Learning to reject meets OOD detection: Are all abstentions created equal?

Harikrishna Narasimhan
Google Research, Mountain View
hnarasimhan@google.com

Aditya Krishna Menon
Google Research, New York
adityakmenon@google.com

Wittawat Jitkrittum
Google Research, New York
wittawat@google.com

Sanjiv Kumar
Google Research, New York
sanjivk@google.com

February 1, 2023

Abstract

Learning to reject (L2R) and out-of-distribution (OOD) detection are two classical problems, each involving the detection of certain abnormal samples: the goal is to detect "hard" samples on which to abstain in L2R, and "outlier" samples not drawn from the training distribution in OOD detection. Intriguingly, despite being developed in parallel literatures, both problems share a simple baseline: the *maximum softmax probability* (MSP) score. Despite this connection, there is limited understanding of precisely how these problems relate. In this paper, we formally relate these problems, and show how they may be jointly solved. We first show that while MSP is theoretically optimal for L2R, it can be *sub-optimal* for OOD detection in some important settings. We then characterize the Bayes-optimal classifier for a unified formulation that generalizes both L2R and OOD detection. Based on this, we design a plug-in approach for learning to abstain on both inlier and OOD samples, while constraining the total abstention budget. Experiments on benchmark OOD datasets demonstrate that our approach yields competitive classification and OOD detection performance compared to baselines from both literatures.

1 Introduction

In typical classification problems, one seeks to learn a classifier that is capable of making accurate predictions on *all* samples. However, in many real-world applications, it may be beneficial to allow the classifier to *abstain* from predicting on "hard" (e.g., ambiguously labelled) examples. This *learning to reject* (L2R) or *learning with abstention* paradigm has been well-studied (Chow, 1970; Bartlett & Wegkamp, 2008; Cortes et al., 2016b; Ramaswamy et al., 2018; Thulasidasan et al., 2019; Ni et al., 2019; Charoenphakdee et al., 2021), with more recent works considering a variant where the classifier is allowed to defer to an expert (Raghu et al., 2019b,a; Mozannar & Sontag, 2020; Okati et al., 2021; Verma & Nalisnick, 2022; Narasimhan et al., 2022).

A distinct problem that has also received much attention is *out-of-distribution (OOD) detection*, where the goal is to detect samples that are "outliers", in the sense of not being drawn from a classifier's training distribution (Hendrycks & Gimpel, 2017; Lee et al., 2018; Ren et al., 2019; Bitterwolf et al., 2022; Katz-Samuels et al., 2022). A special case of this problem, referred to as *open-set classification*, seeks to identify samples that belong to a class not seen during training (Bendale & Boult, 2016; Vaze et al., 2021).

While conceptually similar — intuitively, it is desirable for a classifier to abstain on OOD samples — L2R and OOD detection have been developed in parallel literatures. Intriguingly, however, both problems share a simple yet effective baseline: one thresholds the maximum softmax probability from a standard classifier, with low probability samples adjudged to be abstention-worthy, or likely outliers. This is known as *Chow's rule* (Chow, 1970) for L2R, and the *maximum softmax probability* (*MSP*) score (Hendrycks & Gimpel, 2017) for OOD detection.

Given this connection, one may ask: how precisely do these problems relate, and can they be jointly solved? Surprisingly, while L2R and OOD detection are generally acknowledged to be related (Yang et al., 2021), these questions have not received systematic study (see §2). In this paper, we make the following contributions towards addressing this issue:

(i) We prove that the MSP baseline can be *sub-optimal* for important special cases of OOD detection, such as open-set classification (§3, Lemmas 3.1, 3.2).

(ii) We outline a formulation that unifies both the L2R and OOD detection paradigms, and show that the resulting Bayes-optimal classifier involves a *sample-dependent threshold* on the MSP (§4, Lemma 4.1).

(iii) We propose a plug-in algorithm that seeks to approximate the Bayes-optimal classifier, and showcase its application to learning an OOD-aware classifier with a budget constraint on the total proportion of abstentions (§5).

(iv) Our experiments on benchmark OOD detection image datasets show that our plug-in approach yields competitive classification and OOD detection performance at any desired abstention rate, compared to both standard learning to reject and OOD detection baselines (§6).

2 Background and notation

Given instances $\mathcal{X}$, labels $\mathcal{Y} \doteq [L]$, and a distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$, multi-class classification concerns learning a classifier $h \colon \mathcal{X} \to \mathcal{Y}$ with minimal misclassification error $R_{\mathrm{err}}(h) \doteq \mathbb{P}(y \neq h(x))$. Such a classifier is typically implemented via $h(x) = \mathrm{argmax}_{y \in [L]} f_y(x)$, where $f \colon \mathcal{X} \to \mathbb{R}^L$ scores the affinity of each label to a given instance. One typically learns f from a training sample $S = \{(x_n, y_n)\}_{n \in [N]}$ drawn from $\mathbb{P}$, via minimisation of the *empirical risk* $\hat{R}(f; \ell) \doteq \frac{1}{N} \sum_{n \in [N]} \ell(y_n, f(x_n))$ for *loss* $\ell \colon [L] \times \mathbb{R}^L \to \mathbb{R}_+$.

2.1 Learning to reject (L2R)

Standard classification assumes one makes a prediction on *all* samples, regardless of how confident the model is. In the *learning to reject* (*L2R*) setting, one has the option of *abstaining* from making a prediction (denoted by predicting $\perp$), at the expense of incurring a cost $c_0 \in [0, 1]$ (Bartlett & Wegkamp, 2008). Intuitively, one may wish to abstain on "hard" (e.g., ambiguously labelled) samples, which could then be forwarded to an expert (e.g., a human labeller).

Formally, one seeks to learn classifier $h \colon \mathcal{X} \to \mathcal{Y}$ and *rejector* $r \colon \mathcal{X} \to \{0, 1\}$, where $r(x) = 1$ denotes "rejected" samples where we abstain from making a prediction. Our goal is to minimise the *abstention-aware misclassification error*:

$$R_{\mathrm{rej}}(h, r) \doteq \mathbb{P}(y \neq h(x), r(x) = 0) + c_0 \cdot \mathbb{P}(r(x) = 1). \quad (1)$$

Intuitively, this is the standard misclassification error on non-rejected samples, plus the cost c_0 times the fraction of rejected samples. The *Bayes-optimal* rejector minimising (1) follows *Chow's rule* (Chow, 1970;

Ramaswamy et al., 2018):

$$r^*(x) = \llbracket \max_{y \in [L]} \mathbb{P}(y \mid x) < 1 - c_0 \rrbracket . \tag{2}$$

Intuitively, we abstain on "hard" samples where the most likely label is ambiguous, i.e., it has low $\mathbb{P}(y \mid x)$.

Chow's rule (2) is not directly practical, since it requires knowledge of $\mathbb{P}(y \mid x)$. The simplest L2R baseline is the *confidence-based* method (Ni et al., 2019), which applies Chow's rule to a *plug-in estimate* $\hat{\mathbb{P}}(y \mid x)$ of $\mathbb{P}(y \mid x)$: given a scorer $f \colon \mathcal{X} \to \mathbb{R}^L$ learned from a training sample, one computes softmax probability estimates $\hat{\mathbb{P}}(y \mid x) \propto \exp(f_y(x))$. Samples with $\max_{y \in [L]} \hat{\mathbb{P}}(y \mid x) < 1 - c_0$ are then rejected. Alternatively, one may modify the training loss ℓ to yield a consistent estimate of Chow's rule (Bartlett & Wegkamp, 2008; Ramaswamy et al., 2018; Charoenphakdee et al., 2021; Gangrade et al., 2021), or learn an explicit rejection function jointly with the classifier (Cortes et al., 2016a; Thulasidasan et al., 2019; Mozannar & Sontag, 2020).

2.2 Out-of-distribution (OOD) detection

The out-of-distribution (OOD) detection problem concerns distinguishing instances from an "inlier" distribution $\mathbb{P}_{\text{in}}$ and an "outlier" distribution $\mathbb{P}_{\text{out}}$ (Hendrycks & Gimpel, 2017). This is a natural auxiliary task in real-world deployment of machine learning classifiers, owing to the possibility of such models producing high-confidence predictions even on completely arbitrary inputs (Nguyen et al., 2015; Hendrycks & Gimpel, 2017), and scoring outlier samples higher than inlier samples (Nalisnick et al., 2019).

Formally, one seeks a binary OOD detector $r \colon \mathcal{X} \to \{0, 1\}$ that predicts whether or not a given sample comes from $\mathbb{P}_{\text{out}}$, so as to minimise the *OOD detection error*

$$R_{\text{ood}}(r) \doteq \alpha \cdot \mathbb{P}_{\text{in}}(r(x) = 1) + \beta \cdot \mathbb{P}_{\text{out}}(r(x) = 0), \tag{3}$$

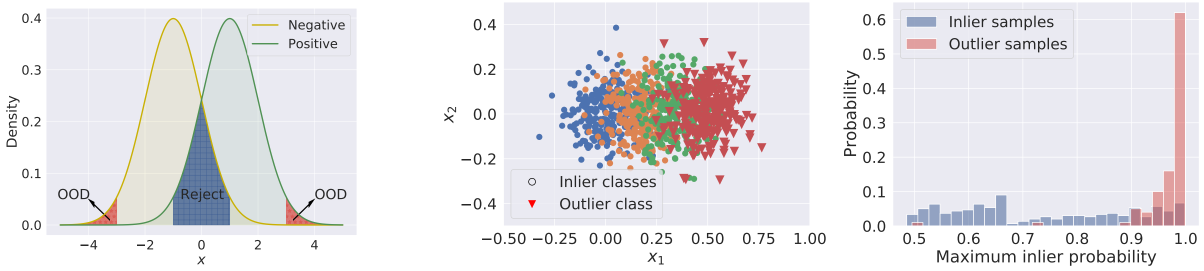
for suitable constants $\alpha, \beta \in [0, 1]$. This is a *cost-sensitive* misclassification error (Elkan, 2001), where samples from the outlier (inlier) distribution are treated as positive (negative). One may learn r by suitably thresholding an *OOD scorer* $s \colon \mathcal{X} \to \mathbb{R}$, to give $r(x) = \llbracket s(x) < t \rrbracket$. Via standard results (Scott, 2012), the Bayes-optimal detector for (3) thresholds the *density ratio* $\frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{out}}(x)}$:

$$r^*(x) = \left\llbracket \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{out}}(x)} < \frac{\beta}{\alpha} \right\rrbracket . \tag{4}$$

A popular metric for evaluating OOD detection methods is the fraction of outlier samples misclassified when 95% of inlier samples are classified as inlier Hendrycks et al. (2019); Bitterwolf et al. (2022). Minimizing this evaluation metric is equivalent to minimizing (3) for suitable (possibly distribution-dependent) choices of α and β.

The viability of OOD detection is strongly determined by the nature of the training data. There are three canonical settings considered in the literature: (i) one only observes samples from $\mathbb{P}_{\text{in}}$, and has no information on $\mathbb{P}_{\text{out}}$; (ii) one observes samples from $\mathbb{P}_{\text{in}}$, and some *training outlier* distribution $\mathbb{P}'_{\text{out}}$; (iii) one observes samples from $\mathbb{P}_{\text{in}}$, and an *unlabeled mixture* of samples from $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}$.

A remarkably effective baseline for OOD detection that requires only inlier samples is the *maximum softmax probability* (Hendrycks & Gimpel, 2017), possibly with temperature scaling and data augmentation (Liang et al., 2018). Given a scorer $f \colon \mathcal{X} \to \mathbb{R}^L$ learned from inlier samples alone, one computes softmax probability estimates $\hat{\mathbb{P}}_{\text{in}}(y \mid x) \propto \exp(f_y(x))$. Samples with $\max_{y \in [L]} \hat{\mathbb{P}}_{\text{in}}(y \mid x) < \frac{\beta}{\alpha}$ are then deemed to be OOD. Recent works found that the maximum *logit* (without softmax normalisation) may be preferable in some settings (Vaze et al., 2021; Hendrycks et al., 2022), possibly with normalisation (Wei et al., 2022). These may be recovered as a limiting case of energy-based approaches (Liu et al., 2020a).



(a) Uniform outlier distribution $\mathbb{P}_{\text{out}}$.

(b) Open-set classification.

Figure 1: Example of two settings where the maximum softmax probability (MSP) baseline fails for OOD detection. Setting **(a)** considers *low-density OOD detection*, where positive and negative samples drawn from a one-dimensional Gaussian distribution. Samples *away* from the origin will have $\mathbb{P}(x) \sim 0$, and are thus outliers under the Bayes-optimal OOD detector. However, the MSP baseline will deem samples *near* the origin to be outliers, as these have maximal $\max_y \mathbb{P}(y \mid x)$. This illustrates the distinction between abstentions favoured by L2R (low label certainty) and OOD detection (low density). Setting **(b)** considers *open-set classification* where there are $L = 4$ total classes, with the fourth class (denoted by ▼) assumed to comprise outliers not seen during training. Each class-conditional is an isotropic Gaussian (left). Note that the maximum *inlier* class-probability $\mathbb{P}_{\text{in}}(y \mid x)$ scores OOD samples significantly *higher* than ID samples (right). Thus, the MSP baseline, which declares samples with low $\max_y \mathbb{P}_{\text{in}}(y \mid x)$ as outliers, will perform poorly.

3 L2R meets OOD detection: a shared baseline (and when it fails)

We now show that the L2R and OOD problems share a common baseline; however, this baseline may *fail* for some important special cases of OOD detection.

3.1 Maximum softmax probability meets Chow's rule

For inlier and outlier distributions $\mathbb{P}_{\text{in}}, \mathbb{P}_{\text{out}}$, consider the L2R problem on $\mathbb{P}_{\text{in}}$ and OOD detection on $\mathbb{P}_{\text{in}}, \mathbb{P}_{\text{out}}$. We begin with an intriguing observation: while developed in parallel strands of literature, the *confidence-based L2R baseline and maximum softmax probability OOD baseline are identical*. Indeed, both baselines involve learning probability estimates $\hat{\mathbb{P}}_{\text{in}}(y|x)$, and using the score $\max_{y \in [L]} \hat{\mathbb{P}}_{\text{in}}(y|x)$ to determine whether to reject a sample, or deem it OOD.

Does this imply that one may similarly employ other L2R methods for OOD detection, and vice-versa? Conceptually, L2R and OOD detection appear related: both problems seek to identify certain "abnormal" samples. However, the notion of "abnormal" is subtly different, as may be seen from their Bayes-optimal solutions (2), (4): in L2R, we identify samples where $\max_{y \in [L]} \mathbb{P}_{\text{in}}(y \mid x)$ is small, while in OOD detection, we identify samples where $\frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{out}}(x)}$ is small.

The absence of any consideration of $\mathbb{P}_{\text{out}}(x)$ in L2R suggests the two problems may not be interchangeable after all. However, despite sharing a common baseline, and the two problems generally acknowledged to be related (Yang et al., 2021), there is surprisingly limited understanding of how precisely they relate to (and differ from) each other.

We now show that in two special cases of OOD detection (which impose specific assumptions on $\mathbb{P}_{\text{out}}$), the MSP baseline can be strongly *sub-optimal* for OOD detection.

3.2 MSP may fail for uniform $\mathbb{P}_{\text{out}}$

As a warm-up, we first consider *density-based OOD detection*, where we assume $\mathcal{X} \subset \mathbb{R}^D$ and $\mathbb{P}_{\text{out}}$ is uniform. Assuming without loss of generality that the volume of $\mathcal{X}$ is normalised to 1, the density ratio $\frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{out}}(x)} = \mathbb{P}_{\text{in}}(x)$. Thus, from (4), the Bayes-optimal OOD detector for (3) is simply

$$r^*(x) = \left[\!\!\left[\mathbb{P}_{\text{in}}(x) < \frac{\beta}{\alpha} \right]\!\!\right];$$

that is, we deem *low-density* samples to be OOD. By contrast, the MSP baseline deems samples where $\max_{y \in [L]} \mathbb{P}_{\text{in}}(y \mid x) < t$ to be OOD; that is, we deem *low label-certainty* samples to be worthy of abstention.

How well do these two rules align? Observe that their decisions are based on fundamentally different quantities: the marginal density $\mathbb{P}_{\text{in}}(x)$, and the class-probability $\mathbb{P}_{\text{in}}(y \mid x)$. Consequently, it is not hard to see that these rules might be at odds: for a fixed joint distribution $\mathbb{P}_{\text{in}}(x, y)$, one may independently vary $\mathbb{P}_{\text{in}}(x)$ and $\mathbb{P}_{\text{in}}(y \mid x)$ such that low marginal density corresponds to *high* label-certainty. Figure 4 illustrates this in a simple one-dimensional setting.

The above makes precise observations that are implicit in prior work (e.g., Liu et al. (2020a)), but may appear to be an artifact of the highly specific structure assumed on $\mathbb{P}_{\text{out}}$. However, we now show that even in OOD detection settings that appear tailored to MSP, the latter can fail.

3.3 MSP may fail for open-set classification

Our second setting is *open-set classification* (Scheirer et al., 2013), wherein the OOD samples come from an unseen class. For example, suppose the inlier distribution $\mathbb{P}_{\text{in}}$ has support on all but one class $\mathcal{Y}' = \mathcal{Y} - \{L\}$, and the outlier distribution $\mathbb{P}_{\text{out}}$ has support on the remaining class L. We wish to reject samples from the unseen class at test time. Formally, suppose the underlying distribution over samples from both the known and unseen classes is $\mathbb{P}^*(x, y) = \pi^*(y) \cdot \mathbb{P}^*(x \mid y)$. The outlier distribution is $\mathbb{P}_{\text{out}}(x) = \mathbb{P}^*(x \mid y = L)$, while the training distribution is

$$\mathbb{P}_{\text{in}}(x \mid y) = \mathbb{P}^*(x \mid y) \quad \pi_{\text{in}}(y) = \frac{1(y \neq L)}{1 - \pi^*(L)} \cdot \pi^*(y).$$

That is, the class-conditional distribution remains unchanged, but we rescale the label frequencies to only exclude the contribution of the unseen label.

Now suppose our goal is to minimize the OOD detection error (3) for suitable α, β. The optimal detector is as follows.

Lemma 3.1. *Under the open-set setting, the Bayes-optimal classifier for the OOD detection error* (3) *takes the form:*

$$\begin{aligned} r^*(x) = 1 &\iff \mathbb{P}^*(y = L \mid x) \geq \frac{\alpha}{\alpha + \beta} \\ &\iff \max_{y' \neq L} \mathbb{P}_{\text{in}}(y' \mid x) \geq \frac{\alpha + \beta}{\beta} \cdot \max_{y' \neq L} \mathbb{P}^*(y' \mid x). \end{aligned}$$

Equipped with the Bayes-optimal classifier for the open-set setting, let us now consider the MSP baseline, which deems a sample to be OOD iff $\max_{y' \neq L} \mathbb{P}_{\text{in}}(y' \mid x) < t_{\text{MSP}}$ for some $t_{\text{MSP}} \in [0, 1]$. Importantly, here we consider the class-probability under the *inlier* distribution $\mathbb{P}_{\text{in}}$.

On close inspection at both the classifiers, one can see that MSP can make predictions that starkly disagree with the optimal rule. More specifically, Lemma 3.1 shows that the optimal decision is to reject when the maximum softmax probability (with respect to $\mathbb{P}_{\text{in}}$) is *higher* than some (sample-dependent) threshold. This is however the precise *opposite* of the MSP baseline, which rejects when the maximum probability is *lower* than some threshold.

What is the reason for this stark discrepancy? Intuitively, the issue is that we would like to threshold the maximum of $\mathbb{P}^*(y \mid x)$, *not* $\mathbb{P}_{\text{in}}(y \mid x)$; however, these two distributions may not align, as the latter includes a normalisation term that causes unexpected behaviour when we threshold. We illustrate this with a concrete example (see also Figure 1(b)).

Example (failure of MSP baseline). Consider a setting where the class probabilities $\mathbb{P}^*(y' \mid x)$ are equal for all the known classes $y' \neq L$. This implies that $\mathbb{P}_{\text{in}}(y' \mid x) = \frac{1}{L-1}, \forall y' \neq L$. The Bayes-optimal classifier rejects a sample when $\mathbb{P}^*(L \mid x) > \frac{\alpha}{\alpha+\beta}$. On the other hand, MSP rejects a sample iff the threshold $t_{\text{MSP}} < \frac{1}{L-1}$. Notice that rejection decision is *independent* of the unknown class density $\mathbb{P}^*(L \mid x)$, and therefore will not agree with the Bayes-optimal classifier in general.

The following lemma formalizes the above observation.

Lemma 3.2. *Suppose the MSP baseline is implemented as* $\max_{y \neq L} \mathbb{P}_{\text{in}}(y \mid x) < t_{\text{MSP}}$ *for* $t_{\text{MSP}} \in (0, 1)$. *Then for any choice of* t_{MSP}, *there exists an underlying conditional-class distribution* $\mathbb{P}^*(y \mid x)$ *for which the Bayes-optimal classifier* $\mathbb{P}^*(y = L \mid x) \geq t$ *disagrees with MSP* $\forall t \in (0, 1)$.

3.4 Discussion and extensions

Given the common use (and success) of the MSP for OOD detection, it is natural to ask when it *does* work. One setting where the MSP score will match the Bayes classifier is for all samples x, (i) one of the classes has a sufficiently $\mathbb{P}^*(y \mid x)$, and (ii) when it is the unseen class that receives the highest conditional probability, all other classes have near-equal and small probability, i.e.: $\mathbb{P}^*(y \mid x) \approx \frac{\beta}{\alpha+\beta} \cdot \frac{1}{L-1}$.

One may also ask whether using the maximum *logit* rather than softmax probability can prove successful in the open-set setting. Unfortunately, as this similarly does not include information about $\mathbb{P}_{\text{out}}$, it can also fail; see Appendix D.6. For the same reason, rejectors that threshold the margin between the highest and the second-highest probabilities, instead of the maximum class probability, can also fail. The use of other L2R methods such as Mozannar & Sontag (2020) may not be successful either, because the optimal solutions for these methods have the same form as MSP.

4 Joint L2R and OOD detection

Given the failure of the MSP baseline for general OOD detection, one may ask: is there some other method that may fare better at jointly handling the L2R and OOD detection problems? To answer this question, it is useful to spell out a joint classification risk. We do so for a general setting, where the OOD samples are not necessarily from an unseen class, but can come from an arbitrary distribution $\mathbb{P}_{\text{out}}$.

4.1 OOD-aware misclassification error

Recall that OOD detection concerns learning a detector $r \colon \mathcal{X} \to \{0, 1\}$. For practical deployment in classification settings, it is natural to use this detector to abstain from making a prediction on samples deemed to be OOD (Bitterwolf et al., 2022; Xia & Bouganis, 2022). As with the L2R problem, one may additionally allow for abstention on in-distribution (but "hard") samples. Intuitively, these two abstentions have different costs: e.g., there may be monetary cost for abstaining on an inlier sample, but a cost to reputation due to an embarrassingly wrong prediction for *failing* to abstain on OOD samples.

To formalise this, we assume that a classifier is allowed the additional option of predicting that a sample is OOD, denoted by $\perp$. Furthermore, there is a cost $\alpha \in \mathbb{R}$ associated with classifying an inlier example as OOD, and a cost $\beta \in \mathbb{R}$ associated with classifying an OOD example as inlier. Our goal is to then learn an *abstention-augmented* classifier $\bar{h} \colon \mathcal{X} \to \mathcal{Y} \cup \{\perp\}$ to minimise the following *joint* objective:

$$\begin{aligned} R_{\text{joint}}(\bar{h}) \doteq \; & \mathbb{P}_{\text{in}}(y \neq \bar{h}(x), \bar{h}(x) \neq \perp) \\ & + \alpha \cdot \mathbb{P}_{\text{in}}(\bar{h}(x) = \perp) + \beta \cdot \mathbb{P}_{\text{out}}(\bar{h}(x) \neq \perp). \end{aligned} \tag{5}$$

This simply combines the usual misclassification error of h with explicit penalties for OOD detection errors. Similar objectives have been implicitly considered in Bitterwolf et al. (2022); Katz-Samuels et al. (2022).

Remark. Notice that interpreting the prediction $\perp$ as abstention, the first two terms are precisely the standard learning to reject risk applied to the inlier distribution $\mathbb{P}_{\text{in}}$. In other words, when we set β to 0, the above joint risk reduces to that for L2R, i.e., to the abstention-aware risk in (1).

4.2 OOD detection as sample-dependent Chow's rule

Below, we derive the Bayes-optimal classifier for (5).

Lemma 4.1 (Bayes-optimal classifier for (5))**.** *The minimizer of* (5) *predicts for any $x \in \mathcal{X}$ with $\mathbb{P}_{\text{in}}(x) > 0$:*

$$\bar{h}^*(x) = \begin{cases} \perp & \text{if } \max_{y \in [L]} \mathbb{P}_{\text{in}}(y|x) < t(x) \\ \operatorname{argmax}_{y \in [L]} \mathbb{P}_{\text{in}}(y|x) & \text{else} \end{cases}, \quad \text{where } \; t(x) \doteq 1 - \alpha + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}. \tag{6}$$

For any instance x with $\mathbb{P}_{\text{in}}(x) = 0$, $\bar{h}^*(x) = \perp$.

Similar to the Bayes-optimal classifier in (2) for L2R, the above also thresholds the inlier conditional-class probability to decide which samples to abstain on. A key difference however is that we use an *instance-dependent* threshold: the abstention rule uses both inlier $\mathbb{P}_{\text{in}}(y \mid x)$ *and* the inverse density ratio $\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}$ to decide when to abstain.

Observe that when $\beta = 0$ and $\alpha \in [0, 1]$, we get the Bayes-optimal classifier for learning to reject in (2), with a constant threshold $t(x) = 1 - \alpha$. On the other hand, if $\beta > \alpha \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{out}}(x)}$ for a given x, we will always abstain.

Open-set classification redux. We now revisit the open-set classification problem introduced in Figure 1(b). Recall that here, the maximum inlier class-probabilities $\max_y \mathbb{P}_{\text{in}}(y \mid x)$ are higher for outlier compared to inlier samples. Figure 2 shows the distribution of the sample-dependent thresholds $t(x)$ on $\max_y \mathbb{P}_{\text{in}}(y \mid x)$, as contrasted with the constant threshold $1 - \alpha$ applied by vanilla MSP. The results are intuitive: we place a *higher* threshold on outlier samples, so that they are correctly rejected by $\bar{h}^*$ despite having high $\max_y \mathbb{P}_{\text{in}}(y \mid x)$. In Appendix D.7, we visualise the impact of varying α, β on abstentions in this example.

Recently, Xia & Bouganis (2022) also considered the problem of jointly performing OOD detection, and abstaining on hard samples. However, they did not explicate the failure cases of the MSP baseline; did not explicate the form of the Bayes-optimal solution; and proposed a heuristic solution.

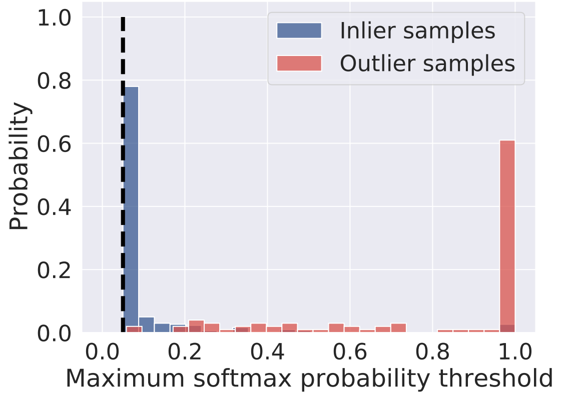


Figure 2: Distribution of instance-dependent maximum softmax probability threshold for the synthetic open-set classification problem in §3.3, for costs $\alpha = 0.95$ and $\beta = 0.1$. As is intuitive, we place a *higher* threshold on outlier as compared to inlier samples, as opposed to the static threshold employed by vanilla MSP (black).

5 Learning a joint OOD-aware classifier

We now turn our attention to learning a classifier that minimizes the joint risk in (5) for user-specified costs α and β. We adopt the setting of Katz-Samuels et al. (2022), and assume access to a labeled inlier sample S_{in} drawn i.i.d. from $\mathbb{P}_{\text{in}}$ and an unlabeled sample, S_{mix} consisting of a mixture of inlier and outlier samples drawn i.i.d. from $\mathbb{P}_{\text{mix}} = \pi_{\text{mix}} \cdot \mathbb{P}_{\text{in}} + (1 - \pi_{\text{mix}}) \cdot \mathbb{P}_{\text{out}}$. We will also assume access to a handful of examples S^*_{in} certified to be strictly inlier, i.e., with $\mathbb{P}_{\text{out}}(x) = 0, \forall x \in S^*_{\text{in}}$.

5.1 A plug-in classification approach

The form of the Bayes-optimal classifier from the previous section suggests a simple two-step procedure to learning a joint classifier: (i) construct estimators for the inlier class probabilities $\mathbb{P}_{\text{in}}(y \mid x)$ and the inverse density ratio $\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}$ using the training sample; (ii) plug estimates of these quantities into (6) to approximate the Bayes classifier.

Surrogate minimization. For the first step, we adopt a surrogate loss from Bitterwolf et al. (2022), to jointly estimate $\mathbb{P}_{\text{in}}(y \mid x)$, and the unlabeled-to-inlier density ratio $\frac{\mathbb{P}_{\text{mix}}(x)}{\mathbb{P}_{\text{in}}(x)}$. Specifically, for classification logits $f \colon \mathcal{X} \to \mathbb{R}^L$ and an OOD scorer $s \colon \mathcal{X} \to \mathbb{R}$, we minimize:

$$R_{\text{dce}}(f, s) \; = \mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{in}}}\left[\ell_{\text{ce}}(y, f(x))\right] + \mathbb{E}_{x\sim\mathbb{P}_{\text{in}}}\left[\ell_{\text{log}}(-1, s(x))\right] + \mathbb{E}_{x\sim\mathbb{P}_{\text{mix}}}\left[\ell_{\text{log}}(+1, s(x))\right], \qquad (7)$$

where $\ell_{\text{ce}} \colon [L] \times \mathbb{R}^L \to \mathbb{R}_+$ is the softmax cross-entropy loss and $\ell_{\text{log}} \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$ is the logistic loss. In words, this the usual softmax cross-entropy loss on the inlier samples with a loss that discriminates between the inlier and the unlabeled samples. It is easy to see that the optimal logits $f^*_y(x) \propto \log(\mathbb{P}_{\text{in}}(y \mid x))$ and $s^*(x) = \log\left(\frac{\mathbb{P}_{\text{mix}}(x)}{\mathbb{P}_{\text{in}}(x)}\right)$. In this step, we do not impose separate costs for the OOD detection errors, as the goal is to estimate a density ratio; see Appendix B.2 for a discussion on alternate surrogate losses.

Let $\hat{f}$ and $\hat{s}$ denote the logits and OOD scorer obtained by minimizing the objective in (7) with samples S_{in} and S_{mix}; we then have estimates $\hat{\mathbb{P}}_{\text{in}}(y|x) \propto \exp(\hat{f}_y(x))$ for the inlier class probabilities, and estimate $\exp(\hat{s}(x))$ for $\frac{\mathbb{P}_{\text{mix}}(x)}{\mathbb{P}_{\text{in}}(x)}$.

Density ratio transformation. To translate our estimate for $\frac{\mathbb{P}_{\text{mix}}(x)}{\mathbb{P}_{\text{in}}(x)}$ to that for $\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}$, we apply a simple transformation:

Algorithm 1 Plug-in classification algorithm

```
**Input:** Labeled set $S_{\text{in}} \sim \mathbb{P}_{\text{in}}$, Unlabeled set $S_{\text{mix}} \sim \mathbb{P}_{\text{mix}}$, Strictly inlier set $S^*_{\text{in}}$ with $\mathbb{P}_{\text{out}}(x) = 0, \forall x \in S^*_{\text{in}}$
**Parameters:** Costs $\alpha, \beta$
**Surrogate opt:** Solve $\hat{f}, \hat{s} \in \underset{f,r}{\operatorname{argmin}}\, \hat{R}_{\text{dce}}(f, s)$,
    where $\hat{R}_{\text{dce}}(f, r) \doteq \frac{1}{|S_{\text{in}}|} \sum_{(x,y) \in S_{\text{in}}} \ell_{\text{ce}}(y, f(x))$
        $+ \frac{1}{|S_{\text{in}}|} \sum_{(x,y) \in S_{\text{in}}} \ell_{\log}(-1, s(x)) + \frac{1}{|S_{\text{mix}}|} \sum_{x \in S_{\text{mix}}} \ell_{\log}(-1, s(x))$.
**Inlier class probabilities:** $\hat{\mathbb{P}}_{\text{in}}(y|x) \doteq \frac{\exp(\hat{f}_y(x))}{\sum_{y'} \exp(\hat{f}_{y'}(x))}$
**Mixture proportion:** $\hat{\pi}_{\text{mix}} \doteq \frac{1}{|S^*_{\text{in}}|} \sum_{x \in S^*_{\text{in}}} \exp(\hat{s}(x))$
**Density ratio:** $\hat{\gamma}_{\text{ood}}(x) \doteq \frac{1}{1-\hat{\pi}_{\text{mix}}} \cdot (\exp(\hat{s}(x)) - \hat{\pi}_{\text{mix}})$
**Plug-in classifier:** Plug estimates $\hat{\mathbb{P}}_{\text{in}}(y|x)$, $\hat{\gamma}_{\text{ood}}(x)$, and costs $\alpha, \beta$ into (8), and construct classifier $\hat{h}$
**Output:** $\hat{h}$
```

Lemma 5.1. *For any* $\pi_{\text{mix}} < 1$,

$$\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)} = \frac{1}{1 - \pi_{\text{mix}}} \cdot \left(\frac{\mathbb{P}_{\text{mix}}(x)}{\mathbb{P}_{\text{in}}(x)} - \pi_{\text{mix}} \right).$$

The above transformation requires knowing the mixing proportion π_{mix} of inlier samples in the unlabeled dataset. To estimate π_{mix}, we observe that for a example x that is deemed to be strictly inlier, i.e., for which $\mathbb{P}_{\text{out}}(x) = 0$, we have $\frac{\mathbb{P}_{\text{mix}}(x)}{\mathbb{P}_{\text{in}}(x)} = \pi_{\text{mix}}$. Therefore, using the small sample S^*_{in} of strictly inlier examples, we can estimate $\hat{\pi}_{\text{mix}} = \frac{1}{|S^*_{\text{in}}|} \sum_{x \in S^*_{\text{in}}} \exp(\hat{s}(x))$. The inverse density ratio can then be estimated as $\hat{\gamma}_{\text{ood}}(x) = \frac{1}{1-\hat{\pi}_{\text{mix}}} \cdot (\exp(\hat{s}(x)) - \hat{\pi}_{\text{mix}})$.

Plug-in classifier. Plugging the above estimates into (6) then gives us an approximation to the Bayes classifier:

$$\hat{h}(x) = \begin{cases} \perp & \text{if } \max_{y \in [L]} \hat{\mathbb{P}}_{\text{in}}(y \mid x) < \hat{t}(x) \\ \underset{y \in [L]}{\operatorname{argmax}}\, \hat{f}_y(x) & \text{else} \end{cases}, \quad \text{where } \hat{t}(x) \doteq 1 - \alpha + \beta \cdot \hat{\gamma}_{\text{ood}}(x). \tag{8}$$

We summarize the details of this procedure in Algorithm 1.

5.2 Incorporating budget constraints

In practice, it may be more natural for a user to specify a budget on the total amount of abstentions, rather than a cost α on the inlier abstention rate. Our formulation easily accommodates this change. More formally, suppose we want to learn a classifier $\bar{h} : \mathcal{X} \to [L] \cup \{\perp\}$ to solve:

$$\begin{aligned} \min_{h}\ & \mathbb{P}_{\text{in}}(y \neq \bar{h}(x), \bar{h}(x) \neq \perp) + c_{\text{ood}} \cdot \mathbb{P}_{\text{out}}(\bar{h}(x) \neq \perp) \\ \text{s.t. }\ & \mathbb{P}^*(\bar{h}(x) = \perp) \leq B_{\text{abs}}, \end{aligned} \tag{9}$$

where $\mathbb{P}^*(x) = (1 - \pi_{\text{out}}) \cdot \mathbb{P}_{\text{in}}(x) + \pi_{\text{out}} \cdot \mathbb{P}_{\text{out}}(x)$ denotes the target distribution with known proportions of inlier and OOD samples, $B_{\text{abs}} \in [0, 1]$ is the abstention budget, and $c_{\text{ood}} \in \mathbb{R}_+$ is the cost for failing to abstain on OOD samples. For a large cost c_{ood}, this constrained problem encourages a classifier to abstain first on OOD samples, and then use the remaining budget to abstain on "hard" inlier samples.

Appealing to Lagrangian theory, we can show that the above problem is equivalent to minimizing the joint error in (5) for a particular choice of the cost parameters.

$$\mathbb{P}_{\text{in}}(y \neq \bar{h}(x), \bar{h}(x) \neq \perp) + \lambda \cdot (1 - \pi_{\text{out}}) \cdot \mathbb{P}_{\text{in}}(\bar{h}(x) = \perp) + (c_{\text{ood}} - \lambda \cdot \pi_{\text{out}}) \cdot \mathbb{P}_{\text{out}}(\bar{h}(x) \neq \perp),$$

for Lagrange multiplier $\lambda \in \mathbb{R}_+$.

Note that for any fixed λ, we can use the procedure in the Section 5.1 to construct a plug-in classifier that minimizes the above joint risk. Finding the optimal λ then reduces to an inexpensive threshold search. In fact, we only need to implement the surrogate minimization step of the plug-in procedure once to estimate the relevant probabilities; we can then construct multiple plug-in classifiers for different values of λ, and choose among those satisfying the budget constraint, the one that minimizes the objective in (9).

Remark. The previous work of Katz-Samuels et al. (2022) also seeks to solve an optimization problem with explicit constraints on abstention rates. However, there are some subtle, but important, technical differences between their formulation and ours, which we explain in Appendix C.

6 Experimental results

We evaluate the efficacy of the plug-in approach in solving the constrained classification task in (9) on benchmark image classification datasets (Hendrycks & Gimpel, 2017; Bitterwolf et al., 2022). The majority of our evaluations will be on the setting described in Section 5, where during training, we receive an unlabeled sample from $\mathbb{P}_{\text{mix}} = \pi_{\text{mix}} \cdot \mathbb{P}_{\text{in}} + (1 - \pi_{\text{mix}}) \cdot \mathbb{P}_{\text{out}}$, and during test time, we evaluate the classifier on a sample drawn from $\mathbb{P}^{\text{eval}}_{\text{out}} = \mathbb{P}_{\text{out}}$. Following Katz-Samuels et al. (2022), we also consider a variant of this setting, where the test OOD distribution is different from the one available during training, i.e., $\mathbb{P}^{\text{eval}}_{\text{out}} \neq \mathbb{P}_{\text{out}}$.

Datasets. We use CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and ImageNet (Deng et al., 2009) as the inlier datasets, and SVHN (Netzer et al., 2011), Places365 (Zhou et al., 2017), CelebA (Liu et al., 2015) and OpenImages (Krasin et al., 2017) as the OOD datasets. We train ResNet-56 on CIFAR and EfficientNet-B0 on ImageNet. See Appendix D for details about the hyper-parameter choices.

Mixing proportions. The training set we use contains 50% labeled samples and 50% unlabeled samples, of which π_{in} fraction of examples are inlier. We hold out 5% of the original inlier test set and use it as the "strictly inlier" sample needed to estimate the proportion π_{mix} of inlier samples in the unlabeled set. We repeat the experiments for datasets generated with different values of π_{mix}. In addition to this, we also vary the proportion $\pi^{\text{eval}}_{\text{out}}$ of OOD samples in the test set, a quantity we assume is known to the trainer.

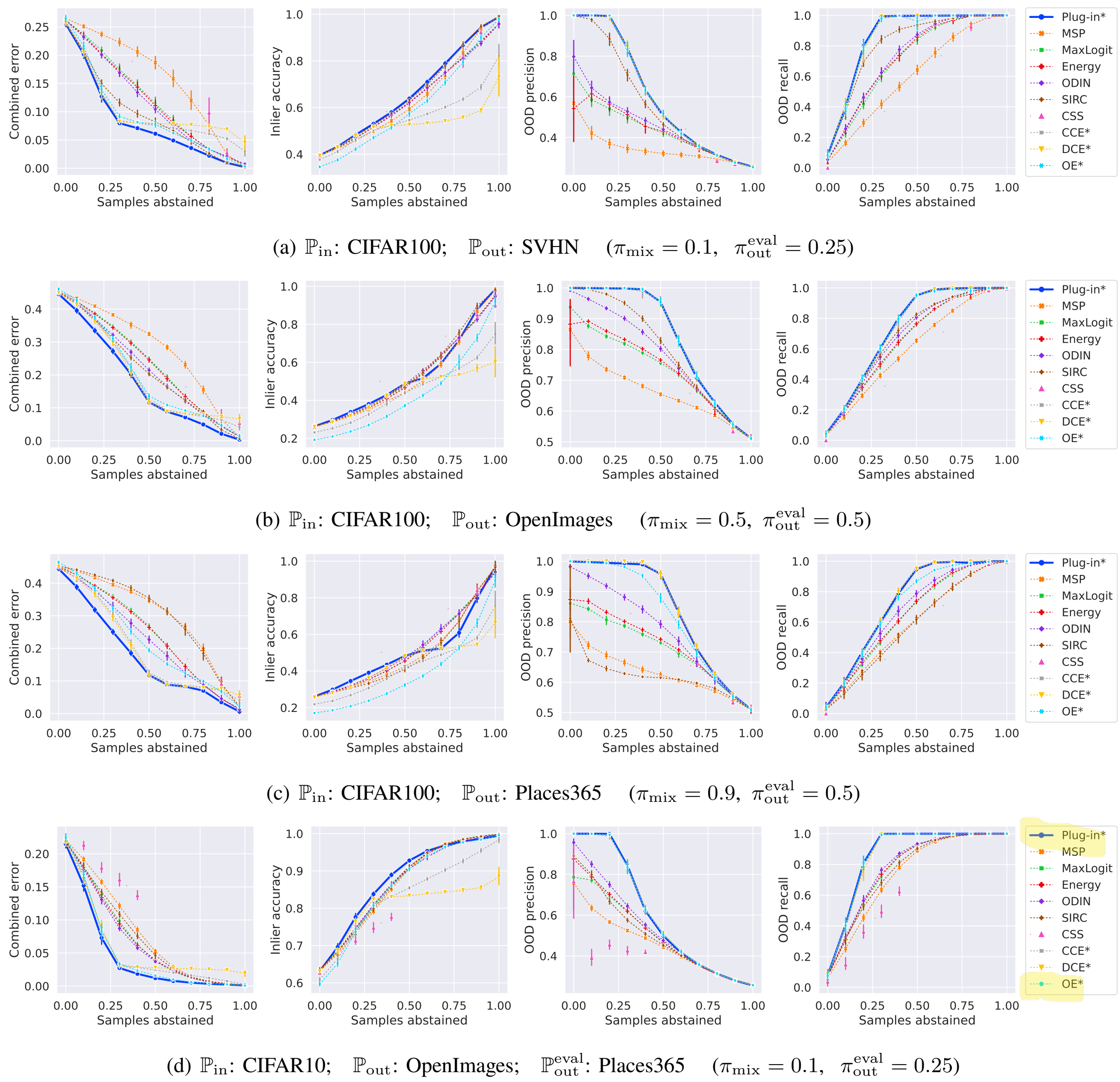
Evaluation metrics. Our classifiers are trained to satisfy a budget constraint on the fraction of abstentions. For a fixed proportion of abstentions, we measure both the classification and OOD detection performance using the following metrics:

$$\text{inlier-acc}(\bar{h}) = \frac{\sum_{(x,y) \in S_{\text{in}}} [\![\bar{h}(x) \neq \perp, y = \bar{h}(x)]\!]}{\sum_{x \in S_{\text{all}}} [\![\bar{h}(x) \neq \perp]\!]}$$

$$\text{ood-precision}(\bar{h}) = \frac{\sum_{(x,y) \in S_{\text{out}}} [\![\bar{h}(x) = \perp]\!]}{\sum_{x \in S_{\text{all}}} [\![\bar{h}(x) = \perp]\!]}$$

$$\text{ood-recall}(\bar{h}) = \frac{\sum_{x \in S_{\text{out}}} [\![\bar{h}(x) = \perp]\!]}{|S_{\text{out}}|},$$

where $S_{\text{all}} = \{x : (x, y) \in S_{\text{in}}\} \cup S_{\text{out}}$ is the combined set of inlier and OOD instances. The *conditional inlier accuracy* is the fraction of non-abstained samples on which the prediction was correct. The *OOD precision* is



(a) $\mathbb{P}_{\text{in}}$: CIFAR100; $\mathbb{P}_{\text{out}}$: SVHN $(\pi_{\text{mix}} = 0.1,\ \pi^{\text{eval}}_{\text{out}} = 0.25)$

(b) $\mathbb{P}_{\text{in}}$: CIFAR100; $\mathbb{P}_{\text{out}}$: OpenImages $(\pi_{\text{mix}} = 0.5,\ \pi^{\text{eval}}_{\text{out}} = 0.5)$

(c) $\mathbb{P}_{\text{in}}$: CIFAR100; $\mathbb{P}_{\text{out}}$: Places365 $(\pi_{\text{mix}} = 0.9,\ \pi^{\text{eval}}_{\text{out}} = 0.5)$

(d) $\mathbb{P}_{\text{in}}$: CIFAR10; $\mathbb{P}_{\text{out}}$: OpenImages; $\mathbb{P}^{\text{eval}}_{\text{out}}$: Places365 $(\pi_{\text{mix}} = 0.1,\ \pi^{\text{eval}}_{\text{out}} = 0.25)$

Figure 3: Plots of classification and OOD detection metrics as a function of the fraction of abstained samples (averaged over 5 trials). These include both stationary settings (a–c) with $\mathbb{P}^{\text{eval}}_{\text{out}} = \mathbb{P}_{\text{out}}$, and a non-stationary setting with $\mathbb{P}^{\text{eval}}_{\text{out}} \neq \mathbb{P}_{\text{out}}$, for different proportions of inlier samples π_{mix} in the unlabeled set, and different proportions of OOD samples $\pi^{\text{eval}}_{\text{out}}$ in the test set. For the conditional risk, *lower values are better*. For all other metrics, *higher values are better*. A '*' against a method indicates it trains on both inlier and unlabeled samples. Higher value of π_{mix} make the unlabeled sample less reliable. We set $c_{\text{ood}} = 5$. In Appendix D.5, we report AUC and FPR95.

the fraction of abstained samples which were OOD. The *OOD recall* is the fraction of OOD samples on which the model abstained.

To measure the combined performance on the classification and OOD detection tasks, we additionally evaluate the *OOD-weighted combined error* in (5). Below is a version of the metric conditioned on the non-abstained samples, and normalized to not exceed 1:

$$\text{comb-error}(\bar{h}) = \frac{1}{Z} \cdot \Big(\sum_{(x,y) \in S_{\text{in}}} [\![\bar{h}(x) \neq \perp, y = \bar{h}(x)]\!] + \ c_{\text{ood}} \cdot \sum_{x \in S_{\text{out}}} [\![\bar{h}(x) \neq \perp]\!] \Big),$$

where $Z = (1 + c_{\text{ood}}) \cdot \sum_{x \in S_{\text{all}}} [\![\bar{h}(x) \neq \perp]\!]$. In all our experiments, we set $c_{\text{ood}} = 5$. We also report ranking metrics such as AUC for the learned OOD detection scorers.

Baselines. We include Chow's rule (or MSP) and the cost-sensitive softmax (CSS) loss of (Mozannar & Sontag, 2020) as representative baselines from the L2R literature. Both these methods use only the inlier samples for training. From the OOD literature, we include both methods which train only on the inlier samples, namely, MaxLogit (Hendrickx et al., 2021), energy-based scorer (Hendrickx et al., 2021), ODIN (Hendrickx et al., 2021), and SIRC (Xia & Bouganis, 2022), and methods which additionally use unlabeled samples, such as the coupled CE loss (CCE) of Thulasidasan et al. (2021), the de-coupled CE loss (DCE) of Bitterwolf et al. (2022), and the outlier exposure (OE) loss of Hendrycks et al. (2019). The closest among these to our approach is SIRC, which also seeks to treat inlier abstentions differently from OOD abstentions, but uses a heuristic post-hoc rule.

With each baseline, we tune the relevant threshold or cost parameter to achieve the desire rate of abstention on a held-out validation set (more details in Appendix D). For the plug-in method, as noted in Section 5.2, we vary the parameter λ to control the amount of abstentions.

Results. We present representative results in Figure 3 for both the stationary setting, where $\mathbb{P}_{\text{out}}^{\text{eval}} = \mathbb{P}_{\text{out}}$ ((a)–(c)), and the non-stationary setting, where $\mathbb{P}_{\text{out}}^{\text{eval}} \neq \mathbb{P}_{\text{out}}$ (d), and include additional results in Appendix D. In each case, we plot the classification and OOD metrics as a function of the percentage of abstained samples.

One can see a few general trends. The combined error decreases with more abstentions; the inlier accuracy increases with abstentions. The OOD precision is the highest initially when the abstentions are on the OOD samples, but decreases when the OOD samples are exhausted, and the abstentions are on the inlier samples; the opposite is true for OOD recall.

The main takeaway from these results is that, the proposed plug-in method is competitive for most operating points. It yields *lower or similar values in combined weighted error* compared to the baselines, while *matching the best performing OOD detectors* on precision and recall. In contrast, the baselines either perform well on classification accuracy, or on the OOD metrics, but usually not on both. Perhaps unsurprisingly, the methods that only use the inlier sample for training (e.g., MSP or ODIN) fare well on the inlier accuracy, but under-perform on the OOD metrics. The methods which use the unlabeled samples for training (CCE, DCE, OE) fare well on the OOD metrics, but suffer on the inlier accuracy in the higher abstention regime.

Finally, in keeping with our sub-optimality results, MSP is seen to under-perform on the OOD metrics. However, it is among the best performing methods on inlier accuracy in settings where π_{mix} is high (i.e., where the unlabeled training set contains a higher fraction of inlier samples). While in these high-noise settings, the plug-in approach sometimes yields a lower inlier accuracy than methods trained only on the inlier samples, it is still very competitive on the combined (weighted) error metric; this showcases its efficacy at optimizing the desired trade-off between classification and OOD detection performance.

7 Discussion and future work

We have thus provided a general framework that unifies L2R and OOD detection, and proposed a plug-in classifier for learning to jointly abstain on both inlier and OOD samples. While our analysis builds on the MSP baseline, it is worth noting that there are many other effective OOD detection approaches; in future work, it would be of interest to further study the viability of augmenting these approaches with a L2R capability. For example, these include approaches based on k-nearest neighbour (Sun et al., 2022), the likelihood ratio against a background sample (Ren et al., 2019), and methods that use a sample of "outliers" to train softmax based classifiers (Dhamija et al., 2018; Chen et al., 2021). It would also be interesting to *improve* the OOD generalization of the learned classifier, and study if this may help in identifying samples worthy of rejection.

References

Bartlett, P. L. and Wegkamp, M. H. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008.

Bendale, A. and Boult, T. E. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

Bitterwolf, J., Meinke, A., Augustin, M., and Hein, M. Breaking down out-of-distribution detection: Many methods based on OOD training data estimate a combination of the same core quantities. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2041–2074. PMLR, 17–23 Jul 2022.

Charoenphakdee, N., Cui, Z., Zhang, Y., and Sugiyama, M. Classification with rejection based on cost-sensitive classification. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1507–1517. PMLR, 18–24 Jul 2021.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. Atom: Robustifying out-of-distribution detection using outlier mining. *In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021.

Chow, C. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1): 41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Cortes, C., DeSalvo, G., and Mohri, M. Boosting with abstention. *Advances in Neural Information Processing Systems*, 29:1660–1668, 2016a.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhamija, A. R., Günther, M., and Boult, T. E. Reducing network agnostophobia. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9175–9186, Red Hook, NY, USA, 2018. Curran Associates Inc.

Elkan, C. The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001.

Gangrade, A., Kag, A., and Saligrama, V. Selective classification via one-sided prediction. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2179–2187. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/gangrade21a.html.

Hendrickx, K., Perini, L., der Plas, D. V., Meert, W., and Davis, J. Machine learning with a reject option: A survey. *CoRR*, abs/2107.11277, 2021.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., and Song, D. Scaling out-of-distribution detection for real-world settings. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8759–8773. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hendrycks22a.html.

Katz-Samuels, J., Nakhleh, J. B., Nowak, R., and Li, Y. Training OOD detectors in their natural habitats. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10848–10865. PMLR, 17–23 Jul 2022.

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryiAv2xAZ.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Deng, H., and Ju, Q. FastBERT: a self-distilling bert with adaptive inference time. In *Proceedings of ACL 2020*, 2020b.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Mozannar, H. and Sontag, D. Consistent estimators for learning to defer to an expert. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 13–18 Jul 2020.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=H1xwNhCcYm.

Narasimhan, H., Jitkrittum, W., Menon, A. K., Rawat, A. S., and Kumar, S. Post-hoc estimators for learning to defer to an expert. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_jg6Sf6tuF7.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015. doi: 10.1109/CVPR.2015.7298640.

Ni, C., Charoenphakdee, N., Honda, J., and Sugiyama, M. On the calibration of multiclass classification with rejection. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2582–2592, 2019.

Okati, N., De, A., and Rodriguez, M. Differentiable learning under triage. *Advances in Neural Information Processing Systems*, 34, 2021.

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., and Mullainathan, S. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019a.

Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, B., Mullainathan, S., and Kleinberg, J. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pp. 5281–5290. PMLR, 2019b.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530 – 554, 2018. doi: 10.1214/17-EJS1388.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. *Likelihood Ratios for Out-of-Distribution Detection*, pp. 14707—-14718. Curran Associates Inc., Red Hook, NY, USA, 2019.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., and Boult, T. E. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Scott, C. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6(none):958 – 992, 2012. doi: 10.1214/12-EJS699. URL https://doi.org/10.1214/12-EJS699.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. Combating label noise in deep learning using abstention. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6234–6243, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Thulasidasan, S., Thapa, S., Dhaubhadel, S., Chennupati, G., Bhattacharya, T., and Bilmes, J. A. An effective baseline for robustness to distributional shift. *CoRR*, abs/2105.07107, 2021. URL https://arxiv.org/abs/2105.07107.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021.

Verma, R. and Nalisnick, E. Calibrated learning to defer with one-vs-all classifiers. *arXiv preprint arXiv:2202.03673*, 2022.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. Mitigating neural network overconfidence with logit normalization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23631–23644. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wei22d.html.

Xia, G. and Bouganis, C.-S. Augmenting softmax information for selective classification with out-of-distribution data. *ArXiv*, abs/2207.07506, 2022.

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

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A Proofs

We will find it useful to derive the following quantities.

$$\begin{aligned}
\mathbb{P}_{\text{in}}(x, y) &= \pi_{\text{in}}(y) \cdot \mathbb{P}_{\text{in}}(x \mid y) \\
&= \frac{1(y \neq L)}{1 - \pi^*(L)} \cdot \pi^*(y) \cdot \mathbb{P}^*(x \mid y) \\
&= \frac{1(y \neq L)}{1 - \pi^*(L)} \cdot \mathbb{P}^*(x, y) \\
\mathbb{P}_{\text{in}}(x) &= \sum_{y \in [L]} \mathbb{P}_{\text{in}}(x, y) \\
&= \sum_{y \in [L]} \pi_{\text{in}}(y) \cdot \mathbb{P}_{\text{in}}(x \mid y) \\
&= \frac{1}{1 - \pi^*(L)} \sum_{y \neq L} \pi^*(y) \cdot \mathbb{P}^*(x \mid y) \\
&= \frac{1}{1 - \pi^*(L)} \sum_{y \neq L} \mathbb{P}^*(y \mid x) \cdot \mathbb{P}^*(x) \\
&= \frac{\mathbb{P}^*(y \neq L \mid x)}{1 - \pi^*(L)} \cdot \mathbb{P}^*(x) \\
\mathbb{P}_{\text{in}}(y \mid x) &= \frac{\mathbb{P}_{\text{in}}(x, y)}{\mathbb{P}_{\text{in}}(x)} \\
&= \frac{1(y \neq L)}{1 - \pi^*(L)} \cdot \frac{1 - \pi^*(L)}{\mathbb{P}^*(y \neq L \mid x)} \cdot \frac{\mathbb{P}^*(x, y)}{\mathbb{P}^*(x)} \\
&= \frac{1(y \neq L)}{\mathbb{P}^*(y \neq L \mid x)} \cdot \mathbb{P}^*(y \mid x).
\end{aligned}$$

Proof of Lemma 3.1. The first part follows from standard results in cost-sensitive learning (Elkan, 2001):

$$r^*(x) = 1 \iff \mathbb{P}^*(y = L \mid x) \geq \frac{\alpha}{\alpha + \beta}.$$

We further have for threshold $t_{\text{OPT}} \doteq \frac{\alpha}{\alpha+\beta}$,

$$\begin{aligned}
\mathbb{P}^*(y = L \mid x) \geq t_{\text{OPT}} &\iff \mathbb{P}^*(y \neq L \mid x) \leq 1 - t_{\text{OPT}} \\
&\iff \frac{1}{\mathbb{P}^*(y \neq L \mid x)} \geq \frac{1}{1 - t_{\text{OPT}}} \\
&\iff \frac{\max_{y' \neq L} \mathbb{P}^*(y' \mid x)}{\mathbb{P}^*(y \neq L \mid x)} \geq \frac{\max_{y' \neq L} \mathbb{P}^*(y' \mid x)}{1 - t_{\text{OPT}}} \\
&\iff \max_{y' \neq L} \mathbb{P}_{\text{in}}(y' \mid x) \geq \frac{\max_{y' \neq L} \mathbb{P}^*(y' \mid x)}{1 - t_{\text{OPT}}}.
\end{aligned}$$

That is, we want to reject when the maximum softmax probability is *higher* than some (sample-dependent) threshold. □

Proof of Lemma 3.2. Let's fix $\epsilon \in (0, 1)$. We consider two cases for threshold t_{chow}:

Case (i): $t_{\text{chow}} \leq \frac{1}{L-1}$. Consider a distribution where for all instances x, $\mathbb{P}^*(y = L \mid x) = 1 - \epsilon$ and $\mathbb{P}^*(y' \mid x) = \frac{\epsilon}{L-1}, \forall y' \neq L$. Then the Bayes-optimal classifier accepts any instance x for all thresholds $t \in \big(0, 1-\epsilon\big)$. In contrast, Chow's rule would compute $\max_{y \neq L} \mathbb{P}_{\text{in}}(y \mid x) = \frac{1}{L-1}$, and thus reject all instances x.

Case (ii): $t_{\text{chow}} > \frac{1}{L-1}$. Consider a distribution where for all instances x, $\mathbb{P}^*(y = L \mid x) = \epsilon$ and $\mathbb{P}^*(y' \mid x) = \frac{1-\epsilon}{L-1}, \forall y' \neq L$. Then the Bayes-optimal classifier would reject any instance x for thresholds $t \in \big(\epsilon, 1\big)$, whereas Chow's rule would accept all instances.

Taking $\epsilon \to 0$ completes the proof. □

Proof of Lemma 4.1. We first define a joint marginal distribution $\mathbb{P}_{\text{comb}}$ that samples from $\mathbb{P}_{\text{in}}(x)$ and $\mathbb{P}_{\text{out}}(x)$ with equal probabilities. We then rewrite the objective in (5) in terms of the joint marginal distribution:

$$\begin{aligned}
&R_{\text{joint}}(h) \\
&= \mathbb{E}_{x \sim \mathbb{P}_{\text{comb}}}\Bigg[\mathbb{E}_{y|x \sim \mathbb{P}_{\text{in}}}\left[\frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \mathbf{1}(y \neq h(x), h(x) \neq \perp)\right] + \alpha \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \mathbf{1}(h(x) = \perp) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \mathbf{1}(h(x) \neq \perp)\Bigg] \\
&= \mathbb{E}_{x \sim \mathbb{P}_{\text{comb}}}\Bigg[\sum_{y \in [L]} \mathbb{P}_{\text{in}}(y|x) \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \mathbf{1}(y \neq h(x), h(x) \neq \perp) + \alpha \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \mathbf{1}(h(x) = \perp) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \mathbf{1}(h(x) \neq \perp)\Bigg].
\end{aligned}$$

The conditional risk that a classifier h incurs when abstaining (i.e., predicting $\perp$) on a fixed instance x is given by:

$$\alpha \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)}.$$

The conditional risk associated with predicting a base class $y \in [L]$ on instance x is given by:

$$\frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot (1 - \mathbb{P}_{\text{in}}(y|x)) + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{comb}}(x)}$$

The Bayes-optimal classifier then predicts the label with the lowest conditional risk. When $\mathbb{P}_{\text{in}}(x) = 0$, this amounts to predicting abstain $\perp$. When $\mathbb{P}_{\text{in}}(x) > 0$, the optimal classifier predicts $\perp$ when:

$$\begin{gathered}
\alpha \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} < \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \min_{y \in [L]} (1 - \mathbb{P}_{\text{in}}(y|x)) + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{comb}}(x)} \\
\Longleftrightarrow \\
\alpha \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} < \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} - \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \max_{y \in [L]} \mathbb{P}_{\text{in}}(y|x) + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{comb}}(x)} \\
\Longleftrightarrow \\
\frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \max_{y \in [L]} \mathbb{P}_{\text{in}}(y|x) < \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} - \alpha \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{comb}}(x)} \\
\Longleftrightarrow \\
\max_{y \in [L]} \mathbb{P}_{\text{in}}(y|x) < 1 - \alpha + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}.
\end{gathered}$$

Otherwise, the classifier predicts $\operatorname{argmax}_{y \in [L]} \mathbb{P}_{\text{in}}(y|x)$, as desired. □

Proof of Lemma 5.1. Expanding the right-hand side, we have:

$$\frac{1}{1-\pi_{\text{mix}}} \cdot \left(\frac{\mathbb{P}_{\text{mix}}(x)}{\mathbb{P}_{\text{in}}(x)} - \pi_{\text{mix}} \right) = \frac{1}{1-\pi_{\text{mix}}} \cdot \left(\frac{\pi_{\text{mix}} \cdot \mathbb{P}_{\text{in}}(x) + (1-\pi_{\text{mix}}) \cdot \mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)} - \pi_{\text{mix}} \right) = \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)},$$

as desired. □

B OOD detection objectives

Suppose we have two sets of samples: a labelled sample from an "inlier" distribution $\mathbb{P}_{\text{in}}(x, y)$, and an unlabelled sample from an out-of-distribution (OOD) "outlier" distribution $\mathbb{P}_{\text{out}}(x)$. Our goal is to estimate the inlier probability model $\mathbb{P}_{\text{in}}(y \mid x)$, as well as the likelihood ratio $\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}$ of any sample being an outlier. Intuitively, at inference time we would like to abstain from making a prediction on outlier samples.

B.1 Decoupled objective

The surrogate objective we employ in our plug-in classification method seeks to learn logits $f \colon \mathcal{X} \to \mathbb{R}^L$, and a scorer $s \colon \mathcal{X} \to \mathbb{R}$ to minimise

$$R_{\text{dce}}(f, s) = \mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{in}}} \left[\ell_{\text{ce}}(y, f(x))\right] + \mathbb{E}_{x\sim\mathbb{P}_{\text{in}}} \left[\ell_{\text{bin}}(-1, s(x))\right] + \mathbb{E}_{x\sim\mathbb{P}_{\text{out}}} \left[\ell_{\text{bin}}(+1, s(x))\right],$$

where $\ell_{\text{ce}} \colon [L] \times \mathbb{R}^L \to \mathbb{R}_+$ is the softmax cross-entropy and $\ell_{\text{bin}} \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$ is a binary surrogate loss. In words, we combine the usual softmax cross-entropy loss on the inlier samples with a loss that discriminates between the inlier and outlier samples. Such an objective was considered in Bitterwolf et al. (2022). It is easy to see that the Bayes-optimal logits will yield probabilities $p_y^*(x) = \mathbb{P}_{\text{in}}(y \mid x)$. Further, for strictly proper composite ℓ_{bin}, the Bayes-optimal $s^*(x)$ will yield an invertible transformation of the likelihood ratio $\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}$.

B.2 Coupled objective

Another natural objective learns logits $\bar{f} \colon \mathcal{X} \to \mathbb{R}^{L+1}$ so as to minimise (Thulasidasan et al., 2021)

$$\begin{aligned} R_{\text{cce}}(\bar{f}) &= \pi_{\text{in}} \cdot \mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{in}}} \left[\ell_{\text{ce}}(y, \bar{f}(x))\right] + (1-\pi_{\text{in}}) \cdot \mathbb{E}_{x\sim\mathbb{P}_{\text{out}}} \left[\ell_{\text{ce}}(\perp, \bar{f}(x))\right] \\ &= \mathbb{E}_{(x,\bar{y})\sim\mathbb{P}_{\text{comb}}} \left[\ell_{\text{ce}}(\bar{y}, \bar{f}(x))\right], \end{aligned}$$

where π_{in} is the fraction of inliers versus outliers, and $\mathbb{P}_{\text{comb}}$ is a mixture over inliers and outliers given by

$$\begin{aligned} \mathbb{P}_{\text{comb}}(\bar{y}) &= \pi_{\text{in}} \cdot \mathbb{P}_{\text{in}}(\bar{y}) \text{ if } \bar{y} \neq \perp \\ \mathbb{P}_{\text{comb}}(\bar{y}) &= 1 - \pi_{\text{in}} \qquad \text{if } \bar{y} = \perp \\ \mathbb{P}_{\text{comb}}(x \mid \bar{y}) &= \mathbb{P}_{\text{in}}(x \mid \bar{y}) \quad \text{if } \bar{y} \neq \perp \\ \mathbb{P}_{\text{comb}}(x \mid \bar{y}) &= \mathbb{P}_{\text{out}}(x) \qquad \text{if } \bar{y} = \perp. \end{aligned}$$

Here, we will obtain probabilities $\bar{p}_{\bar{y}}^*(x) = \mathbb{P}_{\text{comb}}(\bar{y} \mid x)$, which simplifies to

$$\begin{aligned} \mathbb{P}_{\text{comb}}(\bar{y} \mid x) &= \frac{\mathbb{P}_{\text{comb}}(x \mid \bar{y}) \cdot \mathbb{P}_{\text{comb}}(\bar{y})}{\mathbb{P}_{\text{comb}}(x)} \\ &= \begin{cases} \mathbb{P}_{\text{in}}(\bar{y} \mid x) \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot \pi_{\text{in}} & \text{if } \bar{y} \neq \perp \\ \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{comb}}(x)} \cdot (1 - \pi_{\text{in}}) & \text{if } \bar{y} = \perp. \end{cases} \end{aligned}$$

Observe that $\mathbb{P}_{\text{comb}}(\perp \mid x)$ estimates a transformation of the likelihood ratio $\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}$: indeed,

$$\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{comb}}(x)} = \frac{\mathbb{P}_{\text{out}}(x)}{\pi_{\text{in}} \cdot \mathbb{P}_{\text{in}}(x) + (1-\pi_{\text{in}}) \cdot \mathbb{P}_{\text{out}}(x)} = \frac{\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}}{\pi_{\text{in}} + (1-\pi_{\text{in}}) \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}}.$$

It is however perhaps less desirable for the regular class-probability estimates $\mathbb{P}_{\text{comb}}(y \mid x)$ to be additionally modified by $\frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{comb}}(x)}$: for inlier samples, this means that we do not actually get estimates of $\mathbb{P}_{\text{in}}(y \mid x)$, unlike the previous formulation. For this reason, we choose the de-coupled CCE objective for our plug-in algorithm.

C Relationship to Katz-Samuels et al. (2022)

Like us, Katz-Samuels et al. (2022) also seek to jointly learn a classifier and an OOD scorer, with constraints on the classification and abstention rates, given access to samples from $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{mix}}$. For a joint classifier $\bar{h} : \mathcal{X} \to [L] \cup \{\perp\}$, their formulation can be written as:

$$\begin{aligned} & \min_{h} \ \mathbb{P}_{\text{out}}\left(\bar{h}(x) \neq \perp\right) \\ \text{s.t.} \quad & \mathbb{P}_{\text{in}}\left(\bar{h}(x) = \perp\right) \leq \kappa \\ & \mathbb{P}_{\text{in}}\left(\bar{h}(x) \neq y, \, \bar{h}(x) \neq \perp\right) \leq \tau, \end{aligned} \tag{10}$$

for given targets $\kappa, \tau \in (0,1)$.

While $\mathbb{P}_{\text{out}}$ is not directly available, Katz-Samuels et al. provide a simple solution to solving (10) using only access to $\mathbb{P}_{\text{mix}}$ and $\mathbb{P}_{\text{in}}$. They show that under some mild assumptions, replacing $\mathbb{P}_{\text{out}}$ with $\mathbb{P}_{\text{mix}}$ in the above problem does not alter the optimal solution. The intuition behind this is that when the first constraint on the inlier abstention rate is satisfied with equality, we have $\mathbb{E}_{x \sim \mathbb{P}_{\text{mix}}}\left[\mathbf{1}\left(h(x) \neq \perp\right)\right] = \pi_{\text{mix}} \cdot (1-\alpha) + (1-\pi_{\text{mix}}) \cdot \mathbb{E}_{x \sim \mathbb{P}_{\text{out}}}\left[\mathbf{1}\left(h(x) \neq \perp\right)\right]$, and minimizing this objective is equivalent to minimizing the OOD objective in (10).

This simple trick of replacing $\mathbb{P}_{\text{out}}$ with $\mathbb{P}_{\text{mix}}$ will only work when we have an explicit constraint on the inlier abstention rate, and will not work for the formulation we are interested in (9). This is because in our formulation, we impose a budget on the overall abstention rate (as this is a more intuitive quantity that a practitioner may want to constraint), and do not explicitly control the abstention rate on $\mathbb{P}_{\text{in}}$.

In comparison to Katz-Samuels et al. (2022), the plug-in based approach we prescribe is more general, and can be applied to optimize any objective that involves as a weighted combination of the mis-classification error and the abstention rates on the inlier and OOD samples. This includes both the budget-constrained problem we consider in (9), and the constrained problem of Katz-Samuels et al. in (10).

D Additional Experiments

D.1 Hyper-parameter choices

We provide details of the learning rate (LR) schedule and other hyper-parameters used in our experiments.

Dataset	Model	LR	Schedule	Epochs	Batch size
CIFAR 10	CIFAR ResNet 56	1.0	anneal	256	1024
CIFAR 100	CIFAR ResNet 56	1.0	anneal	256	1024
ImageNet	EfficientNet B0	0.1	cosine	90	1024

We use SGD with momentum as the optimization algorithm for all models. For annealing schedule, the specified learning rate (LR) is the initial rate, which is then decayed by a factor of ten after each epoch in a specified list. For CIFAR, these epochs are 15, 96, 192 and 224. For ImageNet, these epochs are 5, 30, 60, and 80.

D.2 Baseline details

We provide further details about the baselines we compare with. The following baselines are trained on only the inlier data.

- *MSP or Chow's rule*: Train a scorer $f : \mathcal{X} \to \mathbb{R}^L$ using CE loss, and threshold the MSP to decide to abstain (Chow, 1970; Hendrycks & Gimpel, 2017).
- *MaxLogit*: Same as above, but instead threshold the maximum logit $\max_{y \in [L]} f_y(x)$ (Hendrickx et al., 2021).
- *Energy score*: Same as above, but threshold the energy function $-\log \sum_y \exp(f_y(x))$ (Liu et al., 2020b).
- *ODIN*: Train a scorer $f : \mathcal{X} \to \mathbb{R}^L$ using CE loss, and uses a combination of input noise and temperature-scaled MSP to decide when to abstain Hendrickx et al. (2021).
- *SIRC*: Train a scorer $f : \mathcal{X} \to \mathbb{R}^L$ using CE loss, and compute a post-hoc deferral rule that combines the MSP score with the L_1-norm of the embedding layer from the scorer f (Xia & Bouganis, 2022).
- *CSS*: Optimize the cost-sensitive softmax L2R loss of Mozannar & Sontag (2020) using only the inlier dataset to learn a scorer $f \colon \mathcal{X} \to \mathbb{R}^{L+1}$, augmented with a rejection score $f_\perp(x)$, and abstain iff $f_\perp(x) > \max_{y' \in [L]} f_{y'}(x) + t$, for threshold t.

The following baselines additional use the unlabeled data containing a mix of inlier and OOD samples.

- *Coupled CE (CCE)*: Train a scorer $f \colon \mathcal{X} \to \mathbb{R}^{L+1}$, augmented with a rejection score $f_\perp(x)$ by optimizing the CCE loss of Thulasidasan et al. (2021), and abstain iff $f_\perp(x) > \max_{y' \in [L]} f_{y'}(x) + t$, for threshold t.
- *De-coupled CE (DCE)*: Same as above but uses the DCE loss of Bitterwolf et al. (2022) for training.
- *Outlier Exposure (OE):* Train a scorer using the OE loss of Hendrycks et al. (2019) and threshold the MSP.

D.3 Illustration of MSP failure

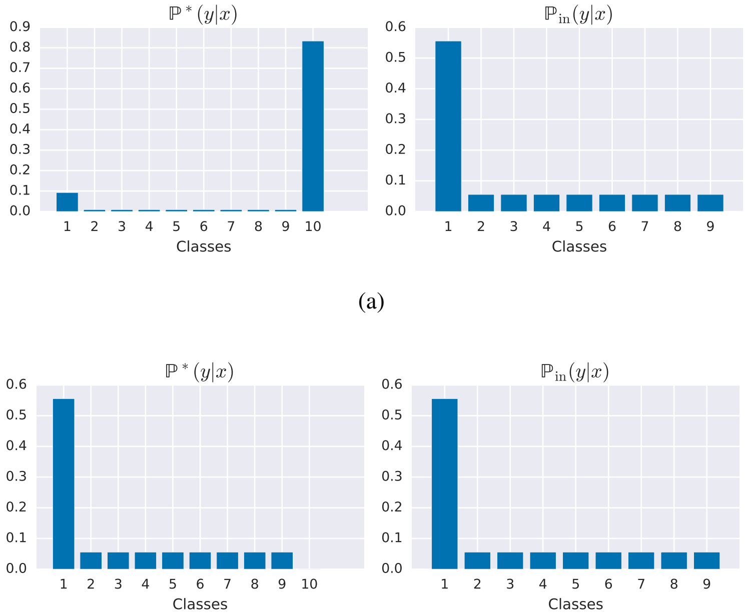
Figure 4 shows a graphical illustration of the example discussed in §3.3, wherein the MSP baseline can fail for open-set classification.

D.4 Additional experimental plots

We present additional experimental plots in Figure 5 and 6.

D.5 AUC and FPR95 metrics

Table 1 reports the ROC-AUC and FPR@95TPR metrics for different OOD scorers.



(a)

(b)

Figure 4: Examples of two open-set classification settings (a) and (b) with $L = 10$ classes, where the inlier class distributions $\mathbb{P}_{\text{in}}(y \mid x) = \frac{\mathbb{P}^*(y|x)}{\mathbb{P}^*(y \neq 10|x)}$ over the first 9 classes are identical, but the unknown class density $\mathbb{P}^*(10|x)$ is significantly different. Consequently, the MSP baseline, which relies only on the inlier class probabilities, will output the same rejection decision for both settings, whereas the Bayes-optimal classifier, which rejects by thresholding $\mathbb{P}^*(10|x)$, may output different decisions for the two settings.

D.6 Maximum logit for open-set classification

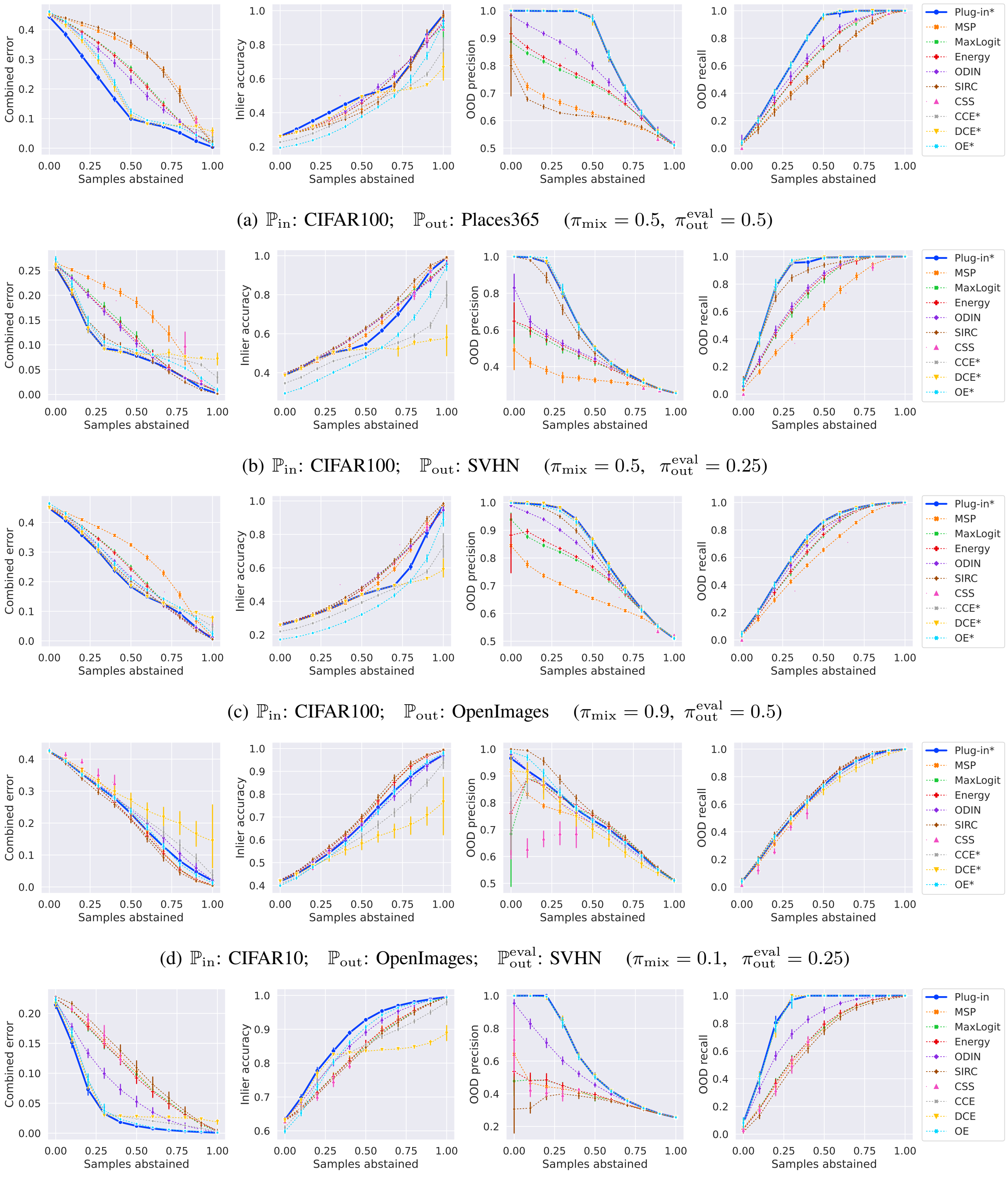
For the same setting as Figure 1, we show in Figure 7 the maximum logit computed over the inlier distribution. As with the maximum probability, the outlier samples tend to get a higher score than the inlier samples.

D.7 Impact of varying abstention costs α, β

Our joint objective that allows for abstentions on both "hard" and "outlier" samples is controlled by parameters α, β. These reflect the costs on not correctly abstaining on samples from either class of anomalous sample. Figure 8 and 9 show the impact of varying these parameters while the other is fixed, for the synthetic open-set classification example of Figure 1(b). The results are intuitive: varying α tends to favour abstaining on samples that are at the class boundaries, while varying β tends to favour abstaining on samples from the outlier class. Figure 10 confirms that when *both* α, β are varied, we achieve abstentions on both samples at the class boundaries, and samples from the outlier class.

D.8 Impact of β on OOD detection performance

For the same setting as Figure 1, we consider the OOD detection performance of the score $s(x) = \max_{y \in [L]} \mathbb{P}_{\text{in}}(y \mid x) - \beta \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{out}}(x)}$ as β is varied. Note that thresholding of this score determines the Bayes-optimal classifier (Lemma 4.1). Rather than pick a fixed threshold, we use this score to compute the AUC-ROC for detecting whether a sample is from the outlier class, or not. As expected, as β increases — i.e., there is greater penalty on not rejecting an OOD sample — the AUC-ROC improves.



(a) $\mathbb{P}_{\rm in}$: CIFAR100; $\mathbb{P}_{\rm out}$: Places365 $(\pi_{\rm mix} = 0.5,\ \pi^{\rm eval}_{\rm out} = 0.5)$

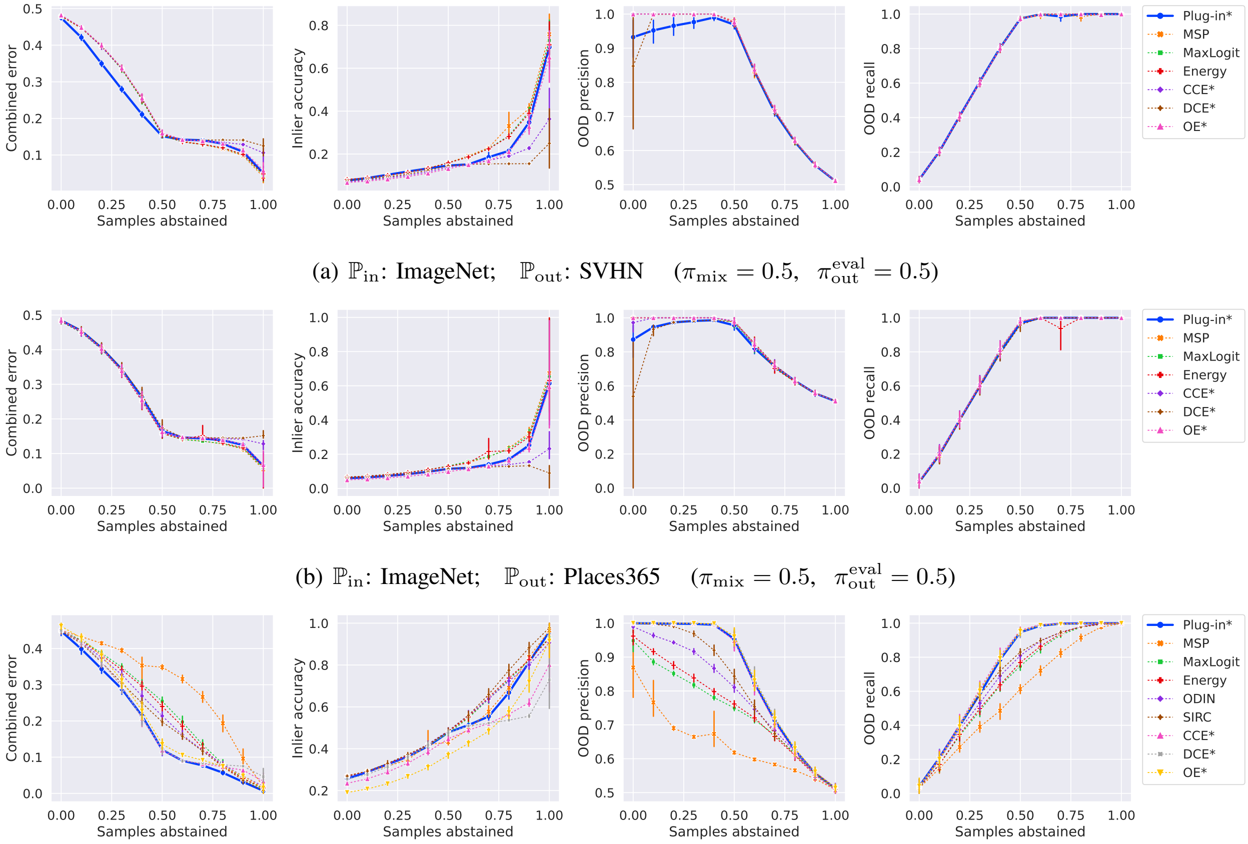
(b) $\mathbb{P}_{\rm in}$: CIFAR100; $\mathbb{P}_{\rm out}$: SVHN $(\pi_{\rm mix} = 0.5,\ \pi^{\rm eval}_{\rm out} = 0.25)$

(c) $\mathbb{P}_{\rm in}$: CIFAR100; $\mathbb{P}_{\rm out}$: OpenImages $(\pi_{\rm mix} = 0.9,\ \pi^{\rm eval}_{\rm out} = 0.5)$

(d) $\mathbb{P}_{\rm in}$: CIFAR10; $\mathbb{P}_{\rm out}$: OpenImages; $\mathbb{P}^{\rm eval}_{\rm out}$: SVHN $(\pi_{\rm mix} = 0.1,\ \pi^{\rm eval}_{\rm out} = 0.25)$

(e) $\mathbb{P}_{\rm in}$: CIFAR10; $\mathbb{P}_{\rm out}$: OpenImages; $\mathbb{P}^{\rm eval}_{\rm out}$: CelebA $(\pi_{\rm mix} = 0.1,\ \pi^{\rm eval}_{\rm out} = 0.25)$

Figure 5: Plots of classification and OOD detection metrics as a function of the fraction of abstained samples (averaged over 5 trials).For the conditional risk, *lower values are better*. For all other metrics, *higher values are better*. A '*' against a method indicates it trains on both inlier and unlabeled samples. Higher value of $\pi_{\rm mix}$ make the unlabeled sample less reliable. We set $c_{\rm ood} = 5$.



(a) $\mathbb{P}_{\text{in}}$: ImageNet; $\mathbb{P}_{\text{out}}$: SVHN $(\pi_{\text{mix}} = 0.5,\ \pi_{\text{out}}^{\text{eval}} = 0.5)$

(b) $\mathbb{P}_{\text{in}}$: ImageNet; $\mathbb{P}_{\text{out}}$: Places365 $(\pi_{\text{mix}} = 0.5,\ \pi_{\text{out}}^{\text{eval}} = 0.5)$

(c) $\mathbb{P}_{\text{in}}$: CIFAR100; $\mathbb{P}_{\text{out}}$: OpenImages; $\mathbb{P}_{\text{out}}^{\text{eval}}$: Places $(\pi_{\text{mix}} = 0.5,\ \pi_{\text{out}}^{\text{eval}} = 0.5)$

Figure 6: Additional plots of classification and OOD detection metrics as a function of the fraction of abstained samples (averaged over 5 trials).For the conditional risk, *lower values are better*. For all other metrics, *higher values are better*. A ‘*’ against a method indicates it trains on both inlier and unlabeled samples. Higher value of π_{mix} make the unlabeled sample less reliable. We set $c_{\text{ood}} = 5$.

	$\mathbb{P}_{\text{out}}$: SVHN				$\mathbb{P}_{\text{out}}$: Places365				$\mathbb{P}_{\text{out}}$: OpenImages			
	$\pi_{\text{mix}} = 0.5$		$\pi_{\text{mix}} = 0.9$		$\pi_{\text{mix}} = 0.5$		$\pi_{\text{mix}} = 0.9$		$\pi_{\text{mix}} = 0.5$		$\pi_{\text{mix}} = 0.9$	
Method	AUC	FPR95	AUC	FPR95	AUC	FPR95	AUC	FPR95	AUC	FPR95	AUC	FPR95
Plug-in*	0.988	0.046	0.948	0.245	0.998	0.002	0.988	0.029	0.992	0.031	0.938	0.360
MSP	0.654	0.738	0.652	0.741	0.686	0.737	0.683	0.742	0.724	0.689	0.724	0.692
MaxLogit	0.801	0.506	0.800	0.508	0.811	0.534	0.811	0.533	0.836	0.507	0.837	0.505
Energy	0.814	0.493	0.816	0.490	0.821	0.526	0.821	0.523	0.846	0.498	0.847	0.493
ODIN	0.826	0.483	0.827	0.484	0.874	0.478	0.873	0.481	0.887	0.462	0.887	0.463
SIRC	0.936	0.366	0.936	0.370	0.674	0.722	0.673	0.719	0.907	0.482	0.909	0.478
CCE*	0.991	0.032	0.958	0.210	0.999	0.001	0.993	0.029	0.993	0.022	0.940	0.343
DCE*	0.988	0.047	0.949	0.245	0.999	0.001	0.993	0.027	0.992	0.032	0.938	0.366
OE*	0.994	0.008	0.961	0.243	0.999	0.006	0.951	0.278	0.991	0.025	0.929	0.406

Table 1: AUC and FPR@95TPR metrics for different OOD detectors with CIFAR-100 as the inlier dataset and different OOD datasets. The same OOD data set is used during both training and evaluation. The plug-in scorer is generally seen to be competitive on both metrics. The methods marked with ‘*’ are trained only on the inlier datasets.

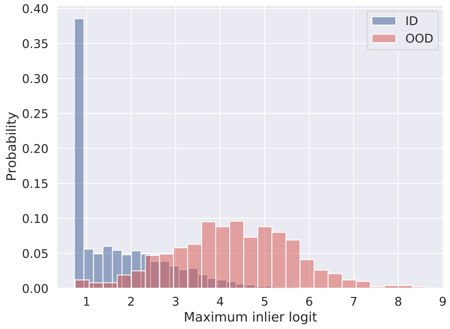


Figure 7: For the same setting as Figure 1, we show the maximum logit computed over the inlier distribution. As with the maximum probability, the outlier samples tend to get a higher score than the inlier samples.

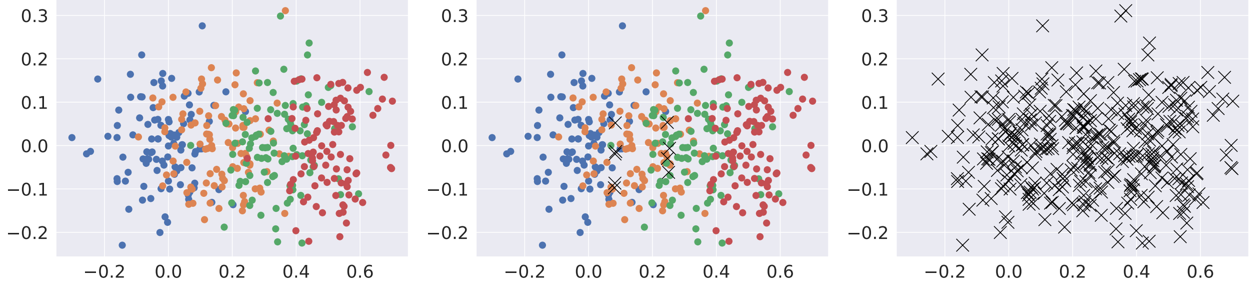


Figure 8: Impact of varying α for a fixed $\beta = 0.0$. The left plot shows the standard dataset, with $\alpha = 1.0$. For intermediate $\alpha = 0.5$ (middle), we abstain (denoting by $\times$) only on the samples at the class boundaries. For $\alpha = 0.0$ (right), we abstain on all samples.

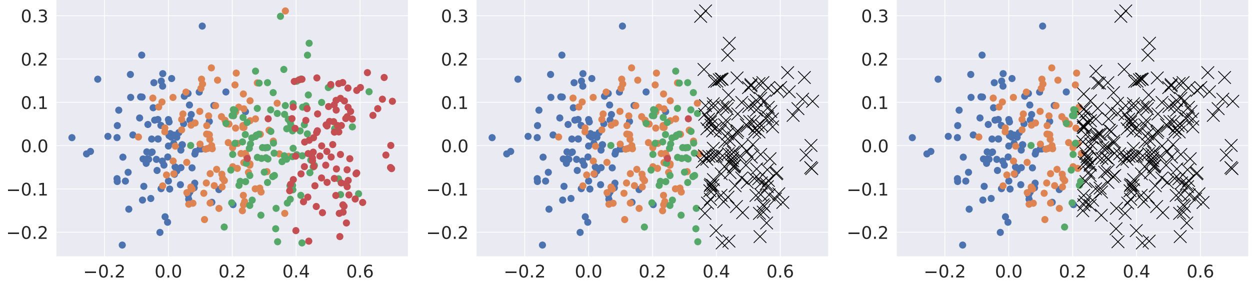


Figure 9: Impact of varying β for a fixed $\alpha = 1.0$. The left plot shows the standard dataset, with $\beta = 0.0$. For intermediate $\beta = 1.0$ (middle), we abstain (denoting by $\times$) only on the samples from the outlier class. For larger $\beta = 10.0$ (right), we start abstaining on inlier samples as well.

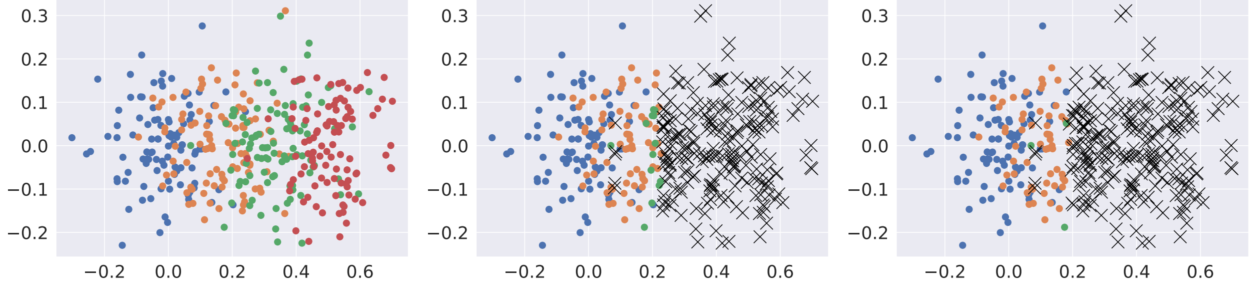


Figure 10: Impact of varying both α and β. The left plot shows the standard dataset, with $\alpha = 1.0, \beta = 0.0$. Setting $\alpha = 0.5, \beta = 1.0$ (middle) and $\alpha = 0.5, \beta = 10.0$ (right) is shown to favour abstaining (denoting by $\times$) on *both* the samples at class boundaries, and the outlier samples.

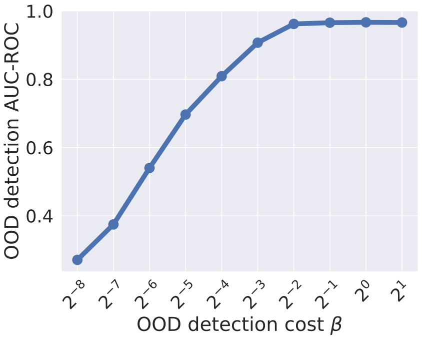


Figure 11: For the same setting as Figure 1, we consider the OOD detection performance of the score $s(x) = \max_{y \in [L]} \mathbb{P}_{\text{in}}(y \mid x) - \beta \cdot \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{out}}(x)}$ as β is varied. Specifically, we use this score to compute the AUC-ROC for detecting whether a sample is from the outlier class, or not. As expected, as β increases, the AUC-ROC improves.