsilon_{12}/(2c_2), \mathcal{F}_1, L_2(\mathbb{P}_{n_1})\big)}{n_1}} d_{\varepsilon_{12}} \\
&\leq 4\epsilon_0 + \big(96(s + \sqrt{2})s' c_0 \sqrt{\log(2p^2)} + 192p\big) \frac{\log((\sqrt{2}s + 2)/\epsilon_0)}{\sqrt{n_1}} + 384p\sqrt{\sqrt{2}s + 2} \frac{\log(4epn_1)}{\sqrt{n_1}} \\
&= \frac{4\sqrt{2}s + 8}{epn_1} + \left(96(s + \sqrt{2})s' c_0 \sqrt{\log(2p^2)} + 192p + 384p\sqrt{\sqrt{2}s + 2}\right) \frac{\log(4epn_1)}{\sqrt{n_1}},
\end{aligned}$$

where the last equality holds due to the choice $\epsilon_0 := \frac{\sqrt{2}s+2}{4epn_1}$. Together with the Proof of Theorem 3 and the fact of $\widehat{\mathfrak{R}}_{n_1}(\mathcal{F}^+_{s,s'}(\gamma)) \leq \widehat{\mathfrak{R}}_{n_1}(\mathcal{F}_2 \circ \mathcal{F}_1)$, with probability at least $1 - \zeta$, we have

$$\begin{aligned}
\mathbb{E}_Q\left[\ell(f(\boldsymbol{X})) \mid Y = 1\right] \leq \frac{1}{n_1} \sum_{i=1}^{n_1} \ell(f(\boldsymbol{x}_i)) + 3(\sqrt{2}s + 2)c\sqrt{\frac{2\log\frac{2}{\zeta}}{n_1}} + \frac{8\sqrt{2}s + 16}{epn_1} \\
+ 192\left((s + \sqrt{2})s' c_0 \sqrt{\log(2p^2)} + 2p + 4p\sqrt{\sqrt{2}s + 2}\right) \frac{\log(4epn_1)}{\sqrt{n_1}}.
\end{aligned}$$

■

Before proving Theorem 4, we prove the below proposition.

Proposition 12 *Let $\upsilon(\gamma) = \inf_{f \in \mathcal{F}^+_{s,s'}(\gamma)} \mathcal{R}_\ell(f)$, then υ is a non-increasing convex function on $[0,1]$.*

Proof υ is non-increasing because of the definition of infimum. We now focus on the convexity. $\mathcal{F}^+_{s,s'}(\gamma)$ is compact due to continuouity and boundedness of ℓ and f, therefore, there exits a $f^\gamma \in \mathcal{F}^+_{s,s'}(\gamma)$ such that $\upsilon(\gamma) = \mathcal{R}_\ell(f^\gamma)$.

Let $\upsilon(\gamma_1) = \mathcal{R}_\ell(f^{\gamma_1}), \upsilon(\gamma_2) = \mathcal{R}_\ell(f^{\gamma_2})$ and define $\gamma_\theta = \theta\gamma_1 + (1-\theta)\gamma_2, f_\theta = \theta f^{\gamma_1} + (1-\theta)f^{\gamma_2}$ for any $\theta \in (0,1)$, since ℓ is convex, then we have

$$E_+[\ell \circ f_\theta] \leq \theta E_+[\ell \circ f^{\gamma_1}] + (1-\theta)E_+[\ell \circ f^{\gamma_2}] \leq \theta\gamma_1 + (1-\theta)\gamma_2 = \gamma_\theta$$

and hence

$$\begin{aligned}\upsilon(\theta\gamma_1 + (1-\theta)\gamma_2) &= \upsilon(\gamma_\theta) \\ &\leq \mathcal{R}_\ell(f_\theta) \\ &\leq \theta\mathcal{R}_\ell(f^{\gamma_1}) + (1-\theta)\mathcal{R}_\ell(f^{\gamma_2}) \\ &= \theta\upsilon(\gamma_1) + (1-\theta)\upsilon(\gamma_2).\end{aligned}$$

Therefore, υ is convex. ■

Proof [Proof of Theorem 4] For any $0 \leq \gamma - \varepsilon_0 < \gamma - \varepsilon < 1$, based on the properties of $\upsilon(\cdot)$ we have

$$\begin{aligned}\frac{\upsilon(\gamma-\varepsilon_0) - \upsilon(\gamma-\varepsilon)}{\varepsilon - \varepsilon_0} &\leq \frac{\upsilon(\gamma-\varepsilon) - \upsilon(\gamma)}{-\varepsilon} \\ \upsilon(\gamma-\varepsilon) - \upsilon(\gamma) &\leq \frac{\varepsilon}{\varepsilon_0 - \varepsilon}\left(\upsilon(\gamma-\varepsilon_0) - \upsilon(\gamma-\varepsilon)\right).\end{aligned}$$

Now take $\varepsilon_0 = \gamma$, we obtain

$$\upsilon(\gamma-\varepsilon) - \upsilon(\gamma) \leq \frac{\varepsilon}{\gamma-\varepsilon}\left(2 + \frac{\delta}{2}\right) \tag{16}$$

because we have $f \equiv 1 + \frac{\delta}{2}$ satisfy the ℓ-type I error and then $\mathcal{R}_\ell \equiv 2 + \frac{\delta}{2}$.

Let's first define $A = \left\{\mathbb{E}_Q\left[\ell(f(\boldsymbol{X})) \mid Y = 1\right] - \frac{1}{n_1}\sum_{i=1}^{n_1}\ell(f(\boldsymbol{x}_i)) < \varepsilon\right\}$, where

$$\varepsilon = \frac{(\sqrt{2}s+2)c\kappa\left(2 + 3\sqrt{2\log\frac{2}{\zeta}}\right)}{\sqrt{n_1}}.$$

Based on the proof for Theorem 3, we have $\mathbb{P}[A] \geq 1-\zeta$.

$$\begin{aligned}
\mathcal{R}_\ell(\hat{f}) - \inf_{f\in\mathcal{F}^+_{s,s'}(\gamma)} R_\ell(f) &= \mathcal{R}_\ell(\hat{f}) - \inf_{f\in\widehat{\mathcal{F}}^+_{s,s'}(\gamma-\varepsilon)} \mathcal{R}_\ell(f) \\
&\quad + \inf_{f\in\widehat{\mathcal{F}}^+_{s,s'}(\gamma-\varepsilon)} \mathcal{R}_\ell(f) - \inf_{f\in\mathcal{F}^+_{s,s'}(\gamma-2\varepsilon)} \mathcal{R}_\ell(f) \\
&\quad + \inf_{f\in\mathcal{F}^+_{s,s'}(\gamma-2\varepsilon)} \mathcal{R}_\ell(f) - \inf_{f\in\mathcal{F}^+_{s,s'}(\gamma)} R_\ell(f) \\
&\leq 2 \sup_{f\in\widehat{\mathcal{F}}^+_{s,s'}(\gamma)} \left| \mathcal{R}_\ell(f) - \frac{1}{m}\sum_{j=1}^m \ell(-f(\boldsymbol{x}_j)) \right| \\
&\quad + 0 \\
&\quad + \frac{2\varepsilon}{\gamma-2\varepsilon}\left(2+\frac{\delta}{2}\right) \quad \text{by Inequality (16)}
\end{aligned} \tag{17}$$

Note that empirical minimizer $\hat{f} \in \widehat{\mathcal{F}}^+_{s,s'}(\gamma-\varepsilon) \subset \mathcal{F}^+_{s,s'}(\gamma)$. Define $\bar{f} := \operatorname{arginf}_{f\in\widehat{\mathcal{F}}^+_{s,s'}(\gamma-\varepsilon)} \mathcal{R}_\ell(f)$, then the first part in the first line on the right of the Inequality (17) bounded by twice of supremum is due to

$$\begin{aligned}
\mathcal{R}_\ell(\hat{f}) - \mathcal{R}_\ell(\bar{f}) &= \mathcal{R}_\ell(\hat{f}) - \frac{1}{m}\sum_{j=1}^m \ell(-\hat{f}(\boldsymbol{x}_j)) + \frac{1}{m}\sum_{j=1}^m \ell(-\hat{f}(\boldsymbol{x}_j)) - \frac{1}{m}\sum_{j=1}^m \ell(-\bar{f}(\boldsymbol{x}_j)) \\
&\quad + \frac{1}{m}\sum_{j=1}^m \ell(-\bar{f}(\boldsymbol{x}_j)) - \mathcal{R}_\ell(\bar{f}) \\
&\leq \mathcal{R}_\ell(\hat{f}) - \frac{1}{m}\sum_{j=1}^m \ell(-\hat{f}(\boldsymbol{x}_j)) + 0 - \left[\mathcal{R}_\ell(\bar{f}) - \frac{1}{m}\sum_{j=1}^m \ell(-\bar{f}(\boldsymbol{x}_j)) \right] \\
&\leq 2 \sup_{f\in\widehat{\mathcal{F}}^+_{s,s'}(\gamma-\varepsilon)} \left| \mathcal{R}_\ell(f) - \frac{1}{m}\sum_{j=1}^m \ell(-f(\boldsymbol{x}_j)) \right| \\
&\leq 2 \sup_{f\in\widehat{\mathcal{F}}^+_{s,s'}(\gamma)} \left| \mathcal{R}_\ell(f) - \frac{1}{m}\sum_{j=1}^m \ell(-f(\boldsymbol{x}_j)) \right|.
\end{aligned}$$

The second line on the right of the Inequality (17) can be bounded by 0 since $\mathcal{F}^+_{s,s'}(\gamma-2\varepsilon) \subset \widehat{\mathcal{F}}^+_{s,s'}(\gamma-\varepsilon)$ with probability $1-\delta$ based on the statement (1) in Theorem 4.

Therefore, with probability $1-2\zeta$,

$$\mathcal{R}_\ell(\hat{f}) - \inf_{f\in\mathcal{F}^+_{s,s'}(\gamma)} R_\ell(f) \leq \frac{2(\sqrt{2}s+2)c\kappa(2+3\sqrt{2\log\frac{2}{\zeta}})}{\sqrt{m}} + \frac{(4+\delta)\varepsilon}{\gamma-2\varepsilon}. \qquad \blacksquare$$

Proof [Proof of Theorem 6] We mainly follow the proof of the Theorem in Chen et al. (2018). Suppose $\|\boldsymbol{x}\|_\infty = \kappa_0 < \infty$, loss function ℓ is differentiable with Lipschitz constant c.

Let $\hat{f}$ be the empirical risk minimizer. Based on Corollary 4.36 (RKHSs of differentiable kernels) in Steinwart and Christmann (2008) we have $\frac{\partial f(\boldsymbol{x})}{\partial d_t} \leq \frac{\sqrt{2}\kappa_0}{\sigma}$ and hence $\frac{\partial \ell(f)}{\partial d_t}$ is still Lipschitz with Lipschitz constant $c' = \frac{\sqrt{2}\kappa_0 c}{\sigma}$. Then similarly to proofs of previous Theorem 1 and 2, with probability at least $1 - 3\zeta$, we have

$$\left| \frac{\partial}{\partial d_t} \left\{ \mathbb{E}_{\mathcal{Q}}[\ell(-\hat{f}(\boldsymbol{X}))] - \frac{1}{m} \sum_{j=1}^{m} \ell(-\hat{f}(\boldsymbol{x}_j)) \right\} \right| \leq \frac{(\sqrt{2}s+2)c'\kappa(2+3\sqrt{2\log\frac{2}{\zeta}})}{\sqrt{m}},$$

and

$$\left| \frac{\partial}{\partial d_t} \left\{ \mathbb{E}_{\mathcal{Q}}[\ell(-f^*(\boldsymbol{X}))] - \mathbb{E}_{\mathcal{Q}}[\ell(-\hat{f}(\boldsymbol{X}))] \right\} \right| \leq \frac{2(\sqrt{2}s+2)c'\kappa(2+3\sqrt{2\log\frac{2}{\zeta}})}{\sqrt{m}} + \frac{(4+\delta)\varepsilon}{\gamma - 2\varepsilon} + D_s,$$

where the approximation error $D_{s,s'} := \inf_{f \in \mathcal{F}^+_{s,s'}(\gamma)} \frac{\partial}{\partial d_t}\mathbb{E}_{\mathcal{Q}}[\ell(-f(\boldsymbol{X}))] - \frac{\partial}{\partial d_t}\mathbb{E}_{\mathcal{Q}}[\ell(-f^*(\boldsymbol{X}))] \to 0$ as $s \to \infty, s' \to p$. Therefore, by the triangle inequality, with probability at least $1 - 3\zeta$, we have

$$\begin{aligned} &\left| \frac{\partial}{\partial d_t} \left\{ \mathbb{E}_{\mathcal{Q}}[\ell(-f^*(\boldsymbol{X}))] - \frac{1}{m} \sum_{j=1}^{m} \ell(-\hat{f}(\boldsymbol{x}_j)) \right\} \right|_{d_t=0, d_{t'}=d^*_{t'}, t\neq t'} \\ &\leq \frac{3(\sqrt{2}s+2)c'\kappa(2+3\sqrt{2\log\frac{2}{\zeta}})}{\sqrt{m}} + \frac{(4+\delta)\varepsilon}{\gamma - 2\varepsilon} + D_{s,s'}. \end{aligned}$$

Consequently, for those important features $\boldsymbol{x}_{\cdot,t}$ we have

$$\begin{aligned} &\frac{\partial}{\partial d_t} \frac{1}{m} \sum_{j=1}^{m} \ell(-\hat{f}(\boldsymbol{x}_j)) \Big\} \Big|_{d_t=0, d_{t'}=d^*_{t'}, t\neq t'} \\ < \; &\frac{\partial \mathbb{E}_{\mathcal{Q}}[\ell(-f^*(\boldsymbol{X}))]}{\partial d_t} \Big|_{d_t=0, d_{t'}=d^*_{t'}, t\neq t'} + O\left(\max(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{m}}, D_{s,s'}) \right), \end{aligned}$$

and for those noise features $\boldsymbol{x}_{\cdot,t}$ we have

$$\begin{aligned} &\frac{\partial}{\partial d_t} \frac{1}{m} \sum_{j=1}^{m} \ell(-\hat{f}(\boldsymbol{x}_j)) \Big\} \Big|_{d_t=0, d_{t'}=d^*_{t'}, t\neq t'} \\ \geq \; &\frac{\partial \mathbb{E}_{\mathcal{Q}}[\ell(-f^*(\boldsymbol{X}))]}{\partial d_t} \Big|_{d_t=0, d_{t'}=d^*_{t'}, t\neq t'} - O\left(\max(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{m}}, D_{s,s'}) \right). \end{aligned}$$

Finally, together with the assumption in Theorem 6 for important and unimportant features, we have $\mathbb{P}\left[\operatorname{sign}(\hat{d}_t) = \operatorname{sign}(d^*_t)\right] \longrightarrow 1,\ t \in [p]$. ■

References

Genevera I Allen. Automatic feature selection via weighted kernels and regularization. *Journal of Computational and Graphical Statistics*, 22(2):284–299, 2013.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840, 2008.

Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pages 1507–1517. PMLR, 2021.

Jingxiang Chen, Chong Zhang, Michael R Kosorok, and Yufeng Liu. Double sparsity kernel learning with automatic variable selection and data extraction. *Statistics and its interface*, 11(3):401, 2018.

Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112 (520):1684–1696, 2017.

CK Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Christophe Denis and Mohamed Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *arXiv preprint arXiv:1507.07235*, 2015.

Christophe Denis and Mohamed Hebiri. Confidence sets with expected sizes for multiclass classification. *The Journal of Machine Learning Research*, 18(1):3571–3598, 2017.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018.

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in neural information processing systems*, pages 703–711, 2014.

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394, 2015.

Lutz Dümbgen, Bernd-Wolfgang Igl, and Axel Munk. P-values for classification. *Electronic Journal of Statistics*, 2(none), Jan 2008. ISSN 1935-7524. doi: 10.1214/08-ejs245. URL `http://dx.doi.org/10.1214/08-EJS245`.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The annals of statistics*, 32(3):928–961, 2004.

Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. Learning bounds for open-set learning. In *International Conference on Machine Learning*, pages 3122–3132. PMLR, 2021.

Saurabh Garg, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under open set label shift. *Advances in Neural Information Processing Systems*, 35:22531–22546, 2022.

Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. Rlsbench: Domain adaptation under relaxed label shift. In *International Conference on Machine Learning*, pages 10879–10928. PMLR, 2023.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3614–3631, 2020.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR, 2016.

Leying Guan and Robert Tibshirani. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 84(2): 524, 2022.

Blaise Hanczar and Michèle Sebag. Combination of one-class support vector machines for classification with reject option. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 547–562. Springer, 2014.

Yotam Hechtlinger, Barnabás Póczos, and Larry Wasserman. Cautious deep learning. *arXiv preprint arXiv:1805.09460*, 2018.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Radu Herbei and Marten H Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721, 2006.

Vilen Jumutc and Johan AK Suykens. Supervised novelty detection. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 143–149. IEEE, 2013.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR, 2022.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Konstantin Kirchheim, Marco Filax, and Frank Ortmeier. Pytorch-ood: A library for out-of-distribution detection based on pytorch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4351–4360, June 2022.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020.

Wonyul Lee, Ying Du, Wei Sun, David Neil Hayes, and Yufeng Liu. Multiple response regression for gaussian mixture models with known labels. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(6):493–508, 2012.

Jing Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):29–43, 2015.

Si Liu, Risheek Garrepalli, Thomas Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with pac guarantees. In *International Conference on Machine Learning*, pages 3169–3178. PMLR, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Tengyu Ma. Lecture notes for machine learning theory. `https://github.com/tengyuma/cs229m_notes/blob/main/master.pdf`, 2022. Accessed: 2022–06-22.

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613–628, 2018.

Xingye Qiao and Yufeng Liu. Adaptive weighted learning for unbalanced multicategory classification. *Biometrics*, 65(1):159–168, 2009.

Xingye Qiao, Hao Helen Zhang, Yufeng Liu, Michael J Todd, and James Stephen Marron. Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association*, 105(489):401–414, 2010.

Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with a reject option. *arXiv preprint arXiv:1505.04137*, 2015.

Philippe Rigollet and Xin Tong. Neyman-pearson classification, convexity and stochastic constraints. *The Journal of Machine Learning Research*, 12:2831–2855, 2011.

Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019.

Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588, 2000.

Bernhard Schölkopf, Alexander J Smola, and Francis Bach. *Learning with kernels: support vector machines, regularization, optimization, and beyond*, chapter 7, page 209. the MIT Press, 2018.

Clayton Scott and Robert Nowak. A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

Alistair Shilton, Sutharshan Rajasegarar, and Marimuthu Palaniswami. Multiclass anomaly detector: the cs++ support vector machine. *J. Mach. Learn. Res.*, 21:213–1, 2020.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(Feb):211–232, 2005.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33:19276–19289, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Xin Tong, Yang Feng, and Anqi Zhao. A survey on neyman-pearson classification and suggestions for future research. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(2):64–81, 2016.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Wenbo Wang and Xingye Qiao. Learning confidence sets using support vector machines. In *Advances in Neural Information Processing Systems*, pages 4929–4938, 2018.

Wenbo Wang and Xingye Qiao. Set-valued support vector machine with bounded error rates. *Journal of the American Statistical Association*, pages 1–13, 2022.

Yichao Wu, Hao Helen Zhang, and Yufeng Liu. Robust model-free multiclass probability estimation. *Journal of the American Statistical Association*, 105(489):424–436, 2010.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International conference on machine learning*, pages 6872–6881. PMLR, 2019.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Yiming Ying and Ding-Xuan Zhou. Learnability of gaussians with flexible variances. *Journal of Machine Learning Research*, 8(9):249–276, 2007. URL `http://jmlr.org/papers/v8/ying07a.html`.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Chong Zhang, Yufeng Liu, and Zhengxiao Wu. On the effect and remedies of shrinkage on classification probability estimation. *The American Statistician*, 67(3):134–142, 2013a.

Chong Zhang, Wenbo Wang, and Xingye Qiao. On reject and refine options in multicategory classification. *Journal of the American Statistical Association*, 113(522):730–745, 2018.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR, 2013b.

Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.