Towards Calibrated Losses for Adversarial Robust Reject Option Classification

Vrund Shah VRUND.SHAH@RESEARCH.IIIT.AC.IN
Tejas Chaudhari TEJAS.CHAUDHARI@RESEARCH.IIIT.AC.IN
Naresh Manwani NARESH.MANWANI@IIIT.AC.IN
Machine Learning Lab, Kohli Research Block, IIIT Hyderabad - 500032, India

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Robustness towards adversarial attacks is a vital property for classifiers in several applications such as autonomous driving, medical diagnosis, etc. Also, in such scenarios, where the cost of misclassification is very high, knowing when to abstain from prediction becomes crucial. A natural question is which surrogates can be used to ensure learning in scenarios where the input points are adversarially perturbed and the classifier can abstain from prediction? This paper aims to characterize and design surrogates calibrated in "Adversarial Robust Reject Option" setting. First, we propose an adversarial robust reject option loss ℓ_d^γ and analyze it for the hypothesis set of linear classifiers ($\mathcal{H}_{\text{lin}}$). Next, we provide a complete characterization result for any surrogate to be $(\ell_d^\gamma, \mathcal{H}_{\text{lin}})$- calibrated. To demonstrate the difficulty in designing surrogates to ℓ_d^γ, we show negative calibration results for convex surrogates and quasi-concave conditional risk cases (these gave positive calibration in adversarial setting without reject option). We also empirically argue that Shifted Double Ramp Loss (DRL) and Shifted Double Sigmoid Loss (DSL) satisfy the calibration conditions. Finally, we demonstrate the robustness of shifted DRL and shifted DSL against adversarial perturbations on a synthetically generated dataset.

Keywords: Calibrated Surrogates, Reject Option Classification , Adversarial Robustness

1. Introduction

Many machine learning models are susceptible to adversarial attacks (Goodfellow et al., 2014; Szegedy et al., 2013), i.e., imperceivable changes in the data at testing time results in learning of bad classifiers and even hazardous accidents. For example, the presence of an artifact on a traffic sign may lead to inaccurate interpretation of the signal, often arising in autonomous driving. To address such problems, several studies were conducted for learning classifiers with reduced sensitivity to these perturbations (Raghunathan et al., 2018; Wong and Kolter, 2018). The property displayed by these classifiers is "Adversarial Robustness".

To achieve robustness against adversarial attacks, the worst-case loss subject to adversarial perturbations is used. Adversarial robustness to small l_p-norm perturbations has been analyzed in Carlini and Wagner (2017); Madry et al. (2018). Optimization of adversarial loss is hard, and this calls for the use of surrogates. An important property that the surrogates should satisfy is "Consistency" - i.e., minimization of the true risk associated with surrogate loss should lead to the minimization of the true risk associated with target loss. One way to analyze consistency is using "calibration" - point-wise minimization of the conditional risk (Bartlett et al., 2006; Steinwart, 2007). Bao et al. (2020) showed that

convex surrogates are not $\mathcal{H}_{\text{lin}}$-calibrated in the adversarial setting for binary classification. Awasthi et al. (2021a,b) extended calibration results in an adversarial setting to other function classes such as generalized linear models and single-layer ReLU neural networks.

This paper studies the binary classifiers with reject option robust to adversarial perturbations. In high-risk environments (such as medical diagnosis, finance, etc.), the ability to abstain from prediction and incur small costs for it is beneficial as compared to the cost of misclassification. Classifiers with ability to abstain are called "reject option" classifiers (Chow, 1970; Cortes et al., 2016; Ni et al., 2019). While many studies address adversarial robustness in standard classification setting, no attention has been given to the adversarial robustness of reject option classifiers. In Kato et al. (2020); Chen et al. (2023), authors propose an adversarial robust approach for reject option classification and validate it empirically. However, there is no attempt to analyze calibration for this domain.

We analyze calibrated surrogates in a scenario where the inputs are adversarially perturbed, and the classification has an embedded reject option. To our knowledge, this is the foundational work for this domain from the standpoint of calibration analysis.

Key Contributions

- We completely characterize surrogates, which are calibrated in the "adversarial-reject option setting" for linear classifiers. We prove that convex loss functions can not be calibrated surrogates in such scenarios. We also show that surrogate losses with quasi-concave conditional risk are not calibrated.
- We propose that Shifted Double Sigmoid Loss (DSL) and Shifted Double Ramp Loss (DRL) are potential candidates for calibrated surrogates and empirically show that they satisfy calibration conditions.
- We describe the adversarial training procedure using proposed loss functions. We experimentally validate our findings on a synthetic dataset that Shifted DSL and Shifted DRL exhibit robustness against adversarial perturbations.

2. Related Work

2.1. Calibrated Surrogates towards Robust Adversarial Classification

Adversarial robust loss for binary classification is $\max_{\tilde{\mathbf{x}}\in\mathcal{U}(\mathbf{x})} \ell(f(\tilde{\mathbf{x}}), y)$ where ℓ is a loss function and $\mathcal{U}(\mathbf{x})$ is an uncertainty set around $\mathbf{x}$. For $\mathcal{U}(\mathbf{x}) = B_2(\mathbf{x}, \gamma)$, Bao et al. (2020) show that convex surrogates are negatively calibrated to the adversarial robust loss for linear classifiers ($\mathcal{H}_{\text{lin}}$) and gave positive calibration results by introducing the quasi-concavity assumption on the conditional risk. Awasthi et al. (2021a,b) extended calibration results for generalized linear models and single-layer ReLU neural network. Meunier et al. (2022) show that shifted odd losses are calibrated to the adversarial robust loss.

2.2. Robustness towards Adversarial Attacks

Adversarial training, a popular adversarial defense approach, involves adding adversarial examples into the training set. Fast gradient sign method (Goodfellow et al., 2014) for l_∞-norm perturbations, projected gradient descent (Madry et al., 2018) are some popular

algorithms based on adversarial training. The approach in Yang et al. (2019) is based on detecting adversarial examples using distance-based approaches or by feature attrition. To fool the gradient-based adversarial attack methods, Papernot and McDaniel (2018) add discrete or non-differentiable components into the model. Zhang et al. (2022) investigates certified l_∞ robustness from the lens of representing Boolean functions. Randomized smoothing (Cohen et al., 2019) is a technique to convert any classifier that classifies well under Gaussian noise into a new classifier that is certifiably robust to adversarial perturbations.

2.3. Reject Option Classification

Chow (1970) was the seminal work to deal with this approach to classification problems. Generalized hinge loss (Bartlett and Wegkamp, 2008), double ramp loss (Manwani et al., 2013; Shah and Manwani, 2018), max-hinge loss and plus-hinge loss (Cortes et al., 2016) etc. are some of the consistent loss functions for learning with rejection in binary setting. Algorithms using these loss functions are support vector machine (SVM) variants for reject option classifiers. Shah and Manwani (2020); Kalra et al. (2021) proposed a new consistent loss function called double sigmoid loss function for binary reject option classifier. Ramaswamy et al. (2018); Ni et al. (2019); Cao et al. (2022) deal with calibration and consistency in the multiclass reject option classification.

3. Preliminaries

3.1. Notations

For a vector $\mathbf{x} \in \mathbb{R}^d$, let $\|\mathbf{x}\|_p$ denote the l_p-norm. $B_p(\mathbf{x}, r) \stackrel{\text{def}}{=} \{\mathbf{v} \in \mathbb{R}^d \mid \|\mathbf{v} - \mathbf{x}\|_p \leq r\}$ be the d-dimensional closed l_p-ball centered at $\mathbf{x}$ with radius r. The set $\{1, \ldots, n\}$ is denoted by $[n]$. The indicator function corresponding to an event A is denoted by $\mathbb{1}_{\{A\}}$.

3.2. Binary Classification Problem

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the instance space and $\mathcal{Y} = \{1, -1\}$ be the label space. Let $\mathcal{P}$ be a fixed but unknown probability distribution over $\mathcal{X} \times \mathcal{Y}$ from which i.i.d. samples of $(\mathbf{x}, y)$ are drawn. The objective of the classification problem is to learn a function $f : \mathcal{X} \to \mathbb{R}$. For any example, the class label is predicted as $\hat{y} = \text{sign}(f(\mathbf{x}))$. Loss ℓ_{01} captures the difference between the predicted label and the true label, where $\ell_{01}(yf(\mathbf{x})) = \mathbb{1}_{yf(\mathbf{x})\leq 0}$. The objective of the learning algorithm is to find a classifier f^* in the function class $\mathcal{H}$ which minimizes the risk function $\mathcal{R}_{\ell_{01}}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell_{01}(yf(\mathbf{x}))]$. The risk $\mathcal{R}_{\ell_{01}}(f)$ is minimized by Bayes classifier $f^*(\mathbf{x}) = \eta - \frac{1}{2}$, where $\eta = P(Y = 1|\mathbf{x})$. In practice, we use surrogates of ℓ_{01} loss which are easy to optimize. The true risk of a classifier f for a surrogate loss $\ell(yf(\mathbf{x}))$ is $\mathcal{R}_\ell(f) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell(yf(\mathbf{x}))]$. The Bayes $(\ell, \mathcal{H})$-risk is defined by $\mathcal{R}^*_{\ell,\mathcal{H}} = \inf_{f\in\mathcal{H}} \mathcal{R}_{\ell,\mathcal{H}}(f)$.

3.3. Consistency of Surrogate Loss Functions

Definition 1 ***$\mathcal{H}$-Consistency*** *: For a given hypothesis set $\mathcal{H}$ and a target loss function ℓ_1, a surrogate ℓ_2 is said to be $\mathcal{H}$-consistent with respect to ℓ_1, if the following holds:*

$$\mathcal{R}_{\ell_2}(f_n) - \mathcal{R}^*_{\ell_2,\mathcal{H}} \stackrel{n\to+\infty}{\longrightarrow} 0 \Longrightarrow \mathcal{R}_{\ell_1}(f_n) - \mathcal{R}^*_{\ell_1,\mathcal{H}} \stackrel{n\to+\infty}{\longrightarrow} 0 \tag{1}$$

for all probability distributions and sequences of $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$.

Using conditional expectation property, $\mathcal{R}_\ell(f)$ can be written as $\mathcal{R}_\ell(f) = \mathbb{E}_X[\mathcal{C}_{\ell,\mathcal{H}}(f(\mathbf{x}), \eta)]$, where $\eta(\mathbf{x}) = P(y = 1|\mathbf{x})$[1] and

$$\mathcal{C}_{\ell,\mathcal{H}}(f(\mathbf{x}), \eta) = \mathbb{E}_{y|\mathbf{x}}[\ell(yf(\mathbf{x})) \mid \mathbf{x}] = \eta\ \ell(f(\mathbf{x})) + (1-\eta)\ \ell(-f(\mathbf{x})). \tag{2}$$

The minimal conditional risk $\mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x}, \eta)$ (Steinwart, 2007) and pseudo-minimal conditional risk $\mathcal{C}^*_{\ell,\mathcal{H}}(\eta)$ (Bao et al., 2020) are defined as

$$\mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x}, \eta) = \inf_{f\in\mathcal{H}} \mathcal{C}_{\ell,\mathcal{H}}(f(\mathbf{x}), \eta) \quad \text{and} \quad \mathcal{C}^*_{\ell,\mathcal{H}}(\eta) = \inf_{f\in\mathcal{H},\mathbf{x}\in\mathcal{X}} \mathcal{C}_{\ell,\mathcal{H}}(f(\mathbf{x}), \eta) \tag{3}$$

respectively. The corresponding excess-conditional risk is defined as :

$$\Delta\mathcal{C}_{\ell,\mathcal{H}}(f(\mathbf{x}), \eta) = \mathcal{C}_{\ell,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x}, \eta) \tag{4}$$

3.4. Calibration Theory

Definition 2 ***Uniform $\mathcal{H}$-Calibration*** *[Definition 2.15 in (Steinwart, 2007)] : For a given hypothesis set $\mathcal{H}$ and a target loss function ℓ_1, a surrogate ℓ_2 is said to be uniformly $\mathcal{H}$-calibrated with respect to ℓ_1 if, for any $\epsilon > 0$, there exists $\delta > 0$ such that for all $\eta \in [0, 1]$, $f \in \mathcal{H}, \mathbf{x} \in \mathcal{X}$, we have $\mathcal{C}_{\ell_2,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\mathbf{x}, \eta) < \delta \implies \mathcal{C}_{\ell_1,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell_1,\mathcal{H}}(\mathbf{x}, \eta) < \epsilon$.*

Definition 3 ***Uniform Pseudo- $\mathcal{H}$- Calibration*** *(Bao et al., 2020) : For a given hypothesis set $\mathcal{H}$ and a target loss function ℓ_1, a surrogate loss function ℓ_2 is uniformly pseudo-$\mathcal{H}$-calibrated with respect to a ℓ_1 if, for any $\epsilon > 0$, there exists $\delta > 0$ such that for all $\eta \in [0, 1]$ and $f \in \mathcal{H}, \mathbf{x} \in \mathcal{X}$, we have $\mathcal{C}_{\ell_2,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\eta) < \delta \implies \mathcal{C}_{\ell_1,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell_1,\mathcal{H}}(\eta) < \epsilon$.*

Definition 4 ***Uniform Calibration function*** *(Steinwart, 2007) : Given a hypothesis set $\mathcal{H}$, we define the uniform calibration function δ and uniform pseudo-calibration function $\hat{\delta}$ for a pair of losses (ℓ_1, ℓ_2) as follows: for any $\epsilon > 0$*

$$\delta(\epsilon) = \inf_{\eta\in[0,1]} \inf_{f\in\mathcal{H},\mathbf{x}\in\mathcal{X}} \left\{\mathcal{C}_{\ell_2,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\mathbf{x}, \eta) \mid \mathcal{C}_{\ell_1,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell_1,\mathcal{H}}(\mathbf{x}, \eta) \geq \epsilon\right\}$$

$$\hat{\delta}(\epsilon) = \inf_{\eta\in[0,1]} \inf_{f\in\mathcal{H},\mathbf{x}\in\mathcal{X}} \left\{\mathcal{C}_{\ell_2,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\eta) \mid \mathcal{C}_{\ell_1,\mathcal{H}}(f(\mathbf{x}), \eta) - \mathcal{C}^*_{\ell_1,\mathcal{H}}(\eta) \geq \epsilon\right\}.$$

Proposition 5 *[Lemma 2.16 in (Steinwart, 2007)] Given a hypothesis set $\mathcal{H}$, loss ℓ_2 is uniformly $\mathcal{H}$-calibrated (or uniformly pseudo-$\mathcal{H}$-calibrated) with respect to ℓ_1 if and only if its calibration function δ satisfies $\delta(\epsilon) > 0$ (resp. its uniform pseudo-calibration function $\hat{\delta}$ satisfies $\hat{\delta}(\epsilon) > 0$) for all $\epsilon > 0$.*

For the standard binary classification problem, when $\mathcal{H} = \mathcal{H}_{all}$, Bartlett et al. (2006) showed that calibration is both necessary and sufficient for consistency. However, when we restrict ourselves to a hypothesis set ($\mathcal{H} \neq \mathcal{H}_{all}$), then calibration is a necessary but not sufficient condition for consistency. Steinwart (2007) showed that if the loss functions satisfy an additional criteria of $\mathcal{P}$-minimizability, then point-wise minimization of conditional risk (calibration) does yield minimization of true risk (consistency).

1. For the rest of the paper, we adopt the notation as η for $\eta(\mathbf{x})$.

3.5. Calibration with Adversarial Robustness

In the scenario when inputs are adversarially perturbed, designing surrogates that are calibrated and consistent becomes difficult. Bao et al. (2020) showed that convex losses are not calibrated when the hypothesis set consists of linear classifiers. The reason is that the convexity assumption on the surrogate results in the minimizer of the conditional risk being close to the origin (i.e., the non-robust region) for the case when the posterior probability for both classes is close to half ($\eta \approx 0.5$). The robust loss in that case was defined as $\phi_\gamma(f(\mathbf{x})) = \mathbb{1}_{\{yf(\mathbf{x})\leq\gamma\}}$. Negative Calibration result is stated as follows :

Theorem 6 *(Bao et al., 2020) For any margin-based surrogate loss $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$, if ℓ is convex, then ℓ is not pseudo-calibrated wrt $(\phi_\gamma, \mathcal{H}_{lin})$.*

3.6. Reject Option Classifier

A confidence-based binary reject option classifier (Cortes et al., 2016) is comprised of a function $f : \mathcal{X} \to \mathbb{R}$ and a confidence parameter $\rho \in \mathbb{R}_+$. The confidence-based reject option classifier is defined as $h(f(\mathbf{x}), \rho) = 1\ \mathbb{1}_{\{f(\mathbf{x})>\rho\}} + \perp\ \mathbb{1}_{\{|f(\mathbf{x})|\leq\rho\}} + -1\ \mathbb{1}_{\{f(\mathbf{x})<-\rho\}}$. The Reject option loss ℓ_d for the **Confidence-based Rejection model** is defined as :

$$\ell_d(yf(\mathbf{x}), \rho) = \mathbb{1}_{\{yf(\mathbf{x})\leq\text{-}\rho\}} + d\ \mathbb{1}_{\{|f(\mathbf{x})|\leq\rho\}} \tag{5}$$

where $d \in (0, 0.5)$ is the cost of rejection. The generalized Bayes classifier (Chow, 1970) is defined as $f_d^*(\mathbf{x}) = \mathbb{1}_{\{\eta(\mathbf{x})>1-d\}} + \perp\ \mathbb{1}_{\{d\leq\eta(\mathbf{x})\leq 1-d\}} - \mathbb{1}_{\{\eta(\mathbf{x})<d\}}$.

4. Proposed Work: Calibration in the Adversarial Robust Reject Option Setting

In our analysis, we assume $\mathcal{X} = B_2(\mathbf{0}, 1)$ (l_2 unit ball centered at origin). To start, we rewrite loss ℓ_d (see eq. (5)) as a convex combination of two indicator functions as follows.

$$\ell_d(yf(\mathbf{x}), \rho) = (1-d)\ \mathbb{1}_{\{yf(\mathbf{x})<\text{-}\rho\}} + d\ \mathbb{1}_{\{yf(\mathbf{x})\leq\rho\}} \tag{6}$$

An adversarial robust loss corresponding to ℓ_d is $\sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|_2\leq\gamma} \ell_d(yf(\mathbf{x}'), \rho)$. However, analysis of this loss is not easy. So, we define a new adversarial robust loss for confidence based reject option classifier.

Definition 7 ***Adversarial Robust Reject Option Loss:*** *Given d, f and ρ, the adversarial reject option loss ℓ_d^γ for $(\mathbf{x}, y)$ is defined as*

$$\ell_d^\gamma(yf(\mathbf{x}), \rho) = (1-d) \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|_2\leq\gamma} \{\mathbb{1}_{\{yf(\mathbf{x}')<-\rho\}}\} + d \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|_2\leq\gamma} \{\mathbb{1}_{\{yf(\mathbf{x}')\leq\rho\}}\} \tag{7}$$

Note that, in $\ell_d^\gamma(yf(\mathbf{x}), \rho)$, we consider l_2 norm perturbations. It is easy to see that $\ell_d^\gamma(yf(\mathbf{x}), \rho) \geq \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} \ell_d(yf(\mathbf{x}'), \rho)$. We use ℓ_d^γ as target loss function. In this paper, we want to characterize the conditions under which surrogate loss functions are calibrated to the target loss function ℓ_d^γ. The class of linear classifiers is defined as $\mathcal{H}_{\text{lin}} = \{\mathbf{x} \to \mathbf{w}\cdot\mathbf{x} \mid \|\mathbf{w}\| = 1\}$. Now onwards, we consider $\mathcal{H} = \mathcal{H}_{\text{lin}}$. For linear classifiers, loss ℓ_d^γ becomes γ-right shift of ℓ_d loss which is shown in the next proposition.

Proposition 8 *The Adversarial Robust Reject Option Loss ℓ_d^γ for the class of linear classifiers is γ-right shift of ℓ_d loss as follows.*

$$\ell_d^\gamma(yf(\mathbf{x}),\rho) = (1-d)\ \mathbb{1}_{\{yf(\mathbf{x})<-\rho+\gamma\ \}} +\ d\ \mathbb{1}_{\{\ yf(\mathbf{x})\leq\ \rho+\gamma\ \}} \tag{8}$$

4.1. Analysis of the Calibration Function

Here, we derive the calibration function for any margin based surrogate ℓ of loss ℓ_d^γ. The inner risk $\mathcal{C}_{\ell_d^\gamma,\mathcal{H}}$ for ℓ_d^γ can be written as

$$\mathcal{C}_{\ell_d^\gamma,\mathcal{H}}(f(\mathbf{x}),\eta) = \eta\ \ell_d^\gamma(f(\mathbf{x}),\rho) + (1-\eta)\ \ell_d^\gamma(-f(\mathbf{x}),\rho) \tag{9}$$

Let $\alpha = f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x}$. Piece-wise definition of the inner-risk for target loss ℓ_d^γ is as follows.

$$\begin{aligned}\mathcal{C}_{\ell_d^\gamma,\mathcal{H}}(\alpha,\eta) = \eta\ \mathbb{1}_{\{\alpha<-\rho-\gamma\}}\ +\ (\eta+(1-\eta)d)\ \mathbb{1}_{\{-\rho-\gamma\leq\alpha<-\rho+\gamma\}}\ +\ d\ \mathbb{1}_{\{-\rho+\gamma\leq\alpha\leq\rho-\gamma\}}\\ +\ (\eta d+(1-\eta))\ \mathbb{I}_{\{\rho-\gamma<\alpha\leq\rho+\gamma\}}\ +\ (1-\eta)\ \mathbb{I}_{\{\rho+\gamma<\alpha\}}\end{aligned} \tag{10}$$

We now find the excess inner risk $\Delta\mathcal{C}_{\ell_d^\gamma,\mathcal{H}}(\alpha,\eta) = \mathcal{C}_{\ell_d^\gamma,\mathcal{H}}(\alpha,\eta) - \mathcal{C}^*_{\ell_d^\gamma,\mathcal{H}}(\alpha,\eta)$ of the target loss ℓ_d^γ. The following lemma gives a case by case expression for the same.

Lemma 9 *The excess-inner risk for target loss ℓ_d^γ is given by*

$$\Delta\mathcal{C}_{\ell_d^\gamma,\mathcal{H}}(\alpha,\eta) = \mathcal{C}_{\ell_d^\gamma,\mathcal{H}}(\alpha,\eta) - \mathcal{C}^*_{\ell_d^\gamma,\mathcal{H}}(\alpha,\eta) =$$

$$\begin{cases}(\eta-d)\ \mathbb{1}_{\min\{\eta,1-\eta\}-d\geq 0} + |2\eta-1|\ \mathbb{1}_{2\eta-1>0}\ \mathbb{1}_{\min\{\eta,1-\eta\}-d<0} & \text{if } \alpha<-\rho-\gamma\\ (1-d)\ \eta\ \mathbb{1}_{\min\{\eta,1-\eta\}-d\geq 0}+ \\ \{(\eta-(1-\eta)(1-d))\ \mathbb{1}_{2\eta-1>0} + (1-\eta)d\ \mathbb{1}_{2\eta-1<0}\}\mathbb{1}_{\min\{\eta,1-\eta\}-d<0} & \text{if } -\rho-\gamma\leq\alpha<-\rho+\gamma\\ \{(d-(1-\eta))\ \mathbb{1}_{2\eta-1>0} + (d-\eta)\ \mathbb{1}_{2\eta-1<0}\}\mathbb{1}_{\min\{\eta,1-\eta\}-d<0} & \text{if } -\rho+\gamma\leq\alpha\leq\rho-\gamma\\ (1-d)\ (1-\eta)\ \mathbb{1}_{\min\{\eta,1-\eta\}-d\geq 0}+ \\ \{\eta\ d\ \mathbb{1}_{2\eta-1>0} + ((1-\eta)-\eta\ (1-d))\ \mathbb{1}_{2\eta-1<0}\}\mathbb{1}_{\min\{\eta,1-\eta\}-d<0} & \text{if } \rho-\gamma<\alpha\leq\rho+\gamma\\ (1-\eta-d)\ \mathbb{1}_{\min\{\eta,1-\eta\}-d\geq 0} + |2\eta-1|\ \mathbb{1}_{2\eta-1<0}\ \mathbb{1}_{\min\{\eta,1-\eta\}-d<0} & \text{if } \rho+\gamma<\alpha\end{cases}$$

Figure 1 illustrates plots of $\Delta\mathcal{C}_{\ell_d^\gamma,\mathcal{H}}$ vs η for $d=0.2$ and $d=0.4$. The vertical lines in violet correspond to $\eta_{\text{left}} = \frac{1-d}{2-d}$ and $\eta_{\text{right}} = \frac{1}{2-d}$. We see that η_{left} is the intersection of the cases $\rho+\gamma<\alpha$ and $-\rho-\gamma\leq\alpha<-\rho+\gamma$. η_{right} is the intersection of the cases $\alpha<-\rho-\gamma$ and $\rho-\gamma<\alpha\leq\rho+\gamma$. Region change precedes definition change when $d<\eta_{\text{left}}$. At $d=\eta_{\text{left}}$, i.e when $d=\frac{3-\sqrt{5}}{2}\ \approx 0.38$, they coincide and for $d>\eta_{\text{left}}$, definition change occurs before region change. The above graphs highlight the cases when $d<\eta_{\text{left}}(d=0.2)$ and $d>\eta_{\text{left}}(d=0.4)$. These were used to develop the idea for splitting the rejection case into further sub-cases. We observe following properties of conditional inner risk and excess risk.

Symmetry Property of $\mathcal{C}_{\ell,\mathcal{H}}(\alpha,\eta)$ and $\Delta\mathcal{C}_{\ell,\mathcal{H}}(\alpha,\eta)$ for margin based loss functions: For any margin based loss function ℓ, using eq.(2), we can see that $\mathcal{C}_{\ell,\mathcal{H}}(\alpha,\eta) = \eta\ \ell(\alpha) + (1-\eta)\ \ell(-\alpha) = \mathcal{C}_{\ell,\mathcal{H}}(-\alpha,1-\eta)$. More specifically, for $\eta=\frac{1}{2}$, we can see that $\mathcal{C}_{\ell,\mathcal{H}}(\alpha,\frac{1}{2}) = \mathcal{C}_{\ell,\mathcal{H}}(-\alpha,\frac{1}{2})$. Thus, we can conclude that $\mathcal{C}_{\ell,\mathcal{H}}(\alpha,\eta)$ is symmetric about $\eta=\frac{1}{2}$. Similarly,

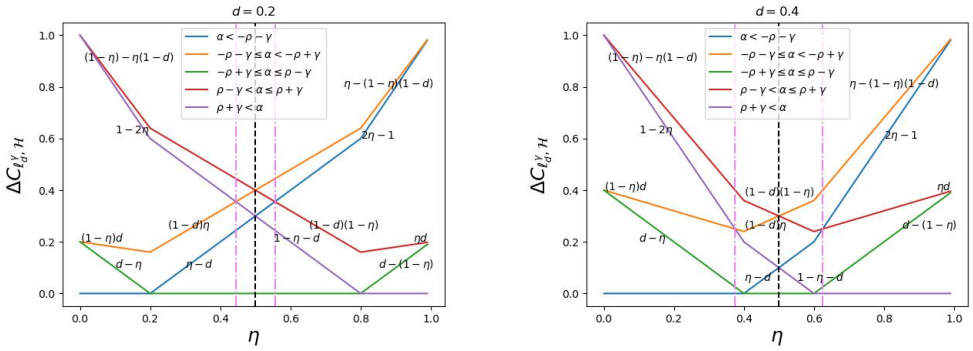


Figure 1: Graph of excess target risk vs η for two different d values.

one can show that $\Delta\mathcal{C}_{\ell,\mathcal{H}}(\alpha,\eta) = \Delta\mathcal{C}_{\ell,\mathcal{H}}(-\alpha, 1-\eta)$. Hence, $\Delta\mathcal{C}_{\ell,\mathcal{H}}(\alpha,\frac{1}{2}) = \Delta\mathcal{C}_{\ell,\mathcal{H}}(-\alpha,\frac{1}{2})$, making $\Delta\mathcal{C}_{\ell,\mathcal{H}}$ symmetric about $\eta = \frac{1}{2}$. Using this, we now characterize calibration of a margin based surrogate ℓ to ℓ_d^γ for linear classifiers.

Theorem 10 *Any margin-based surrogate ℓ is $(\ell_d^\gamma, \mathcal{H})$-calibrated if and only if it satisfies the following :*

$$\inf_{\rho-\gamma<\alpha\leq\|\mathbf{x}\|} \mathcal{C}_{\ell,\mathcal{H}}(\alpha,\frac{1}{2}) > \inf_{0\leq\alpha\leq\|\mathbf{x}\|} \mathcal{C}_{\ell,\mathcal{H}}(\alpha,\frac{1}{2}) \tag{11}$$

$$\inf_{-\|\mathbf{x}\|\leq\alpha\leq\rho+\gamma} \mathcal{C}_{\ell,\mathcal{H}}(\alpha,\eta) > \inf_{-\|\mathbf{x}\|\leq\alpha\leq\|\mathbf{x}\|} \mathcal{C}_{\ell,\mathcal{H}}(\alpha,\eta) \;\; \eta\in(\frac{1}{2},1] \tag{12}$$

Minima Jump Requirement: For $\eta = 0.5$, (11) is the calibration condition which implies that minima should be close to origin, specifically in the interval $[0, \rho-\gamma]$. Calibration to hold for the case of $\eta > 0.5$, (12) should be satisfied implying minima lies beyond $\rho+\gamma$. So, even for a small increase (ξ) in value of η, the minima needs to jump from a region closer to origin to the region lying on the rightmost end of the interval $[-1, 1]$.

4.2. Negative Calibration of Convex Surrogates to ℓ_d^γ

In the adversarial binary classification setting, convex surrogates show negative calibration result (Theorem 6). Reason being that for $\eta = 0.5$, minimizer for the conditional risk falls in the non-robust region $[-\gamma, \gamma]$. However, $\eta = 0.5$ does not yield problems in the adversarial robust reject option case. It is evident from (11) that minima of the conditional risk being closer to origin is in fact needed for calibration (minimizer must lie in $[0, \rho-\gamma]$). For the case $\eta > 0.5$, problems do arise as minimizer is needed at rightmost end of the interval $[-\|\mathbf{x}\|, \|\mathbf{x}\|\,]$. This gives a negative result for calibration as stated in Theorem 11.

Theorem 11 *Let ℓ be a differentiable and convex margin based surrogate to ℓ_d^γ. Then, ℓ is not $(\ell_d^\gamma, \mathcal{H})$-calibrated.*

4.3. Negative Calibration Result When Conditional Risk is Quasi-Concave

Here, we see an important negative result that any surrogate loss of ℓ_d^γ whose conditional risk is quasi-concave, is not calibrated to ℓ_d^γ.

Theorem 12 *No margin-based surrogate ℓ satisfying the property of quasi-concavity of the conditional risk $\mathcal{C}_{\ell,\mathcal{H}}(\alpha,\eta)$ in α, $\forall \eta \in [0,1]$ is $(\ell_d^\gamma, \mathcal{H})$-calibrated.*

This result is in contrast to Bao et al. (2020, Theorem 14) in the context of adversarial binary classification setting (without reject option) which says that any surrogate which has quasi-concave conditional risk is calibrated. The above theorem highlights that results from calibration in adversarial setting doesn't hold straightaway upon adding the reject option. Hence, the extension to calibrated surrogates of ℓ_d^γ is highly non-trivial and challenging.

NOTE: Proof of all the propositions, lemmas and theorems given in this section are provided in the supplementary material.

5. Possible Calibrated Surrogates for ℓ_d^γ

In this section, we present two surrogate loss functions which exhibit the properties required to be $\mathcal{H}$-calibrated to ℓ_d^γ, namely shifted double sigmoid loss and shifted double ramp loss.

5.1. Shifted Double Sigmoid Loss

Double sigmoid loss (DSL) (Shah and Manwani, 2020) was presented in the context of standard reject option classification in the non-adversarial setting. It is shown to be a calibrated surrogate to ℓ_d (Kalra et al., 2021). Double sigmoid loss is defined as :

$$\ell_{\text{ds}}^{\mu}(yf(\mathbf{x}),\rho) = 2\,d\,\sigma(yf(\mathbf{x})-\rho) + 2\,(1-d)\,\sigma(yf(\mathbf{x})+\rho) \tag{13}$$

where $\sigma(a) = \frac{1}{1+e^{\mu a}}$ and $\mu > 0$. We conjecture that shifted DSL is a calibrated loss for ℓ_d^γ.

Definition 13 (Shifted Double Sigmoid Loss) *Given the shift parameter $\beta > 0$, we define the shifted double sigmoid loss as,*

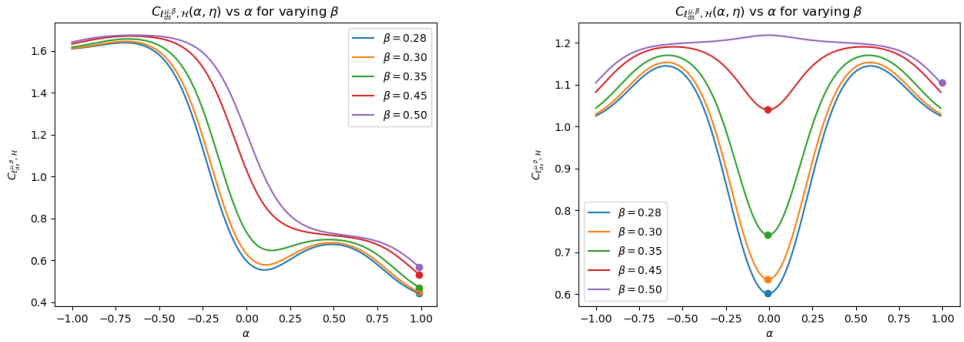
$$\ell_{ds}^{\mu,\beta}(yf(\mathbf{x}),\rho) = \ell_{ds}^{\mu}(yf(\mathbf{x})-\beta,\rho) = 2\,d\,\sigma(yf(\mathbf{x})-\beta-\rho) + 2\,(1-d)\,\sigma(yf(\mathbf{x})-\beta+\rho). \tag{14}$$

As ℓ_d^γ (eq.(8)) is γ-right shifted version of ℓ_d (Proposition 8), for $\ell_{\text{ds}}^{\mu,\beta}$ (eq.(14)) to be surrogate to ℓ_d^γ, the shift (β), has to be at least γ i.e $\beta \geq \gamma$. We can easily see that $\ell_{\text{ds}}^{\mu,\beta}$ is not a convex function of $yf(\mathbf{x})$. Conditional risk associated with $\ell_{\text{ds}}^{\mu,\beta}$ can be written as

$$\mathcal{C}_{\ell_{\text{ds}}^{\mu,\beta},\mathcal{H}}(f(\mathbf{x}),\eta) = \eta\,\ell_{\text{ds}}^{\mu,\beta}(f(\mathbf{x}),\rho) + (1-\eta)\,\ell_{\text{ds}}^{\mu,\beta}(-f(\mathbf{x}),\rho) \tag{15}$$

Analysis of conditional risk of $\ell_{\text{ds}}^{\mu,\beta}$ for varying β values :

Here, we empirically demonstrate the effect of varying shift parameter (β) on the conditional risk to identify the cases when calibration holds. Figure 2 shows plots for varying β for $\eta = 0.6$ and $\eta = 0.5$ respectively. We see that for fixed μ, d, γ - high β values are needed to push the minima towards the right-most end (i.e close to 1, see Figure 2(a)) and low β values keep minima closer to origin (Figure 2(b)). High β values are therefore favourable for (12) and low β values favour (11). For calibration, both the conditions need to be satisfied. $\exists\ \beta$ (depending on μ, γ, d) for which these conditions are satisfied.

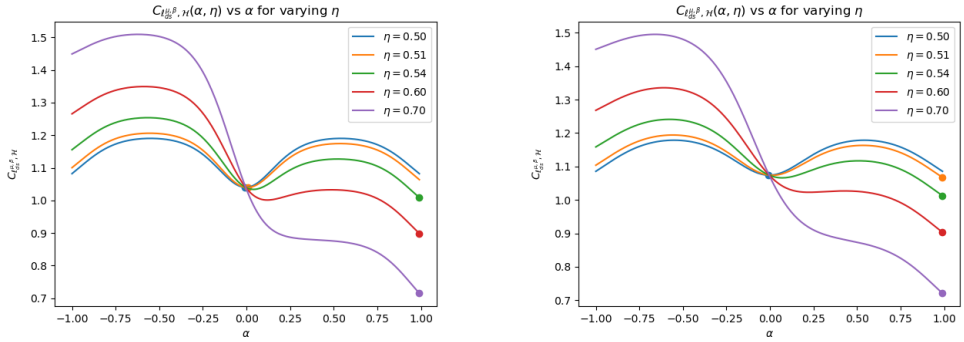


(a) $\eta = 0.6$ (b) $\eta = 0.5$

Figure 2: (a) corresponds to case when $\eta = 0.6$ and (b) corresponds to case when $\eta = 0.5$, for varying β values with fixed $d = 0.2$ and fixed $\mu = 3.0$

Analysis of conditional risk of $\ell_{\mathbf{ds}}^{\mu,\beta}$ for varying η values :

Here, we empirically demonstrate the effect of varying η on the conditional risk to identify the cases when calibration holds. Given below are the plot of (15) vs α for fixed $\beta = 0.45$ and fixed $d = 0.2$ with $\mu = 3.0$ and $\mu = 2.65$ respectively. For the case when $\eta > 0.5$, we



(a) $\mu = 3.0$ (b) $\mu = 2.65$

Figure 3: (a) corresponds to $\mu = 3.0$ and (b) corresponds to $\mu = 2.65$

consider two offsets $\xi_1 = 0.04$ and $\xi_2 = 0.01$ and analyse calibration for $\eta + \xi_1$ and $\eta + \xi_2$. From Figure 3 (a), when $\mu = 3.0$, it is evident that minima for $\eta + \xi_2$ is located close to the origin, thereby violating (11) whereas minima for $\eta + \xi_1$ is located at the rightmost end. Upon reducing the eta value from $\eta + \xi_1$ to $\eta + \xi_2$, we need to reduce the value of μ so that calibration conditions are satisfied. This is evident from Figure 3 (b), where $\mu = 2.65$. We claim that $\exists$ a single μ value that satisfies both calibration conditions no matter how small the offset ξ is taken. Empirically, it is seen that both calibration conditions (11) and (12) are satisfied. The minima lies in $[0, \rho - \gamma]$ for $\eta = 0.5$ and lies beyond $\rho + \gamma$ for $\eta > 0.5$.

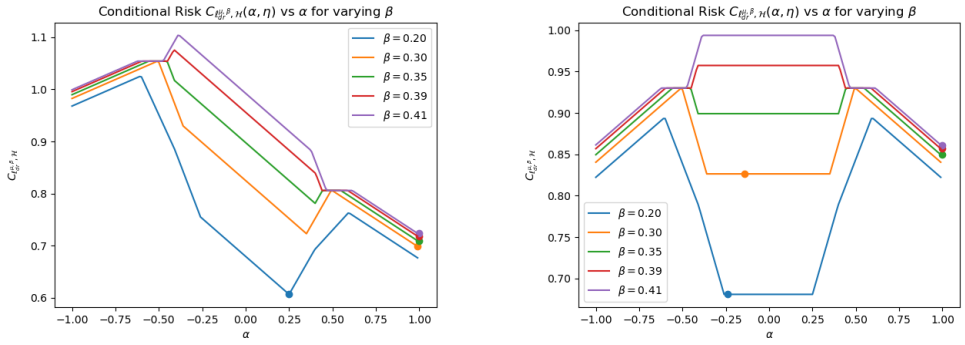
5.2. Shifted Double Ramp Loss

Double ramp loss (DRL) (Manwani et al., 2013) is proposed in the context of standard reject option classification without adversarial attacks. It is shown to be a calibrated surrogate to ℓ_d (Shah and Manwani, 2018). Let $[a]_+ := \max(0, a)$, then DRL is described as,

$$\ell^{\mu}_{\text{dr}}(yf(\mathbf{x}), \rho) = \frac{d}{\mu}\left[[\mu + \rho - yf(\mathbf{x})]_+ - [-\mu^2 + \rho - yf(\mathbf{x})]_+\right] + \frac{1-d}{\mu}\left[[\mu - \rho - yf(\mathbf{x})]_+ - [-\mu^2 - \rho - yf(\mathbf{x})]_+\right]$$

Definition 14 (Shifted Double Ramp Loss) *Given the shift parameter $\beta(\geq 0)$, shifted double ramp loss is defined as $\ell^{\mu,\beta}_{dr}(yf(\mathbf{x}), \rho) = \ell^{\mu}_{dr}(yf(\mathbf{x}) - \beta, \rho)$.*

For shifted DRL $\ell^{\mu,\beta}_{\text{dr}}$ to be a surrogate for ℓ^{γ}_{d}, we require that $\beta \geq \gamma$. Conditional risk associated with $\ell^{\mu,\beta}_{\text{dr}}$ is $\mathcal{C}_{\ell^{\mu,\beta}_{\text{dr}},\mathcal{H}}(f(\mathbf{x}), \eta) = \eta\, \ell^{\mu,\beta}_{\text{dr}}(f(\mathbf{x}), \rho) + (1-\eta)\, \ell^{\mu,\beta}_{\text{dr}}(-f(\mathbf{x}), \rho)$.



(a) $\eta = 0.6$ (b) $\eta = 0.5$

Figure 4: Conditional risk $\mathcal{C}_{\ell^{\mu,\beta}_{\text{dr}},\mathcal{H}}$ versus shift parameter β for $d, \mu, \gamma = \{0.2, 0.55, 0.2\}$ respectively. (a) For $\eta = 0.6$, high β values push minima towards the rightmost end (beyond the $\rho + \gamma$ mark). (b) For $\eta = 0.5$, low β (values that are closer to γ) are good to ensure that minima lies in $[0, \rho - \gamma]$.

Analysis of Conditional risk of $\ell^{\mu,\beta}_{\text{dr}}$ for varying β :

Figure 4 shows plots of $\mathcal{C}_{\ell^{\mu,\beta}_{\text{dr}},\mathcal{H}}$ with varying β values for fixed $d(= 0.2), \mu(= 0.55), \gamma(= 0.2)$. We make following observations. Low β (values that are closer to γ) are good to ensure that minima lies in $[0, \rho - \gamma]$ thereby, satisfying (11) for $\eta = \frac{1}{2}$ (as seen in Figure 4 (b)) but it violates (12) for $\eta > 0.5$ (as seen in Figure 4 (a)). On the other hand, high β values push minima towards the rightmost end (beyond the $\rho + \gamma$ mark), thereby favoring (12) for $\eta > 0.5$ but violating (11) for $\eta = 0.5$ case. Similar to shifted Double Sigmoid case, here too, $\exists\ \beta$ value (depending on μ, d, γ) which satisfies both calibration conditions.

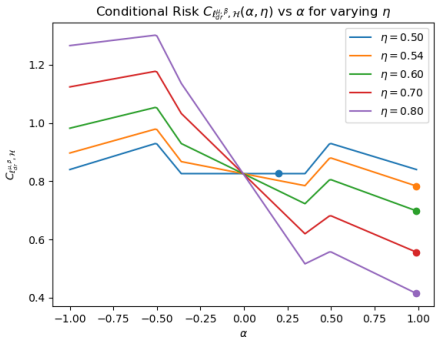


Figure 5: conditional risk $\mathcal{C}_{\ell^{\mu,\beta}_{dr},\mathcal{H}}$ vs α for varying η.

Analysis of Conditional risk of $\ell^{\mu,\beta}_{\mathbf{dr}}$ for varying η

Given below is the plot for varying η for $\beta = 0.3$ and fixed $\mu, d, \gamma = 0.55, 0.2, 0.2$ respectively. For shifted DRL also we observe "Minima-Jump". We see that, in the region around the origin, $\mathcal{C}_{\ell^{\mu,\beta}_{dr},\mathcal{H}}$ remains constant. For every offset ξ around η that we consider, μ has to be adjusted while keeping d, γ fixed to ensure calibration. As seen in Figure 5.2, for $\xi = 0.04$, $\mu = 0.55$ with $\beta = 0.3$ yields calibration.

5.3. Non Quasi-Concavity of $\mathcal{C}_{\ell^{\mu,\beta}_{\mathbf{ds}},\mathcal{H}}(f(\mathbf{x}),\eta)$ and $\mathcal{C}_{\ell^{\mu,\beta}_{\mathbf{dr}},\mathcal{H}}(f(\mathbf{x}),\eta)$

Theorem 12 states that if for a margin based surrogate of ℓ^{γ}_{d}, the conditional risk is quasi-concave in $\alpha \;\; \forall \eta \in [0,1]$, then it is not $(\ell^{\gamma}_{d}, \mathcal{H})$-calibrated. It can be seen easily (refer Fig 3 and Fig 5.2) that $\mathcal{C}_{\ell^{\mu,\beta}_{\text{ds}},\mathcal{H}}(f(\mathbf{x}),\eta)$ and $\mathcal{C}_{\ell^{\mu,\beta}_{\text{dr}},\mathcal{H}}(f(\mathbf{x}),\eta)$ are not quasi-concave in $\alpha(= f(\mathbf{x}))\; \forall \eta$. This property makes $\ell^{\mu,\beta}_{\text{ds}}$ and $\ell^{\mu,\beta}_{\text{dr}}$ candidate surrogates which can be $(\ell^{\gamma}_{d}, \mathcal{H})$-calibrated.

5.4. Adversarial Training using the Double Sigmoid Loss / Double Ramp Loss

In this section, we present a generic algorithm for adversarial training of linear reject option classifier using shifted DSL $\ell^{\mu,\beta}_{\text{ds}}$ and shifted DRL $\ell^{\mu,\beta}_{\text{dr}}$. We explain here adversarial learning using shifted DSL $\ell^{\mu,\beta}_{\text{ds}}$. For shifted DRL, we adopt a similar approach.

Step 1: Train a linear reject option classifier (parameters are $\Theta = [\mathbf{w}, \rho]^T$) on clean data $\mathcal{D}_{\text{clean}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ by minimizing empirical risk under ℓ^{μ}_{ds}. Bias term can be included in the vector $\mathbf{w}$ by appending 1 in the feature vector $\mathbf{x}$. Optimal parameters $\Theta^* = [\mathbf{w}^*, \rho^*]^T$ are obtained as, $\Theta^* = \arg\min_{\Theta} \; \frac{1}{N}\sum_{i=1}^{N} \ell^{\mu}_{\text{ds}}(y_i f(\mathbf{x}_i), \rho)$. using stochastic gradient descent. The optimal classifier is $f^*(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x}$.

Step 2: Generate adversarial data over a subset of examples indexed by set $\mathcal{I} \subseteq [N]$. For any $\mathbf{x}_i, \;\; i \in \mathcal{I}$ and $\gamma > 0$, its adversarial corrupted version $\mathbf{x}^{\gamma}_i$ is generated as $\mathbf{x}^{\gamma}_i = \arg\max_{\mathbf{x}'_i \in B_2(\mathbf{x}_i,\gamma)} \; \ell^{\mu}_{\text{ds}}(y_i \; f^*(\mathbf{x}'_i), \rho^*)$. Projected Gradient Ascent is used to maximise (13), by taking an ascent step in the gradient direction and projecting it onto $B_2(\mathbf{x}_i\gamma)$, in succession. An adversarial dataset is created by adding the perturbed samples to clean data samples $\mathcal{D}_{\text{adv}} = \{(\mathbf{x}^{\gamma}_i, y_i)\}_{i=1}^{N} = \{(\mathbf{x}^{\gamma}_i, y_i), \; i \in \mathcal{I}\} \cup \{(\mathbf{x}_i, y_i), i \in [N] \setminus \mathcal{I}\}$.

Step 3: Train a robust linear reject option classifier by minimizing the empirical risk on adversarial data $\mathcal{D}_{\text{adv}}$ using shifted DSL. The optimal parameters $\Theta^{\gamma} = [\mathbf{w}^{\gamma}, \rho^{\gamma}]^T$ after adversarial training are found as, $\Theta^{\gamma} = \arg\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{ds}}^{\beta,\mu}(y_i(\mathbf{w} \cdot \mathbf{x}_i^{\gamma}), \rho)$.

6. Experiments

6.1. Baselines and Dataset Description

The linear Reject Option Classifiers (ROC) trained using DSL or DRL without shift ($\beta = 0$) are used as base models. Introducing shift ($\beta > 0$) in the corresponding DSL or DRL makes them robust to l_2-norm attacks. We report the performance of the non-robust ($\beta = 0$) and three robust classifiers ($\beta = 0.1, 0.15, 0.25$) on it as shown in Table 1 and Table 2. We do not choose ATRO (Kato et al., 2020) as a baseline as ATRO works on l_∞ perturbations, whereas our work deals with l_2 perturbations.
We generate a linearly separable data ($\in \mathbb{R}^2$) with separation boundary as $\mathbf{x} = 0$. All points should lie in $B_2(\mathbf{0}, 1)$, the unit circle centred at origin. Take rejection width $= 0.5$ and 100 points per class in the reject region and 200 points per class in the non-reject region. Flip the labels of 5% of the samples in the reject region for each class. This data will be used to train the classifiers. For testing, we generate data similarly, except for half the count of training. The perturbation radius of l_2-attack, γ is referred as γ_{train} and γ_{test} for training time and test-time respectively ; $\gamma_{\text{train}} \in \{0.0, 0.1, 0.2\}$ and $\gamma_{\text{test}} \in \{0.0, 0.1, 0.2\}$. Every classifier trained using any value of γ_{train} is tested with all three values of γ_{test}.

6.2. Observations

Table 1 and Table 2 show results for shifted DSL and shifted DRL. For each value of $d, \mu, \gamma_{\text{train}}, \beta$; we report the following metrics evaluated on test-data by averaging over 10 runs: (a) error (b) rejection rate (c) accuracy on the predicted samples.

6.2.1. Effect of Increasing γ_{test} in the Test Time Attack

For a fixed value of γ_{train}, d and β, as test-time attack γ_{test} increases, the error increases. We observe this behavior with both shifted DSL and shifted DRL. This happens due to the following reason. By increasing the γ_{test}, the overlap between the two classes increases, increasing the linear classifier's error. However, we observe the following additional property in the case of shifted DRL. If we train with shifted DRL for a certain γ_{train}, its error at test time does not increase much by increasing γ_{test} as long as $\gamma_{\text{test}} \leq \gamma_{\text{train}}$. Shifted DRL has a flat region for $yf(\mathbf{x}) \in [-\rho + \mu + \beta, \rho - \mu^2 + \beta] \cup [-\infty, -\rho - \mu^2 + \beta] \cup [\rho + \mu + \beta, \infty)$. For smaller γ_{test}, pushing out an example out of these flat regions is hard. For sufficiently large γ_{test}, the loss can increase for such points in two ways. (a) Correctly classified data point in the region $[\rho + \mu + \beta, \infty)$ after γ_{test} perturbation leaves the zero loss region and moves towards rejection region, thereby increasing loss value. (b) The data point is in the region $[-\rho + \mu + \beta, \rho - \mu^2 + \beta]$ where the loss value is $d(1 + \mu)$, and after perturbation moves towards the misclassification region and achieves higher loss values.

γ_{train}	d	Training Loss	Attack $\gamma_{\text{test}}=0$ Err	Acc	RR	$\gamma_{\text{test}}=0.1$ Err	Acc	RR	$\gamma_{\text{test}}=0.2$ Err	Acc	RR
0	0.2	DSL ($\beta=0$)	0.338	0.458	0.53	0.38	0.394	0.514	0.484	0.306	0.41
		DSL ($\beta=0.1$)	0.315	0.353	0.612	0.342	0.308	0.604	0.409	0.24	0.544
		DSL ($\beta=0.15$)	0.251	0.15	0.828	0.26	0.135	0.825	0.28	0.116	0.808
		DSL ($\beta=0.25$)	**0.25**	0.148	0.835	**0.258**	0.132	0.836	**0.275**	0.107	0.829
	0.3	DSL ($\beta=0$)	0.435	0.501	0.315	0.552	0.38	0.202	0.663	0.296	0.106
		DSL ($\beta=0.1$)	0.43	0.5	0.35	0.525	0.394	0.264	0.638	0.304	0.15
		DSL ($\beta=0.15$)	0.378	0.297	0.615	0.430	0.237	0.57	0.509	0.174	0.488
		DSL ($\beta=0.25$)	**0.375**	0.297	0.631	**0.407**	0.253	0.613	**0.486**	0.182	0.529
	0.4	DSL ($\beta=0$)	0.476	0.507	0.176	0.623	0.345	0.109	0.7	0.288	0.035
		DSL ($\beta=0.1$)	0.475	0.506	0.194	0.616	0.353	0.121	0.699	0.288	0.039
		DSL ($\beta=0.15$)	**0.436**	0.448	0.343	0.578	0.32	0.272	0.651	0.261	0.197
		DSL ($\beta=0.25$)	0.457	0.398	0.432	**0.553**	0.291	0.348	**0.634**	0.226	0.259
0.1	0.2	DSL ($\beta=0$)	0.363	0.45	0.456	0.414	0.38	0.434	0.537	0.273	0.327
		DSL ($\beta=0.1$)	0.36	0.447	0.469	0.393	0.393	0.467	0.49	0.298	0.383
		DSL ($\beta=0.15$)	**0.285**	0.336	0.711	**0.302**	0.304	0.712	**0.348**	0.255	0.678
		DSL ($\beta=0.25$)	0.302	0.296	0.664	0.323	0.257	0.667	0.388	0.189	0.611
	0.3	DSL ($\beta=0$)	0.437	0.498	0.319	0.56	0.367	0.216	0.681	0.277	0.097
		DSL ($\beta=0.1$)	0.425	0.5	0.372	0.527	0.38	0.288	0.64	0.29	0.176
		DSL ($\beta=0.15$)	0.426	0.501	0.36	0.524	0.389	0.277	0.65	0.287	0.151
		DSL ($\beta=0.25$)	**0.423**	0.5	0.38	**0.509**	0.396	0.313	**0.627**	0.291	0.202
	0.4	DSL ($\beta=0$)	0.485	0.499	0.155	0.624	0.352	0.098	0.698	0.291	0.034
		DSL ($\beta=0.1$)	0.479	0.501	0.187	0.615	0.356	0.115	0.697	0.290	0.04
		DSL ($\beta=0.15$)	0.478	0.498	0.227	0.611	0.354	0.136	0.698	0.284	0.053
		DSL ($\beta=0.25$)	**0.464**	0.455	0.324	**0.583**	0.321	0.244	**0.667**	0.253	0.161
0.2	0.2	DSL ($\beta=0$)	0.341	0.404	0.518	0.382	0.347	0.502	0.48	0.264	0.405
		DSL ($\beta=0.1$)	**0.2**	0	1	**0.2**	0	1	**0.2**	0	1
		DSL ($\beta=0.15$)	0.218	0.048	0.942	0.219	0.047	0.942	0.219	0.047	0.942
		DSL ($\beta=0.25$)	0.218	0.048	0.942	0.218	0.048	0.942	0.218	0.048	0.942
	0.3	DSL ($\beta=0$)	0.434	0.501	0.32	0.551	0.379	0.217	0.671	0.288	0.098
		DSL ($\beta=0.1$)	0.332	0.203	0.833	0.349	0.166	0.825	0.355	0.157	0.816
		DSL ($\beta=0.15$)	0.315	0.103	0.915	0.323	0.086	0.911	0.326	0.082	0.906
		DSL ($\beta=0.25$)	**0.3**	0	1	**0.3**	0	1	**0.3**	0	1
	0.4	DSL ($\beta=0$)	0.477	0.506	0.168	0.623	0.351	0.1	0.698	0.29	0.034
		DSL ($\beta=0.1$)	0.437	0.5	0.628	0.465	0.435	0.633	0.479	0.414	0.612
		DSL ($\beta=0.15$)	0.43	0.353	0.698	0.453	0.3	0.702	0.462	0.285	0.69
		DSL ($\beta=0.25$)	**0.409**	0.09	0.904	**0.417**	0.08	0.906	**0.418**	0.08	0.904

Table 1: Results with linear reject option classifier with/without shift trained using Double Sigmoid Loss ($\mu = 2.65$).

6.2.2. Effect of Increasing d

For any robust classifier (with fixed $\beta, \gamma_{\text{train}}$), an increase in the cost of rejection d leads to an increase in error and a reduction in the rejection rate. However, at high values of training γ ($= 0.2$), the overlap between the two classes becomes very large, and the classifier starts rejecting almost all samples. This is the common behavior of any reject option classifier.

6.2.3. Effect of Increasing β

Robust classifiers ($\beta = 0.1, 0.15, 0.25$) give less error than their non-robust ($\beta = 0$) counterparts, which is expected from shifted DSL and shifted DRL. However, we observe this behavior when γ_{train} is nonzero in the case of shifted DRL. For $\gamma_{\text{train}} = 0$, changing β does not make any change in the performance of shifted DRL for different values of d and γ_{test}. Also, for fixed d, β and for any $\gamma_{\text{test}} \leq \gamma_{\text{train}}$, the difference between the errors of non-robust classifier and the robust classifier is very small. This gap starts to widen when $\gamma_{\text{test}} > \gamma_{\text{train}}$.

			Attack $\gamma_{\text{test}} = 0$			Attack $\gamma_{\text{test}} = 0.1$			Attack $\gamma_{\text{test}} = 0.2$		
γ_{train}	d	Training Loss	Err	Acc	RR	Err	Acc	RR	Err	Acc	RR
0	0.2	DRL ($\beta = 0$)	**0.451**	0.494	0.175	0.538	0.481	0.114	**0.545**	0.414	0.101
		DRL ($\beta = 0.1$)	**0.451**	0.495	0.173	**0.537**	0.418	0.113	**0.545**	0.414	0.101
		DRL ($\beta = 0.15$)	**0.451**	0.495	0.173	**0.537**	0.418	0.113	**0.545**	0.414	0.101
		DRL ($\beta = 0.25$)	**0.451**	0.495	0.173	**0.537**	0.418	0.113	**0.545**	0.414	0.101
	0.3	DRL ($\beta = 0$)	**0.475**	0.498	0.129	0.561	0.416	0.082	**0.559**	0.419	0.075
		DRL ($\beta = 0.1$)	**0.475**	0.498	0.129	**0.56**	0.416	0.082	**0.559**	0.419	0.075
		DRL ($\beta = 0.15$)	**0.475**	0.498	0.129	**0.56**	0.416	0.082	**0.559**	0.419	0.075
		DRL ($\beta = 0.25$)	**0.475**	0.498	0.129	**0.56**	0.416	0.082	**0.559**	0.419	0.075
	0.4	DRL ($\beta = 0$)	**0.491**	0.497	0.106	0.574	0.416	0.065	0.565	0.424	0.058
		DRL ($\beta = 0.1$)	**0.491**	0.497	0.106	**0.573**	0.414	0.064	**0.564**	0.424	0.058
		DRL ($\beta = 0.15$)	**0.491**	0.497	0.106	**0.573**	0.414	0.064	**0.564**	0.424	0.058
		DRL ($\beta = 0.25$)	**0.491**	0.497	0.106	**0.573**	0.414	0.064	**0.564**	0.424	0.058
0.1	0.2	DRL ($\beta = 0$)	0.39	0.49	0.384	0.404	0.477	0.367	0.427	0.457	0.338
		DRL ($\beta = 0.1$)	**0.276**	0.595	0.752	**0.28**	0.59	0.746	**0.293**	0.579	0.729
		DRL ($\beta = 0.15$)	**0.276**	0.595	0.752	**0.28**	0.59	0.746	**0.293**	0.579	0.729
		DRL ($\beta = 0.25$)	**0.276**	0.595	0.752	**0.28**	0.59	0.746	**0.293**	0.579	0.729
	0.3	DRL ($\beta = 0$)	0.426	0.495	0.381	0.439	0.481	0.362	0.443	0.478	0.357
		DRL ($\beta = 0.1$)	**0.376**	0.695	0.621	**0.387**	0.683	0.605	**0.388**	0.601	0.603
		DRL ($\beta = 0.15$)	**0.376**	0.695	0.621	**0.387**	0.683	0.605	**0.388**	0.601	0.603
		DRL ($\beta = 0.25$)	**0.376**	0.695	0.621	**0.387**	0.683	0.605	**0.388**	0.601	0.603
	0.4	DRL ($\beta = 0$)	0.466	0.492	0.362	0.489	0.468	0.319	0.522	0.435	0.267
		DRL ($\beta = 0.1$)	**0.465**	0.496	0.368	**0.474**	0.485	0.353	**0.493**	0.46	0.321
		DRL ($\beta = 0.15$)	**0.465**	0.496	0.368	**0.474**	0.485	0.353	**0.493**	0.46	0.321
		DRL ($\beta = 0.25$)	**0.465**	0.496	0.368	**0.474**	0.485	0.353	**0.493**	0.463	0.321
0.2	0.2	DRL ($\beta = 0$)	0.359	0.508	0.453	0.359	0.508	0.453	0.359	0.508	0.453
		DRL ($\beta = 0.1$)	**0.229**	0.904	0.895	**0.229**	0.904	0.895	**0.229**	0.904	0.895
		DRL ($\beta = 0.15$)	**0.229**	0.904	0.895	**0.229**	0.904	0.895	**0.229**	0.904	0.895
		DRL ($\beta = 0.25$)	**0.229**	0.904	0.895	**0.229**	0.904	0.895	**0.229**	0.904	0.895
	0.3	DRL ($\beta = 0$)	0.421	0.498	0.454	0.421	0.498	0.454	0.421	0.498	0.454
		DRL ($\beta = 0.1$)	**0.414**	0.5	0.427	**0.414**	0.499	0.427	**0.414**	0.499	0.427
		DRL ($\beta = 0.15$)	**0.414**	0.5	0.427	**0.414**	0.499	0.427	**0.414**	0.499	0.427
		DRL ($\beta = 0.25$)	**0.414**	0.5	0.427	**0.414**	0.5	0.427	**0.414**	0.5	0.427
	0.4	DRL ($\beta = 0$)	0.477	0.47	0.4	0.486	0.459	0.386	0.507	0.436	0.351
		DRL ($\beta = 0.1$)	**0.465**	0.484	0.434	**0.465**	0.484	0.434	**0.465**	0.483	0.433
		DRL ($\beta = 0.15$)	**0.465**	0.484	0.434	**0.465**	0.484	0.434	**0.465**	0.483	0.434
		DRL ($\beta = 0.25$)	**0.465**	0.483	0.433	**0.465**	0.483	0.433	**0.465**	0.483	0.433

Table 2: Results with linear reject option classifier with/without shift trained using Double Ramp Loss ($\mu = 0.95$).

7. Conclusion and Future Work

In this paper, we give a complete characterization of surrogates calibrated to ℓ_d^γ and provide insights on designing them (via extensive analysis of $\ell_{\text{ds}}^{\mu,\beta}$ and $\ell_{\text{dr}}^{\mu,\beta}$) for the hypothesis set $\mathcal{H} = \mathcal{H}_{\text{lin}}$. To the best of our knowledge, this is the first attempt towards analyzing surrogates in the "Adversarial Robust Reject Option" setting for binary classification from the lens of Calibration Theory. The first line of future work is to provide a proof technique for class of surrogates which are $(\ell_d^\gamma, \mathcal{H})$-calibrated using the ideas presented in this work for $\ell_{\text{ds}}^{\mu,\beta}$ and $\ell_{\text{dr}}^{\mu,\beta}$. Calibration analysis for other function classes like generalized linear models ($\mathcal{H}_g$) and single-layer ReLU neural networks ($\mathcal{H}_{\text{NN}}$) is another future research direction.

References

Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses. In *NeurIPS*, volume 34, 2021a.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. A finer calibration analysis for adversarial robustness. *CoRR*, abs/2105.01550, 2021b.

Han Bao, Clay Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *COLT*, volume 125, pages 408–451, 2020.

Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *JMLR*, 9(59):1823–1840, 2008.

Yuzhou Cao, Tianchi Cai, Lei Feng, Lihong Gu, Jinjie GU, Bo An, Gang Niu, and Masashi Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In *NeurIPS*, volume 35, pages 521–534, 2022.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

Jiefeng Chen, Jayaram Raghuram, Jihye Choi, Xi Wu, Yingyu Liang, and Somesh Jha. Stratified adversarial robustness with rejection. In *ICML*, pages 4867–4894. PMLR, 2023.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *COLT*, 2016.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv 1412.6572*, 2014.

Bhavya Kalra, Kulin Shah, and Naresh Manwani. RISAN: robust instance specific deep abstention network. In *UAI*, volume 161, pages 1525–1534, 2021.

Masahiro Kato, Zhenghang Cui, and Yoshihiro Fukuhara. ATRO: adversarial training with a rejection option. *CoRR*, abs/2010.12905, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Naresh Manwani, Kalpit Desai, Sanand Sasidharan, and Ramasubramanian Sundararajan. Double ramp loss based reject option classifier. In *PAKDD*, 2013.

Laurent Meunier, Raphael Ettedgui, Rafael Pinot, Yann Chevaleyre, and Jamal Atif. Towards consistency in adversarial classification. In *NeurIPS*, volume 35, pages 8538–8549, 2022.

Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. In *NeurIPS*, volume 32, 2019.

Nicolas Papernot and Patrick D. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *CoRR*, abs/1803.04765, 2018.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *CoRR*, abs/1801.09344, 2018.

H. G. Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12:530–554, 2018.

Kulin Shah and Naresh Manwani. Sparse and robust reject option classifier using successive linear programming. *CoRR*, abs/1802.04235, 2018.

Kulin Shah and Naresh Manwani. Online active learning of reject option classifiers. In *AAAI*, pages 5652–5659. AAAI Press, 2020.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 08 2007.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*, pages 5286–5295, 2018.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. ML-LOO: detecting adversarial examples with feature attribution. *CoRR*, abs/1906.03499, 2019.

Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective, 2022.