Anomaly detection with semi-supervised classification based on risk estimators

Anonymous authors
Paper under double-blind review

Abstract

A significant limitation of one-class classification anomaly detection methods is their reliance on the assumption that unlabeled training data only contains normal instances. To overcome this impractical assumption, we propose two novel classification-based anomaly detection methods. Firstly, we introduce a semi-supervised shallow anomaly detection method based on an unbiased risk estimator. Secondly, we present a semi-supervised deep anomaly detection method utilizing a nonnegative (biased) risk estimator. We establish estimation error bounds and excess risk bounds for both risk minimizers. Additionally, we propose techniques to select appropriate regularization parameters that ensure the nonnegativity of the empirical risk in the shallow model under specific loss functions. Our extensive experiments provide evidence of the effectiveness of the risk-based anomaly detection methods.

1 Introduction

Anomaly Detection (AD) can be defined as the task of identifying instances that deviates significantly from the majority of the data instances, see e.g., (Chandola et al., 2009; Pang et al., 2020; Ruff et al., 2021) for comprehensive surveys on AD. One important approach for AD is one-class classification Khan & Madden (2014); Tax & Duin (1999). It can be viewed as a specialized binary classification problem aimed at learning a model that distinguishes between positive (normal) and negative (anomalous) classes. This approach assumes that the unlabeled dataset primarily consists of data from the normal class. By utilizing a sufficient amount of normal data, one-class classification AD (OC-AD) methods identify a decision boundary that encompasses all the normal points. For example, the decision boundaries of *shallow* OC-AD methods include a hyperplane with maximum margin Schölkopf et al. (2001), a compact spherical boundary Tax & Duin (1999; 2004), an elliptical boundary (Rousseeuw & Van Driessen, 1999; Rousseeuw, 1985), a pair of subspaces Wang & Cherian (2019), or even a collection of multiple spheres Görnitz et al. (2018). To enhance their applicability in high-dimensional settings, these shallow methods have been extended into *deep* methods Erfani et al. (2016); Ruff et al. (2018).

Unsupervised learning, where only unlabeled data is available, represents the most common setting in AD. Unsupervised AD methods typically assume that the training data consists solely of normal instances Hodge & Austin (2004); Pimentel et al. (2014); Zimek et al. (2012). However, in real-world scenarios, labeled samples may be available alongside the unlabeled dataset, leading to the development of semi-supervised AD methods, including semi-supervised OC-AD methods Görnitz et al. (2009); Munoz-Mari et al. (2010); Ruff et al. (2020). It is important to note that unsupervised/semi-supervised shallow/deep one-class anomaly detection methods do not explicitly handle mixed unlabeled data. This is because they typically assume that there are no anomalous instances present in the unlabeled dataset, which is impractical in real-world scenarios. Classification methods that handle mixed unlabeled data have been extensively studied in the field of learning with positive and unlabeled examples (LPUE or PU learning). In this context, we have access to information on positive and unlabeled data, but negative data is unavailable. PU learning methods have also been utilized as semi-supervised AD methods Bekker & Davis (2020); Blanchard et al. (2010); Chandola et al. (2009); Ju et al. (2020). It is widely recognized that incorporating labeled anomalies, even if only a few instances, can greatly enhance the AD performance Görnitz et al. (2013); Kiran et al. (2018). Semi-supervised AD methods

that consider the availability of negative data have demonstrated highly promising AD performance Han et al. (2022); Ruff et al. (2021; 2020).

To overcome the impractical assumption of OC-AD methods, we adopt the key concept of risk-based PU learning methods du Plessis et al. (2014; 2015); Kiryo et al. (2017); Sakai et al. (2017). These methods propose empirical estimators for the risk associated with the learning problem. In order to improve anomaly detection performance, we focus on the semi-supervised setting where a negative dataset is also available. It is noteworthy that the estimation of risk in anomaly detection is a relatively unexplored subject, distinguished by specific characteristics, especially in terms of error bounds, which are not commonly found in current anomaly detection approaches.

Contributions Our main contributions are summarized as follows.

- Considering AD as a semi-supervised binary classification problem, where we have access to a positive dataset, a negative dataset, and an unlabeled dataset that may contain anomalous examples, we introduce two risk-based AD methods. These methods include a shallow AD approach developed using an unbiased risk estimator and a deep AD method based on a nonnegative risk estimator.
- We develop methods to select suitable regularization that ensures the nonnegativity of the empirical risk in the proposed shallow AD method. This is crucial as negative empirical risk can lead to significant overfitting issues Kiryo et al. (2017).
- We additionally establish estimation error bounds and excess risk bounds for the two risk minimizers, building upon the theoretical findings presented in Kiryo et al. (2017); Niu et al. (2016).
- We conduct extensive experiments on benchmark AD datasets obtained from *Adbench* Han et al. (2022) to compare the performance of our proposed risk-based AD (rAD) methods against various baseline methods.

Organization We discuss related work in Section 2 and provide a brief background on risk estimators in Section 3. We then introduce the two risk estimators that form the basis of our risk-based AD methods in Section 4. Additionally, we present a theoretical analysis in Section 5, present experimental results in Section 6, highlight limitations in Section 7, and conclude the paper in Section 8. All proofs and additional experiments can be found in the supplementary material.

2 Related work

We direct readers to Ruff et al. (2021); Roth et al. (2022) for a comprehensive review of recent advancements in anomaly detection techniques (with a particular focus on industrial anomaly detection). We remark that our primary contribution lies in delving deeper into the approach of risk estimation, a technique that remains relatively unexplored within the context of anomaly detection. In the following, we provide a brief overview of the most relevant works related to our proposed risk-based anomaly detection methods.

AD methods Outlier detection (Hawkins, 1980; Hodge & Austin, 2004), novelty detection (Salu, 1988; Pimentel et al., 2014), and AD are closely related topics. In fact, these terms have been used interchangeably, and solutions to outlier detection and novelty detection are often used for AD and vice versa. AD methods can be generally classified into three types (i) density-based methods, which estimate the probability distribution of normal instances Lecun et al. (2006); Li et al. (2019); Parzen (1962); Pincus (1995), (ii) reconstruction-based methods, which learn a model that fails to reconstruct anomalous instances but succeeds to reconstruct normal instances Dhillon et al. (2004); Hawkins (1974); Hawkins et al. (2002); Huang et al. (2006); Yan et al. (2021), and (iii) one-class classification methods. We refer the readers to Ruff et al. (2021) for a comprehensive review of the three types of AD methods.

PU learning methods Regarding PU learning methods, they can be classified into three categories: biased learning, two-step techniques, and class-prior incorporation. Similarly to one-class classification AD methods, biased PU learning methods make an impractical assumption: they assume/label all unlabeled instances as negative, see e.g., Lee & Liu (2003); Liu et al. (2003). Although the PU learning methods using two-step techniques do not have such assumption, they are heuristics since they first identify “reliable"

negative examples and then apply (semi-)supervised learning techniques to the positive labeled instances and the reliable negative instances, see e.g., Li & Liu (2003); Chaudhari & Shevade (2012). To have some theoretical guarantee, the class-prior incorporation methods need to assume that the class priors are known, see e.g., du Plessis et al. (2014); Elkan & Noto (2008); Hsieh et al. (2019). We refer the readers to Bekker & Davis (2020) and the references therein for more details on the three types of PU learning methods. Methods that rely on risk estimators du Plessis et al. (2014; 2015); Kiryo et al. (2017); Sakai et al. (2017) belong to the third category.

3 Background on risk estimators

Let x and $y \in \{+1, -1\}$ be random variables with joint density $p(x, y)$. The class-conditional densities are $p_p(x) = P(x|y = +1)$ and $p_n(x) = P(x|y = -1)$. Let $\pi_p = p(y = +1)$ and $\pi_n = p(y = -1)$ be the class-prior probabilities for the positive and negative classes. We have $\pi_p + \pi_n = 1$. Suppose the positive ($\mathcal{P}$), negative ($\mathcal{N}$) and unlabeled ($\mathcal{U}$) data are sampled independently as $(\mathcal{P}) = \{x_i^p\}_{i=1}^{n_p} \sim p_p(x), (\mathcal{N}) = \{x_i^n\}_{i=1}^{n_n} \sim p_n(x), (\mathcal{U}) = \{x_i^u\}_{i=1}^{n_u} \sim p(x)$, where

$$p(x) = \pi_p p_p(x) + \pi_n p_n(x). \tag{1}$$

Given ($\mathcal{P}$), ($\mathcal{N}$) and ($\mathcal{U}$), let us consider a binary classification problem from x to y. Suppose $g : \mathbb{R}^d \to \mathbb{R}$ is a decision function that needs to be trained from ($\mathcal{P}$), ($\mathcal{N}$) and ($\mathcal{U}$), and $\ell : \mathbb{R} \times \{+1, -1\} \to \mathbb{R}$ is a loss function that imposes a cost $\ell(t, y)$ if the predicted output is t and the expected output is y. Under loss ℓ, let us denote

$$\mathcal{R}_p^+(g) = \mathbb{E}_{x\sim p_p(x)}[\ell(g(x), +1)], \mathcal{R}_n^+(g) = \mathbb{E}_{x\sim p_n(x)}[\ell(g(x), +1)], \mathcal{R}_u^+(g) = \mathbb{E}_{x\sim p(x)}[\ell(g(x), +1)],$$
$$\mathcal{R}_p^-(g) = \mathbb{E}_{x\sim p_p(x)}[\ell(g(x), -1)], \mathcal{R}_n^-(g) = \mathbb{E}_{x\sim p_n(x)}[\ell(g(x), -1)], \mathcal{R}_u^-(g) = \mathbb{E}_{x\sim p(x)}[\ell(g(x), -1)].$$

Given ℓ and assuming that π_p is known (in practice, π_p can be effectively estimated from ($\mathcal{P}$), ($\mathcal{N}$) and ($\mathcal{U}$) du Plessis & Sugiyama (2013); Saerens et al. (2002)), our goal is to find g that minimizes the risk of g, which is defined by

$$\mathcal{R}(g) := \mathbb{E}_{(x,y)\sim p(x,y)}[\ell(g(x), y)] = \pi_p \mathcal{R}_p^+(g) + \pi_n \mathcal{R}_n^-(g). \tag{2}$$

In ordinary classification, the optimal classifier minimizes the expected misclassification rate that corresponds to using zero-one loss in (2), $\ell_{0\text{-}1}(t, y) = 0$ if $ty > 0$ and $\ell_{0\text{-}1}(t, y) = 1$ otherwise. We denote $I(g) = \mathbb{E}_{(x,y)\sim p(x,y)}[\ell_{0\text{-}1}(g(x), y)]$.

PN risk estimator In supervised learning when we have fully labeled data, $\mathcal{R}(g)$ can be approximated by a PN risk estimator $\hat{\mathcal{R}}_{pn}(g) = \pi_p \hat{\mathcal{R}}_p^+(g) + \pi_n \hat{\mathcal{R}}_n^-(g)$, where

$$\hat{\mathcal{R}}_p^+(g) := \frac{1}{n_p} \sum_{i=1}^{n_p} \ell(g(x_i^p), +1), \quad \hat{\mathcal{R}}_n^-(g) := \frac{1}{n_n} \sum_{i=1}^{n_n} \ell(g(x_i^n), -1). \tag{3}$$

PU risk estimator In PU learning when ($\mathcal{N}$) is unavailable, du Plessis et al. (2014; 2015); Kiryo et al. (2017) propose methods to approximate $\mathcal{R}(g)$ from ($\mathcal{P}$) and ($\mathcal{U}$). From (1) we have $\pi_n \mathcal{R}_n^-(g) = \mathcal{R}_u^-(g) - \pi_p \mathcal{R}_p^-(g)$, which implies that

$$\mathcal{R}(g) = \pi_p(\mathcal{R}_p^+ - \mathcal{R}_p^-) + \mathcal{R}_u^-(g). \tag{4}$$

When ℓ satisfies the symmetric condition $\ell(t, +1) + \ell(t, -1) = 1$ then we have $\mathcal{R}(g) = 2\pi_p \mathcal{R}_p^+(g) - \pi_p + \mathcal{R}_u^-(g)$, which can be approximated by

$$\hat{\mathcal{R}}_{pu}^{(1)}(g) = 2\pi_p \hat{\mathcal{R}}_p^+(g) - \pi_p + \hat{\mathcal{R}}_u^-(g), \tag{5}$$

where $\hat{\mathcal{R}}_p^+(g)$ is defined in (3) and $\hat{\mathcal{R}}_u^-(g) = \frac{1}{n_u} \sum_{i=1}^{n_u} \ell(g(x_i^u), -1)$, see du Plessis et al. (2014). When ℓ satisfies the linear-odd condition $\ell(t, +1) - \ell(t, -1) = -t$ then $\mathcal{R}(g)$ can be approximated by

$$\hat{\mathcal{R}}_{pu}^{(2)}(g) = -\pi_p \frac{1}{n_p} \sum_{i=1}^{n_p} g(x_i^p) + \hat{\mathcal{R}}_u^-(g), \tag{6}$$

see du Plessis et al. (2015). The authors in Kiryo et al. (2017) propose a *non-negative* PU risk estimator

$$\hat{\mathcal{R}}_{pu}^{(3)}(g) = \pi_p \hat{\mathcal{R}}_p^+(g) + \max\{0, \hat{\mathcal{R}}_u^-(g) - \pi_p \hat{\mathcal{R}}_p^-(g)\}, \tag{7}$$

where $\hat{\mathcal{R}}_p^-(g) = \frac{1}{n_p}\sum_{i=1}^{n_p} \ell(g(x_i^p), -1)$. Note that $\hat{\mathcal{R}}_{pu}^{(3)}(g)$ is a biased estimator.

NU risk estimator Similarly, considering NU learning when $(\mathcal{P})$ is unavailable, see Sakai et al. (2017), NU risk estimators can be formulated by combining the equation $\pi_p\mathcal{R}_p^+(g) = \mathcal{R}_u^+(g) - \pi_n\mathcal{R}_n^+(g)$ (which is derived from (1)) and (2) to obtain

$$\mathcal{R}(g) = -\pi_n(\mathcal{R}_n^+ - \mathcal{R}_n^-) + \mathcal{R}_u^+(g). \tag{8}$$

With a loss satisfying the symmetric condition, we have a nonconvex NU risk estimator

$$\hat{\mathcal{R}}_{nu}^{(1)}(g) = 2\pi_n\hat{\mathcal{R}}_n^-(g) - \pi_n + \hat{\mathcal{R}}_u^+(g), \tag{9}$$

where $\hat{\mathcal{R}}_n^-(g)$ is defined in (3) and $\hat{\mathcal{R}}_u^+(g) = \frac{1}{n_u}\sum_{i=1}^{n_u} \ell(g(x_i^u), +1)$. And with a loss satisfying the linear-odd condition, we get a convex NU risk estimator

$$\hat{\mathcal{R}}_{nu}^{(2)}(g) = \pi_n\frac{1}{n_n}\sum_{i=1}^{n_n} g(x_i^n) + \hat{\mathcal{R}}_u^+(g). \tag{10}$$

Finally, Sakai et al. (2017) proposes to use a linear combination between the PN, the NU, and the PU risk ofdu Plessis et al. (2014; 2015).

4 The proposed semi-supervised anomaly detection methods

In the previous section, we presented risk estimators for the PU learning problem where $(\mathcal{N})$ is unavailable. Let us consider the setting where we have access to $(\mathcal{P})$, $(\mathcal{N})$ as well as $(\mathcal{U})$. We perceive semi-supervised AD as a binary classification problem from x to $y \in \{+1, -1\}$, where $+1$ represents the normal class and -1 represents the anomalous class. Our goal is to propose risk estimators for the risk in (2). Specifically, we propose two risk estimators for semi-supervised AD that lead to two risk-based AD methods.

If we take a convex combination of (2) and (8), we obtain

$$\begin{aligned}\mathcal{R}(g) &= a(-\pi_n(\mathcal{R}_n^+ - \mathcal{R}_n^-) + \mathcal{R}_u^+(g)) + (1-a)(\pi_p\mathcal{R}_p^+(g) + \pi_n\mathcal{R}_n^-(g)) \\ &= a\mathcal{R}_u^+(g) + (1-a)\pi_p\mathcal{R}_p^+(g) + \pi_n\mathcal{R}_n^-(g) - a\pi_n\mathcal{R}_n^+,\end{aligned} \tag{11}$$

where $a \in (0, 1)$.

The empirical version of (11) yields the following linear combination of PN and NU risk estimators:

$$\hat{\mathcal{R}}_s^{(2)}(g) = a\hat{\mathcal{R}}_u^+(g) + (1-a)\pi_p\hat{\mathcal{R}}_p^+(g) + \pi_n\hat{\mathcal{R}}_n^-(g) - a\pi_n\hat{\mathcal{R}}_n^+(g). \tag{12}$$

While $\hat{\mathcal{R}}_s^{(2)}(g)$ was also considered in Sakai et al. (2017), they only focused on the set of linear classifiers with two specific losses – the (scaled) ramp loss and the truncated (scaled) squared loss (see (Sakai et al., 2017, Section 4.1)). We consider a more general setting for $\hat{\mathcal{R}}_s^{(2)}$ and also propose methods to choose appropriate regularization for $\hat{\mathcal{R}}_s^{(2)}$ to avoid negative empirical risks. In fact, $\hat{\mathcal{R}}_s^{(2)}$ may take negative values when ℓ is unbounded due to the negative term $-a\pi_n\hat{\mathcal{R}}_n^+(g)$. Theorem 1 summarizes the conditions that guarantee a nonnegative objective.

Inspired by $\hat{\mathcal{R}}_{pu}^{(3)}(g)$ in (7), we also propose the following nonnegative risk estimator:

$$\hat{\mathcal{R}}_s^{(1)}(g) = \pi_n\hat{\mathcal{R}}_n^-(g) + (1-a)\pi_p\hat{\mathcal{R}}_p^+(g) + a\max\{0, \hat{\mathcal{R}}_u^+(g) - \pi_n\hat{\mathcal{R}}_n^+(g)\}, \tag{13}$$

where the max term is introduced since $\mathcal{R}_u^+(g) - \pi_n\mathcal{R}_n^+(g) = \pi_p\mathcal{R}_p^+$ must be nonnegative. Note that $\hat{\mathcal{R}}_{pu}^{(3)}(g)$ was designed for the PU learning problem while we propose $\hat{\mathcal{R}}_s^{(1)}(g)$ for the AD problem which often assumes anomalies are rare. In other words, we put more emphasis on $\hat{\mathcal{R}}_u^+(g)$ rather than $\hat{\mathcal{R}}_u^-(g)$.

In Section 5, we will establish the theoretical estimation error bounds and excess risk bounds for the minimizers of both $\min_{g\in\mathcal{G}}\hat{\mathcal{R}}_s^{(1)}(g)$ and $\min_{g\in\mathcal{G}}\hat{\mathcal{R}}_s^{(2)}(g)$, where $\mathcal{G}$ is some class function. We now present the practical optimization problems involved when using $\hat{\mathcal{R}}_s^{(1)}(g)$ and $\hat{\mathcal{R}}_s^{(2)}(g)$.

Optimization problems Suppose g is parameterized by w, which needs to be learned from $(\mathcal{P})$, $(\mathcal{N})$ and $(\mathcal{U})$. When $\hat{\mathcal{R}}_s^{(1)}$ in (13) is used, the corresponding optimization problem for AD is

$$\begin{aligned}\min_w \Big\{ &\frac{\pi_n}{n_n}\sum_{i=1}^{n_n}\ell(g(x_i^n),-1) + \frac{(1-a)\pi_p}{n_p}\sum_{i=1}^{n_p}\ell(g(x_i^p),+1) \\ &+ a\max\Big\{0, \frac{1}{n_u}\sum_{i=1}^{n_u}\ell(g(x_i^u),+1) - \frac{\pi_n}{n_n}\sum_{i=1}^{n_n}\ell(g(x_i^n),+1)\Big\} + \lambda\mathbf{R}(w)\Big\},\end{aligned} \tag{14}$$

where $\mathbf{R}$ is some regularizer, and $\lambda \geq 0$ is regularization parameter. And when $\hat{\mathcal{R}}_s^{(2)}$ in (12) is used, the corresponding optimization problem is

$$\begin{aligned}\min_w \Big\{ &\frac{a}{n_u}\sum_{i=1}^{n_u}\ell(g(x_i^u),+1) + \frac{(1-a)\pi_p}{n_p}\sum_{i=1}^{n_p}\ell(g(x_i^p),+1) \\ &+ \frac{\pi_n}{n_n}\sum_{i=1}^{n_n}\ell(g(x_i^n),-1) - \frac{a\pi_n}{n_n}\sum_{i=1}^{n_n}\ell(g(x_i^n),+1) + \lambda\mathbf{R}(w)\Big\}.\end{aligned} \tag{15}$$

Unfortunately, the objective of (15) is not guaranteed to be nonnegative due to the negative term $-\frac{a\pi_n}{n_n}\sum_{i=1}^{n_n}\ell(g(x_i^n),+1)$. As pointed out by Kiryo et al. (2017), this can lead to serious overfitting problems. The following theorem provides methods to choose the regularization parameters such that the nonnegativity of the objective of (15) is guaranteed.

Theorem 1 *Suppose there exist positive constants b_1, b_2 and b_3 such that*

$$\ell(t,-1) - \ell(t,+1) \geq -b_1|t|, \quad \textit{and} \quad \ell(t,-1) \geq b_2(b_3 - |t|). \tag{16}$$

(In Table 1 we give examples of loss functions that satisfy (16)*, see their proofs in the supp. material.)*

(i) We have

$$\frac{\pi_n}{n_n}\sum_{i=1}^{n_n}\ell(g(x_i^n),-1) - \frac{a\pi_n}{n_n}\sum_{i=1}^{n_n}\ell(g(x_i^n),+1) \geq (1-a)\pi_n b_2 b_3 - ((1-a)b_2 + ab_1)\frac{\pi_n}{n_n}\sum_{i=1}^{n_n}|g(x_i^n)|.$$

(ii) If we choose λ and $\mathbf{R}$ such that

$$\lambda\mathbf{R}(w) \geq ((1-a)b_2 + ab_1)\frac{\pi_n}{n_n}\sum_{i=1}^{n_n}|g(x_i^n)| - (1-a)\pi_n b_2 b_3 \tag{17}$$

then the objective of (15) *is always nonnegative.*

(iii) Consider the specific case $g(x) = \langle w, \phi(x)\rangle$, where $\phi : \mathbb{R}^d \to \mathbb{R}^q$ is a feature map transformation. The following choices of λ and $\mathbf{R}$ satisfy (17)*.*

- $\mathbf{R}(w) = \|w\|_2^2$ *and* $\lambda \geq \frac{((1-a)b_2+ab_1)^2\pi_n c^2}{4(1-a)b_2 b_3}$*, where* $c = \max\{\|\phi(x_i^n)\|_2 : i = 1,\dots,n_n\}$ *(note that, in practice, we can scale the data to have* $c = 1$*).*
- $\mathbf{R}(w) = \|w\|_1$ *and* $\lambda \geq c_\infty((1-a)b_2 b_3 + ab_1)\pi_n$*, where* $c_\infty = \max\{\|\phi(x_i^n)\|_\infty : i = 1,\dots,n_n\}$ *(in practice, we can scale the data to have* $c_\infty = 1$*).*

We consider both a shallow and deep implementation of the rAD method. In the following, π_p^e and $\pi_n^e = 1 - \pi_p^e$ will denote estimates of the real class-prior probabilities π_p and π_n, respectively.

A shallow rAD method We plug in $g(x) = \langle w, \phi(x)\rangle$ in (15) (the empirical version of (12)), where $\phi : \mathbb{R}^d \to \mathbb{R}^q$ is a feature map transformation, and choose the regularization method proposed in Theorem 1

Table 1: Examples of loss functions satisfying (16)

Name	$\ell(t,y)=\ell(z)$ with $z=ty$	Bounded	(b_1,b_2,b_3)
Hinge loss	$\max\{0,1-z\}$	×	$(2,1,1)$
Double hinge loss	$\max\{0,(1-z)/2,-z\}$	×	$(1,1/2,1)$
Squared loss	$\frac{1}{2}(z-1)^2$	×	$(2,1/2,1/2)$
Modified Huber loss	$\begin{cases}\max\{0,1-z\}^2 & \text{if } z\geq -1\\ -4z & \text{otherwise}\end{cases}$	×	$(4,1,1/2)$
Logistic loss	$\ln(1+\exp(-z))$	×	$(1,1,\ln 2)$
Sigmoid loss	$1/(1+\exp(z))$	✓	$(1,1/2,1)$
Ramp loss	$\max\{0,\min\{1,(1-z)/2\}\}$	✓	$(1,1/2,1)$

(iii). Specifically, we solve the following minimization problem:

$$\min_w \Big\{ \frac{a}{n_u}\sum_{i=1}^{n_u}\ell(w^\top\phi(x_i^u),+1)+\frac{(1-a)\pi_p^e}{n_p}\sum_{i=1}^{n_p}\ell(w^\top\phi(x_i^p),+1) + \frac{\pi_n^e}{n_n}\sum_{i=1}^{n_n}\ell(w^\top\phi(x_i^n),-1)-\frac{a\pi_n^e}{n_n}\sum_{i=1}^{n_n}\ell(w^\top\phi(x_i^n),+1)+\lambda\mathbf{R}(w)\Big\}. \tag{18}$$

A deep rAD method We plug in $g(x)=\phi(x;\mathcal{W})$ in (14) (the empirical version of (13)), where $\mathcal{W}$ is a set of weights of a deep neural network. Specifically, we train a deep neural network by solving the following optimization problem:

$$\min_{\mathcal{W}} \Big\{ \frac{\pi_n^e}{n_n}\sum_{i=1}^{n_n}\ell(\phi(x_i^n;\mathcal{W}),-1)+\frac{(1-a)\pi_p^e}{n_p}\sum_{i=1}^{n_p}\ell(\phi(x_i^p;\mathcal{W}),+1) + a\max\Big\{0,\frac{1}{n_u}\sum_{i=1}^{n_u}\ell(\phi(x_i^u;\mathcal{W}),+1)-\frac{\pi_n^e}{n_n}\sum_{i=1}^{n_n}\ell(\phi(x_i^n;\mathcal{W}),+1)\Big\}+\lambda\mathbf{R}(\mathcal{W})\Big\}, \tag{19}$$

where $\mathbf{R}$ can be any regularizer. Note that we focus on these specific implementations but it is also possible to consider a deep model with (15) or a shallow model with (14). It is noteworthy to mention that in our initial numerical findings, we have observed that the shallow model in (18) frequently yields better results compared to the shallow model with (14), while the deep model in (19) outperforms the deep model with (15).

5 Risk bounds

In this section, we establish the estimation error bound and the excess risk bound for $\hat{g}^1$ and $\hat{g}^2$ which are the empirical risk minimizers obtained by $\min_{g\in\mathcal{G}}\hat{\mathcal{R}}_s^{(1)}(g)$ and $\min_{g\in\mathcal{G}}\hat{\mathcal{R}}_s^{(2)}(g)$, where $\mathcal{G}$ is a function class.

Let g^* be the true risk minimizer, that is, $g^*=\arg\min_{g\in\mathcal{G}}\mathcal{R}(g)$. Throughout this section, we assume that (i) $\mathcal{G}=\big\{g\,\big|\,\|g\|_\infty\leq C_g\big\}$ for some constant C_g, and (ii) there exists $C_\ell>0$ such that $\sup_{|t|\leq C_g}\max_y\ell(t,y)\leq C_\ell$. It is worth noting that the set of linear classifiers with bounded norms and feature maps is a special case of Condition (i)

$$\mathcal{G}=\{g(x)=\langle w,\phi(x)\rangle_{\mathcal{H}}\,\big|\,\|w\|_{\mathcal{H}}\leq C_w,\|\phi(x)\|_{\mathcal{H}}\leq C_\phi\}, \tag{20}$$

where $\mathcal{H}$ is a Hilbert space, ϕ is a feature map, and C_w and C_ϕ are positive constants.

Given g, $\hat{\mathcal{R}}_s^{(2)}(g)$ is an unbiased estimator of $\mathcal{R}(g)$ but $\hat{\mathcal{R}}_s^{(1)}$ is a biased estimator. The following proposition estimates the bias of $\hat{\mathcal{R}}_s^{(1)}$ (see Inequality (21)) and shows that, for a fixed g, $\hat{\mathcal{R}}_s^{(1)}(g)$ and $\hat{\mathcal{R}}_s^{(2)}(g)$ converge to $\mathcal{R}(g)$ with the rate $O\big(\frac{\pi_n}{\sqrt{n_n}}+\frac{\pi_p}{\sqrt{n_p}}+\frac{a}{\sqrt{n_u}}\big)$ (see Inequality (22) and (23)).

Proposition 1 *Consider a classifier g. Suppose there exists $\rho_g > 0$ such that $\mathcal{R}_p^+(g) \geq \rho_g > 0$ and denote $\epsilon_g = a\pi_n C_\ell \exp\left(-\frac{2\pi_p^2\rho_g^2}{C_\ell^2(1/n_u+\pi_n^2/n_n)}\right)$. Then the bias of $\hat{\mathcal{R}}_s^{(1)}(g)$ satisfies*

$$0 \leq \mathbb{E}[\hat{\mathcal{R}}_s^{(1)}(g)] - \mathcal{R}(g) \leq \epsilon_g. \tag{21}$$

Moreover, for any $\delta > 0$, we have the following inequalities hold with probability at least $1-\delta$

$$|\hat{\mathcal{R}}_s^{(2)}(g) - \mathcal{R}(g)| \leq C_\ell\sqrt{\ln(2/\delta)/2}\big(\frac{(1+a)\pi_n}{\sqrt{n_n}} + \frac{(1-a)\pi_p}{\sqrt{n_p}} + \frac{a}{\sqrt{n_u}}\big), \tag{22}$$

and

$$|\hat{\mathcal{R}}_s^{(1)}(g) - \mathcal{R}(g)| \leq C_\ell\sqrt{\ln(2/\delta)/2}\big(\frac{(1+a)\pi_n}{\sqrt{n_n}} + \frac{(1-a)\pi_p}{\sqrt{n_p}} + \frac{a}{\sqrt{n_u}}\big) + \epsilon_g. \tag{23}$$

Estimation error bound The Rademacher complexity of $\mathcal{G}$ for a sample of size n drawn from some distribution q (see e.g., Mohri et al. (2018)) is defined by $\mathfrak{R}_{n,q}(\mathcal{G}) := \mathbb{E}_{Z\sim q^n}[\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}(\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i))]]$, where $Z_1, \ldots, Z_n$ are i.i.d random variables following distribution q, $Z = (Z_1, \ldots, Z_n)$, $\sigma_1, \ldots, \sigma_n$ are independent random variables uniformly chosen from $\{-1, 1\}$, and $\sigma = (\sigma_1, \ldots, \sigma_n)$. Similarly to (Kiryo et al., 2017, Theorem 4), we can establish the following estimation error bound for $\hat{g}^1$.

Theorem 2 (Estimation error bound for $\hat{g}^1$) *We assume that (i) there exists $\rho > 0$ such that $\mathcal{R}_p^+(g) \geq \rho$ for all $g \in \mathcal{G}$, (ii) if $g \in \mathcal{G}$ then $-g \in \mathcal{G}$, and (iii) $t \mapsto \ell(t, 1)$ and $t \mapsto \ell(t, -1)$ are L_ℓ-Lipschitz continuous over $\{t : |t| \leq C_g\}$. Denote $\epsilon = a\pi_n C_\ell \exp\left(-\frac{2\pi_p^2\rho^2}{C_\ell^2(1/n_u+\pi_n^2/n_n)}\right)$. For any $\delta > 0$, the following inequality hold with probability at least $1-\delta$*

$$\begin{aligned}\mathcal{R}(\hat{g}^1) - \mathcal{R}(g^*) \leq\ & 8(1+a)\pi_n L_\ell \mathfrak{R}_{n_n,p_n}(\mathcal{G}) + 8(1-a)\pi_p L_\ell \mathfrak{R}_{n_p,p_p}(\mathcal{G}) + 8aL_\ell\mathfrak{R}_{n_u,p}(\mathcal{G}) + \\ & + 2C_\ell\sqrt{\ln(2/\delta)/2}\Big(\frac{(1+a)\pi_n}{\sqrt{n_n}} + \frac{(1-a)\pi_p}{\sqrt{n_p}} + \frac{a}{\sqrt{n_u}}\Big) + 2\epsilon.\end{aligned} \tag{24}$$

By using basic uniform deviation bound Mohri et al. (2018), the McDiarmid's inequality McDiarmid (1989), and Talagrand's contraction lemma Ledoux & Talagrand (1991), we can prove the following estimation error bound for $\hat{g}^2$.

Theorem 3 (Estimation error bound for $\hat{g}^2$) *Assume that $t \mapsto \ell(t, 1)$ and $t \mapsto \ell(t, -1)$ are L_ℓ-Lipschitz continuous over $\{t : |t| \leq C_g\}$. For any small $\delta > 0$, the following inequality hold with probability at least $1-\delta$*

$$\begin{aligned}\mathcal{R}(\hat{g}^2) - \mathcal{R}(g^*) \leq\ & 4(1-a)\pi_p L_\ell \mathfrak{R}_{n_p,p_p}(\mathcal{G}) + 4(a+1)\pi_n L_\ell \mathfrak{R}_{n_n,p_n}(\mathcal{G}) + 4aL_\ell\mathfrak{R}_{n_u,p}(\mathcal{G}) + \\ & + 2C_\ell\sqrt{\ln(6/\delta)/2}\Big(\frac{(1-a)\pi_p}{\sqrt{n_p}} + \frac{(1+a)\pi_n}{\sqrt{n_n}} + \frac{a}{\sqrt{n_u}}\Big).\end{aligned} \tag{25}$$

Note that Theorem 3 explicitly states the error bound for $\hat{g}^2$ with any loss function that satisfies the Lipschitz continuity assumption. The (scaled) ramp loss and the truncated (scaled) squared loss considered in Sakai et al. (2017) have $L_\ell = 1/2$.

Excess risk bound The excess risk focuses on the error due to the use of surrogates for the 0-1 loss function. Denote $I^* = \inf_{g\in\mathcal{F}} I(g)$ and $\mathcal{R}^* = \inf_{g\in\mathcal{F}} \mathcal{R}(g)$, where $\mathcal{F}$ is the set of all measurable functions. By using (Bartlett et al., 2006, Theorem 1) (see (42) in the supp. material), Theorem 2, and Theorem 3, we can derive the following excess risk bound for $\hat{g}^1$ and $\hat{g}^2$.

Corollary 1 *If ℓ is a classification-calibrated loss (see Definition 1 in the supp. material), then there exists a convex, invertible, and nondecreasing transformation ψ_ℓ with $\psi_\ell(0) = 0$ and the following inequalities hold with probability at least $1-\delta$*

$$I(\hat{g}^1) - I^* \leq \psi_\ell^{-1}(B_1 + \mathcal{R}(g^*) - \mathcal{R}^*), \quad I(\hat{g}^2) - I^* \leq \psi_\ell^{-1}(B_2 + \mathcal{R}(g^*) - \mathcal{R}^*),$$

where B_1 and B_2 are the right hand side of (24) and (25), respectively.

Table 2: Mean (and SE $\times 10^2$) of the AUC over 30 trials. The best means are highlighted in bold. d, n, and π_n denote the feature dimension, the sample size of the dataset, and the ratio of negative samples in the dataset.

dataset (d, n, π_n)	rAD			PU			OC-SVM	semi-OC-SVM
	square	hinge	m-Huber	square	hinge	m-Huber		
pendigits (16, 6870, 0.02)	**0.98**(0.13)	**0.98**(0.13)	**0.98**(0.13)	0.78(3.90)	0.78(3.94)	0.78(3.84)	0.87(0.25)	0.77(2.74)
mammography (6, 11 183, 0.02)	**0.91**(0.35)	**0.91**(0.35)	**0.91**(0.34)	0.84(2.53)	0.84(2.54)	0.84(2.54)	0.76(0.48)	0.58(2.86)
optdigits (64, 5216, 0.03)	0.996(0.06)	0.996(0.07)	**0.997**(0.05)	0.80(1.97)	0.79(2.02)	0.80(1.98)	0.48(0.64)	0.78(2.44)
Stamps (9, 340, 0.09)	**0.82**(3.56)	0.81(4.08)	0.80(4.14)	0.70(4.17)	0.74(4.26)	0.66(4.47)	0.63(1.67)	0.74(3.37)
cardio (21, 1831, 0.10)	0.92(1.65)	0.88(1.73)	**0.93**(1.58)	0.82(2.34)	0.80(2.46)	0.83(2.23)	0.86(0.44)	0.79(1.97)
InternetAds (1555, 1966, 0.19)	0.76(2.90)	0.86(0.38)	**0.87**(0.50)	0.58(3.54)	0.78(0.51)	0.78(0.56)	0.60(0.50)	0.65(1.14)
Cardiotocography (21, 2114, 0.22)	0.89(0.92)	0.87(1.10)	**0.90**(0.87)	0.82(1.40)	0.80(1.64)	0.83(1.25)	0.74(0.35)	0.82(0.79)
magic.gamma (10, 19 020, 0.35)	**0.78**(0.51)	0.77(0.53)	**0.78**(0.50)	0.77(0.66)	0.77(0.68)	0.77(0.65)	0.56(0.18)	0.54(0.39)
SpamBase (57, 4207, 0.40)	**0.94**(0.14)	**0.94**(0.14)	**0.94**(0.13)	0.93(0.25)	0.93(0.23)	0.92(0.27)	0.54(0.40)	0.63(0.84)

6 Experiments

A. Experiments with shallow rAD

Baseline methods and implementation We compare rAD with OC-SVM Schölkopf et al. (2001), semi-supervised OC-SVM Munoz-Mari et al. (2010), and the PU methods using the risk estimator $\hat{\mathcal{R}}_{pu}(g)$ given in (4). Note that $\hat{\mathcal{R}}_{pu}(g) = \hat{\mathcal{R}}_{pu}^{(1)}(g)$ given in (5) if ℓ satisfies the symmetric condition, and $\hat{\mathcal{R}}_{pu}(g) = \hat{\mathcal{R}}_{pu}^{(2)}(g)$ given in (6) if ℓ satisfies the linear-odd condition. We implement rAD and PU methods with 3 losses: squared loss, hinge loss, and modified Huber loss. For rAD, we use l_2 regularization and take $\phi(x) = x$ in (18), i.e. no kernel is used. We set $a = 0.1$ and $\pi_p^e = 0.8$ ($\pi_n^e = 0.2$) as default values for both the shallow rAD and the PU methods. Note that the real π_n of the datasets can be different.

Datasets We test the algorithms on 26 classical anomaly detection benchmark datasets from Han et al. (2022), whose π_n ranges from 0.02 to 0.4. The real π_n of the datasets are given in the first column of Table 2 (and Table 5 in the supp. material). We randomly split each dataset 30 times into train and test data with a ratio of 7:3, i.e. we have 30 trials for each dataset. Then, for each trial, we randomly select 5% of the train data to make the labeled data and keep the remaining 95% as unlabeled data.

Experimental results In Table 2, we report the mean and standard error (SE) of the AUC (area under the ROC curve) over 30 trials of 9 benchmark datasets. The results of the 17 remaining datasets are given in Table 5 in the supp. material. We observe that, on average, rAD outperforms the PU methods and OC-SVM methods. The difference between the AUC of rAD and that of PU is large on the datasets with $\pi_n \leq 0.2$ but it is small when π_n is larger. We also notice that rAD with modified Huber loss often gives better results than rAD with square loss and hinge loss.

Sensitivity analysis for π_p^e With $a = 0.1$, we run shallow rAD on the 30 trials for $\pi_p^e \in \{1-\pi_n, 0.9, 0.7, 0.6\}$ (when $\pi_p^e = 1 - \pi_n$, no approximation is made). The results are reported in Table 3 for 9 benchmark datasets and the results of the 17 remaining datasets are given in Table 6 in the supp. material. From Table 2– 6, we can see that we can obtain good results even if π_p^e is different from π_p. In fact, with $a = 0.1$, we get worse AUC means when π_p^e is close to π_p. The combination $(a, \pi_p^e) = (0.1, 0.8)$ or $(a, \pi_p^e) = (0.1, 0.7)$ seem to be good choices across the datasets. Compared to the other two losses, we found the modified Huber loss to be robust to the values of π_p^e.

Sensitivity analysis for a We run shallow rAD (with fixed $\pi_p^e = 0.8$) on the 30 trials of each dataset for $a \in \{0.3, 0.7, 0.9\}$. The results are reported in Table 4 for the 9 benchmark datasets and the results of the 17

Table 3: AUC means of shallow rAD over 30 trials for different π_p^e. The significant changes in the AUC means are highlighted in bold.

Dataset	square/π_p^e				hinge/π_p^e				m-Huber/π_p^e			
	$1-\pi_n$	0.9	0.7	0.6	$1-\pi_n$	0.9	0.7	0.6	$1-\pi_n$	0.9	0.7	0.6
pendigits	0.96	0.98	0.98	0.98	0.94	0.98	0.98	0.98	0.97	0.98	0.98	0.98
mammography	0.90	0.91	0.91	0.91	0.90	0.91	0.91	0.91	0.90	0.91	0.91	0.91
optdigits	0.96	0.99	0.997	0.997	**0.93**	0.99	0.997	0.997	0.98	0.996	0.998	0.998
Stamps	0.80	0.80	0.82	0.82	0.81	0.81	0.81	0.80	0.80	0.80	0.80	0.80
cardio	0.91	0.91	0.92	0.92	0.87	0.88	0.88	0.89	0.92	0.93	0.94	0.94
InternetAds	0.77	0.77	0.70	**0.60**	0.86	0.85	0.86	0.86	0.87	0.87	0.86	0.86
Cardiotocography	0.89	0.88	0.89	0.89	0.87	0.85	0.87	0.87	0.90	0.90	0.90	0.90
magic.gamma	0.78	0.77	0.78	0.78	0.78	0.77	0.78	0.78	0.78	0.78	0.78	0.78
SpamBase	0.94	0.94	0.94	0.94	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.94

Table 4: AUC means of shallow rAD over 30 trials for different a. The significant changes in the AUC means are highlighted in bold.

Dataset	square/a			hinge/a			m-Huber/a		
	0.3	0.7	0.9	0.3	0.7	0.9	0.3	0.7	0.9
pendigits	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
mammography	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
optdigits	0.997	0.995	0.99	0.996	0.995	0.99	0.997	0.996	0.99
Stamps	0.83	0.82	0.82	0.81	0.82	0.81	0.80	0.81	0.81
cardio	0.92	0.91	0.91	0.88	0.87	0.85	0.93	0.93	0.93
InternetAds	0.79	**0.69**	**0.62**	0.87	0.85	**0.77**	0.83	**0.71**	**0.65**
Cardiotocography	0.87	0.87	0.87	0.86	0.85	0.83	0.90	0.89	0.88
magic.gamma	0.78	0.78	0.78	0.78	0.78	0.77	0.78	0.78	0.78
SpamBase	0.94	0.94	0.93	0.94	0.94	0.93	0.94	0.94	0.94

remaining datasets are given in Table 7 in in the supp. material. From Table 2, 5, 4 and 7, we can see that the AUC means do not decrease significantly when we increase a (except for the dataset InternetAds). Hence, shallow rAD with $\pi_p^e = 0.8$ is also robust to different values of a.

B. Experiments with deep rAD

Baseline methods and implementation We compare deep rAD with the deep semi-supervised AD method (deep SAD) Ruff et al. (2020)) and the PU learning method with nonnegative risk estimator and sigmoid loss (nnPU) Kiryo et al. (2017)). For deep SAD and nnPU, we use default hyperparameter settings and network architectures as in their original implementation by the authors. For deep rAD, we use the same network architectures as deep SAD and we use ADAM to solve the optimization problem in (19). We implement 4 losses for deep rAD: squared loss, sigmoid loss, logistic loss, and modified Huber loss. We set $a = 0.1$ and $\pi_p^e = 0.8$ (thus $\pi_n^e = 0.2$) as default values for deep rAD.

Datasets We test the algorithms on 3 benchmark k-classes-out datasets: MNIST, Fashion-MNIST, and CIFAR-10 (all have 10 classes). We use AD setups following previous works Chalapathy et al. (2018); Ruff et al. (2020): for each $\pi_n \in \{0.01, 0.05, 0.1, 0.2\}$, we set one of the ten classes to be a positive class, letting the remaining nine classes be anomalies and maintaining the ratio between normal instances and anomaly instances such that the setup has the required π_n (so we have 10 setups corresponding to 10 classes). We note that the anomalous data in our generation process can originate from more than one of the nine classes (unlike in the setup of deep SAD where the anomaly is only from one of the nine classes). For each π_n, we repeat this generation process 2 times to get 20 AD setups (or 20 trials). Then, in each trial, we randomly choose γ_l (with $\gamma_l \in \{0.05, 0.1, 0.2\}$) portion of the train data to be labeled and keep the remaining $(1-\gamma_l)$

portion as unlabeled data. Note that we make the labeled data for nnPU only from normal instances. To make labeled data for deep rAD and deep SAD, $(1 - \pi_n)$ portion is taken from the nnPU labeled data (which contain only normal instances), and the remaining π_n portion is taken from the anomalous instances. Hence, the number of labeled anomalous instances for deep rAD and deep SAD is about $(\gamma_l \times \pi_n)$ portion of the train data.

Experiment results In Figure 1, we report the mean and standard deviation (std) of the AUC over 20 trials on the datasets with increasing pollution ratio π_n and default $\gamma_l = 0.05$. The results for $\gamma_l \in \{0.1, 0.2\}$ are given in Figures 4 and 5 in the supp. material. Figures 1, 4 and 5 show that, on average, deep rAD methods provide better AUC than deep SAD and nnPU on CIFAR-10. The AUC difference is significant when π_n is increased (deep rAD and deep SAD have similar performance when $\pi_n = 0.01$). On FMNIST, deep rAD methods are still better than the others when π_n is increased but the AUC improvement is small. On MNIST, deep SAD is best, and deep rAD catch up with it when either π_n or γ_l is increased. Deep rAD with quadratic loss underperforms the other rAD methods on MNIST and FMNIST. On average, deep rAD with logistic loss performs best among the rAD methods. It is also interesting to note that in the presence of anomalies from multiple classes, the performance of deep SAD degrades over the performance reported in Ruff et al. (2020). The degradation is more severe for CIFAR-10.

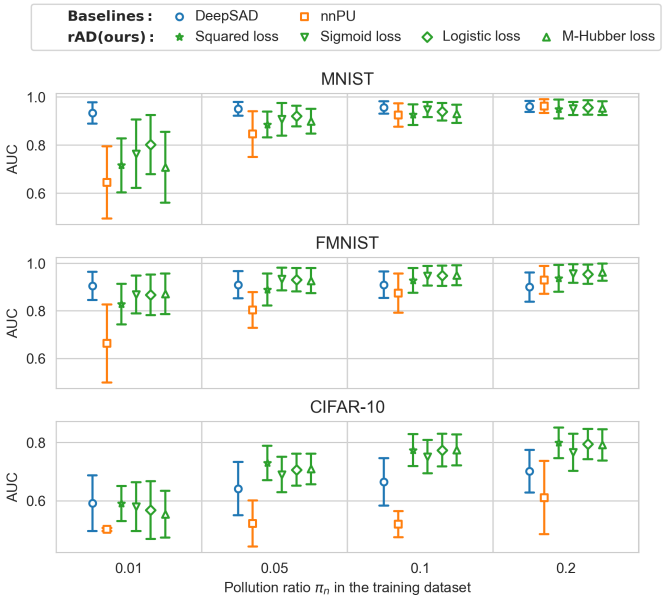

Figure 1: AUC mean and std over 20 trials with various π_n and default $\gamma_l = 0.05$

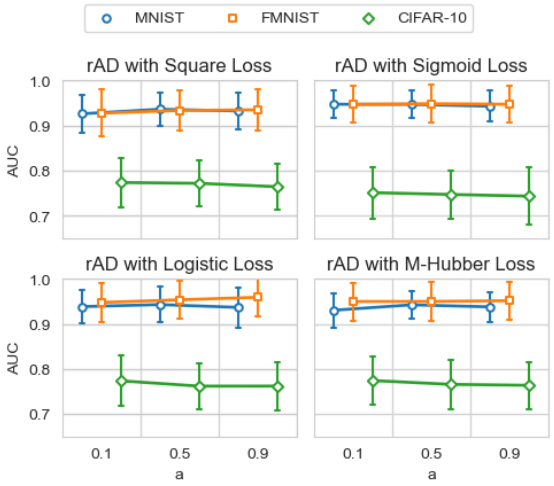

Figure 2: AUC mean and std over 20 trials at various a for the datasets with $\gamma_l = 0.05$ and $\pi_n = 0.1$

To observe the impact of the amount of labeled data, we report the results for the datasets with $\pi_n = 0.1$ and $\gamma_l \in \{0.05, 0.1, 0.2\}$ in Figure 3. We observe that all the methods improve when we increase γ_l. From $\gamma_l = 0.05$ to $\gamma_l = 0.1$ (i.e., 5% more labeled data), deep rAD methods show a significant improvement in performance.

Sensitivity analysis for π_p^e We run deep rAD with $a = 0.1$ on the 20 trials of each dataset for $\pi_p^e \in \{1 - \pi_n, 0.9, 0.7, \pi_n\}$ (when $\pi_p^e = 1 - \pi_n$, it is an exact estimation of π_p, and when $\pi_p^e = \pi_n$, we can say π_p^e is a bad estimation of π_p). We report the result in Table 8 in the supp. material. Again, we see that π_p^e is not necessarily a precise estimation of π_p; and $(a, \pi_p^e) = (0.1, 0.8)$ or $(a, \pi_p^e) = (0.1, 0.7)$ are good settings. These results are consistent with the results of shallow rAD.

Sensitivity analysis for a We fix $\pi_p^e = 0.8$ and run deep rAD with additional values of $a \in \{0.5, 0.9\}$ ($a = 0.1$ is the default setting). We report the results for the datasets with $\pi_n = 0.1$ and $\gamma_l = 0.05$ in Figure 2. The results for the datasets with other values of π_n and γ_l are given in the supp. material. We observe that

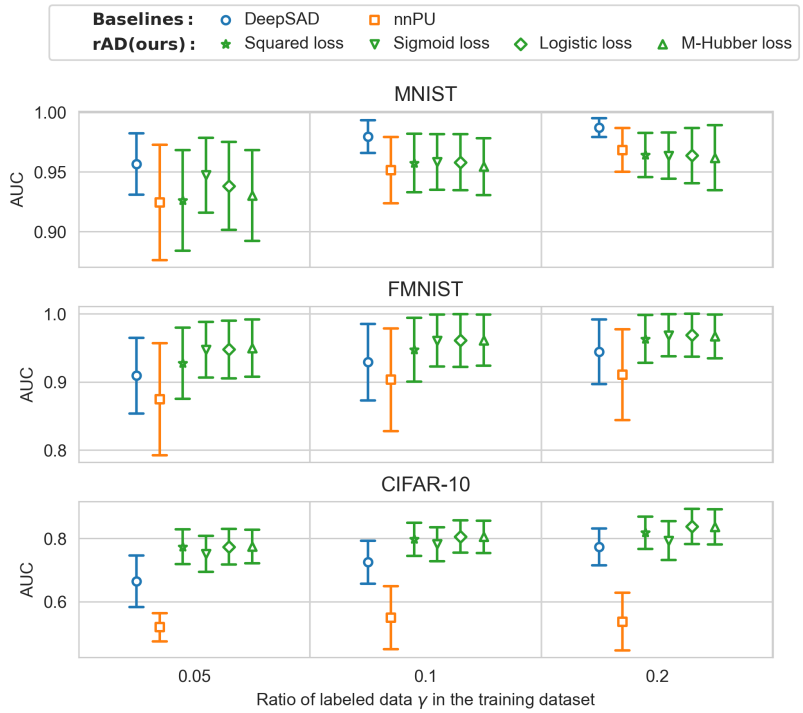

Figure 3: AUC mean and std over 20 trials at various γ_l for fixing $\pi_n = 0.1$

on CIFAR-10, AUC decreases when a is increased; however, the difference is not significant. On FMNIST and MNIST, deep rAD with $\pi_p^e = 0.8$ is quite robust to the change of a.

7 Limitations

On the implementation side, although the experiments have shown that our rAD methods are quite robust to the changes of the parameters a and π_p^e, we still have to tune them to obtain the best AD performance. Furthermore, solving the optimization problem in (19) is challenging for very large-scale dataset since the max operator does not allow parallel computations. On the theoretical side, although the risk bounds are established for the proposed risk minimizers in Section 5, we still need the assumption that π_p and π_n are known in advance, which is a limitation.

8 Conclusion

With semi-supervised classification based on risk estimators, we have introduced a shallow AD method equipped with suitable regularization as well as a deep AD method. Theoretically, we have established the estimation error bounds and the excess risk bounds for the two risk minimizers. Empirically, the shallow AD methods show significant improvement over the baseline methods while the deep AD methods compete favorably with the baselines. Let us conclude the paper by giving some possible future research directions that address the limitation given in Section 7. One possible research direction is to develop a method that can learn the best combination of (a, π_p^e) from the available data. On the other hand, our experiments have shown that using $a = 0.1$, precise estimation of π_p and π_n are not necessarily needed to obtain good accuracy in terms of AUC. Hence, another possible research direction would be to study the theoretical bounds of the risk minimizers with π_p and π_n replaced by some estimates. Finally, investigating effective optimization techniques to tackle the nonconvex Problem (19) is also an important research direction aimed at overcoming the difficulties associated with handling exceedingly large-scale datasets.

References

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. doi: 10.1198/016214505000000907.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109: 719–760, 2020.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(99):2973–3009, 2010. URL `http://jmlr.org/papers/v11/blanchard10a.html`.

Raghavendra Chalapathy, Aditya Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. 02 2018. http://arxiv.org/abs/1802.06360.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection. *ACM Computing Surveys*, 41(3): 1–58, 7 2009. ISSN 0360-0300. doi: 10.1145/1541880.1541882. URL `https://dl.acm.org/doi/10.1145/1541880.1541882`.

Sneha Chaudhari and Shirish Shevade. Learning from positive and unlabelled examples using maximum margin clustering. In *Proceedings of the 19th International Conference on Neural Information Processing - Volume Part III*, ICONIP'12, pp. 465–473, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642344862. doi: 10.1007/978-3-642-34487-9_56. URL `https://doi.org/10.1007/978-3-642-34487-9_56`.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 551–556, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138881. doi: 10.1145/1014052.1014118. URL `https://doi.org/10.1145/1014052.1014118`.

Marthinus du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural networks : the official journal of the International Neural Network Society*, 50C:110–119, 11 2013. doi: 10.1016/j.neunet.2013.11.010.

Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/file/35051070e572e47d2c26c241ab88307f-Paper.pdf`.

Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1386–1394, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/plessis15.html`.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 213–220, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401920. URL `https://doi.org/10.1145/1401890.1401920`.

Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, 58:121–134, 10 2016. ISSN 0031-3203. doi: 10.1016/J.PATCOG.2016.03.028.

Nico Görnitz, Marius Kloft, and Ulf Brefeld. Active and semi-supervised data domain description. In *Lect Notes Artif Intell.*, volume 5781, pp. 407–422, 09 2009. ISBN 978-3-642-04179-2. doi: 10.1007/978-3-642-04180-8_44.

Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 46:235–262, 2013.

Nico Görnitz, Luiz Alberto Lima, Klaus-Robert Müller, Marius Kloft, and Shinichi Nakajima. Support vector data descriptions and k -means clustering: One class? *IEEE Transactions on Neural Networks and Learning Systems*, 29(9):3994–4006, 2018. doi: 10.1109/TNNLS.2017.2737941.

Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Neural Information Processing Systems (NeurIPS)*, 2022.

D. M. Hawkins. *Identification of outliers*. Monographs on applied probability and statistics. Chapman and Hall, London [u.a.], 1980. ISBN 041221900X. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+02435757X&sourceid=fbw_bibsonomy.

Douglas M. Hawkins. The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, 69(346):340–344, 1974. ISSN 01621459.

Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter. Outlier detection using replicator neural networks. In Yahiko Kambayashi, Werner Winiwarter, and Masatoshi Arikawa (eds.), *Data Warehousing and Knowledge Discovery*, pp. 170–180, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-46145-6.

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 10 2004. doi: 10.1023/B:AIRE.0000045502.10941.a9.

Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from positive, unlabeled and biased negative data. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2820–2829. PMLR, 09–15 Jun 2019.

Ling Huang, XuanLong Nguyen, Minos Garofalakis, Michael Jordan, Anthony Joseph, and Nina Taft. In-network pca and anomaly detection. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/file/2227d753dc18505031869d44673728e2-Paper.pdf.

Hyunjun Ju, Dongha Lee, Junyoung Hwang, Junghyun Namkung, and Hwanjo Yu. Pumad: Pu metric learning for anomaly detection. *Information Sciences*, 523:167–183, 2020. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2020.03.021. URL https://www.sciencedirect.com/science/article/pii/S0020025520302012.

Shehroz S. Khan and Michael G. Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374, 2014. doi: 10.1017/S026988891300043X.

Bangalore Kiran, Dilip Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4, 01 2018. doi: 10.3390/jimaging4020036.

Ryuichi Kiryo, Gang Niu, Marthinus D du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

Yann Lecun, Sumit Chopra, and Raia Hadsell. *A tutorial on energy-based learning*. MIT Press, 01 2006.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991. ISBN 978-3-642-20212-4. doi: 10.1007/978-3-642-20212-4_6. URL https://doi.org/10.1007/978-3-642-20212-4_6.

Wee Sun Lee and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 448–455. AAAI Press, 2003. ISBN 1577351894.

Dan Li, Dacheng Chen, Lei Shi, Baihong Jin, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *International Conference on Artificial Neural Networks*, 2019.

Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03): 2003; Acapulco, Mexico*, pp. 587–594, 01 2003.

B. Liu, Y. Dai, X. Li, W.S. Lee, and P.S. Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pp. 179–186, 2003. doi: 10.1109/ICDM.2003.1250918.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, 1989.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018. ISBN 978-0-262-03940-6.

Jordi Munoz-Mari, Francesca Bovolo, Luis Gomez-Chova, Lorenzo Bruzzone, and Gustavo Camp-Valls. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(8):3188–3197, 2010. doi: 10.1109/TGRS.2010.2045764.

Gang Niu, Marthinus C. du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *NIPS'16*, pp. 1207–1215, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2), 7 2020. doi: 10.1145/3439950. URL `http://arxiv.org/abs/2007.02500http://dx.doi.org/10.1145/3439950`.

Emanuel Parzen. On Estimation of a Probability Density Function and Mode on JSTOR. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2013.12.026. URL `https://www.sciencedirect.com/science/article/pii/S016516841300515X`.

R. Pincus. Barnett, v., and lewis t.: Outliers in statistical data. 3rd edition. j. wiley & sons 1994, xvii. 582 pp., £49.95. *Biometrical Journal*, 37(2):256–256, 1995. doi: https://doi.org/10.1002/bimj.4710370219. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.4710370219`.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022. doi: 10.48550/arxiv.2106.08265. URL `https://arxiv.org/abs/2106.08265v2`.

Peter J Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(37):283–297, 1985.

Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999. ISSN 15372723. doi: 10.1080/00401706.1999.10485670. URL `https://www.tandfonline.com/action/journalInformation?journalCode=utch20`.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402. PMLR, 10–15 Jul 2018.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HkgH0TEYwH`.

Lukas Ruff, Jacob Kauffmann, Robert Vandermeulen, Gregoire Montavon, Wojciech Samek, Marius Kloft, Thomas Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, PP:1–40, 02 2021. doi: 10.1109/JPROC.2021.3052449.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural computation*, 14:21–41, 02 2002. doi: 10.1162/089976602753284446.

Tomoya Sakai, Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2998–3006. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/sakai17a.html`.

Yehuda Salu. Models of neural novelty detectors, with similarities to cerebral cortex. *Biosystems*, 21 (2):99–113, 1988. ISSN 0303-2647. doi: https://doi.org/10.1016/0303-2647(88)90003-2. URL `https://www.sciencedirect.com/science/article/pii/0303264788900032`.

Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola, and Robert Williamson. Estimating support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 07 2001. doi: 10.1162/089976601750264965.

David M.J. Tax and Robert P.W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20 (11-13):1191–1199, 11 1999. ISSN 0167-8655. doi: 10.1016/S0167-8655(99)00087-2.

David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54:45–66, 2004. doi: 10.1023/B:MACH.0000008084.60811.49.

Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8200–8210, 2019. doi: 10.1109/ICCV.2019.00829.

Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning Semantic Context from Normal Samples for Unsupervised Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3110–3118, 5 2021. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16420`.

Arthur Zimek, Erich Schubert, and Peer Kröger. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5:363–387, 10 2012. doi: 10.1002/sam.11161.