Birds of an Odd Feather: Guaranteed Out-of-Distribution (OOD) Novel Category Detection

Yoav Wald[1] **Suchi Saria**[1,2]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD
[2]Bayesian Health, New York, NY

Abstract

In this work, we solve the problem of novel category detection under distribution shift. This problem is critical to ensuring the safety and efficacy of machine learning models, particularly in domains such as healthcare where timely detection of novel subgroups of patients is crucial.
To address this problem, we propose a method based on constrained learning. Our approach is guaranteed to detect a novel category under a relatively weak assumption, namely that rare events in past data have bounded frequency under the shifted distribution. Prior works on the problem do not provide such guarantees, as they either attend to very specific types of distribution shift or make stringent assumptions that limit their guarantees.
We demonstrate favorable performance of our method on challenging novel category detection problems over real world datasets.

1 INTRODUCTION

Distribution shifts occur in most real-world scenarios where machine learning (ML) is deployed and these shifts can result from both natural and adversarial changes including differences in data recording protocols, shifts in the underlying population being monitored, or the way the ML tool is being used [Koh et al., 2021, Finlayson et al., 2021, Quinonero-Candela et al., 2008, Saria and Subbaswamy, 2019]. While some shifts do not pose an immediate safety concern, others warrant examination and proper treatment. In this paper, we are concerned with potential risks that arise from the emergence of a novel category or subgroup and will study guarantees around the automated detection of such subgroups under practical, real-world assumptions.

As motivation, consider dataset shift scenarios in the healthcare domain [Finlayson et al., 2021, Table 1]. At the start of the COVID-19 pandemic, a Michigan hospital described how a predictive tool for catching patients at-risk for a life-threatening complication called sepsis started to over-alert and incorrectly flag patients as the underlying population shifted [Finlayson et al., 2021]. Ultimately they had to turn off the tool because of the harms it posed to patients. In this scenario, the tool was scanning and providing predictions on new patient groups (e.g., patients with likely COVID-19) which led to the safety issue. [1]

In this paper we tackle the problem of *Out-of-Distribution (OOD) Novel Category Detection* (also called novel class, or subgroup). We aim to identify novel instances within a dataset that contains both known and novel categories. What sets our approach apart is that we not only account for the introduction of a new category but also allow *other distribution shifts between the new and previously observed data*. This aspect is of utmost importance when it comes to monitoring risks in real-world applications, as distributions tend to change over time and new data continually emerges.

Returning to the healthcare scenario described above, besides the introduction of new patient subgroups related to COVID-19, the baseline population itself shifted because the types of patients coming into the hospital evolved over the course of the pandemic. Early on, only those with urgent needs visited. Over time, those with longer term needs and planned surgeries began to use the hospital. Requiring the baseline distribution to remain constant over time is a highly restrictive assumption and in many real-world settings, we need the ability to detect novel categories without enforcing this assumption [Koh et al., 2021, Finlayson et al., 2021] (that is, the ability to work OOD). To this end, we develop a method with guarantees on classification error of the novel category, that hold under a wide range of distribution shifts. Formal guarantees are particularly important in

[1]A trivial solution might be to filter out new likely subgroups including COVID-19 patients from the list of patients that the tool was allowed to make predictions on. However, because patient diagnoses were not available upon presentation to the hospital, this was inadequate.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

safety-critical applications, where these can help substantiate trust and confidence by users and regulators. Many approaches have been devised for detection of novel categories, mainly in the Open World learning literature (e.g. [Panareda Busto and Gall, 2017, Han et al., 2019, Xu et al., 2019]). These methods are often applied in complex scenarios, yet give little to no theoretical guarantees. On the other hand, methods that are rigorously justified apply to scenarios without distribution shifts, and are hence substantially simpler [Blanchard et al., 2010, Garg et al., 2022, Liu et al., 2018], or less suitable for novel category detection than the setting we pursue here [He et al., 2018]. We make the following contributions:

- We propose a new learning algorithm for the problem of OOD novel category detection (i.e. novel category detection when the baseline distribution shifts). The method builds on approaches for constrained learning [Eban et al., 2017, Agarwal et al., 2018, Chamon et al., 2022, Cotter et al., 2019a, Donini et al., 2018] and seeks to maximize the number of points correctly detected as novel, while keeping false detections below a certain rate.
- We provide guarantees on the error of the learned model that hold under a certain assumption, namely, that rare events in past data have bounded frequency under the new distribution. Prior works either provide much weaker guarantees or rely on stringent assumptions. Works that study the label-shift scenario assume that the only change is in frequency of known and labelled subgroups [Garg et al., 2022, Shanmugam and Pierson, 2021]. In our healthcare scenario, such methods require defining all possible patient subgroups that can shift, labelling the membership of patients in them, and accurately estimating the change in their frequency. Methods based on this strong assumption can also become impractical considering the tedious labelling and amount of data required. Other approaches require access to perfectly accurate density ratios between the distribution of past and current data (or propensity scores, that cannot necessarily be estimated from data) [Bekker et al., 2019, Gerych et al., 2022, Jain et al., 2020], which limits both theoretical guarantees and their performance in many settings, for instance those involving high-dimensional data where density ratio estimation is challenging [Sugiyama et al., 2012, Chapter 8].
- Finally, we show favorable performance of the algorithm on challenging novel category detection tasks that we simulate over real world datasets.

The problem we study is related to OOD detection and Open-World learning [Ruff et al., 2021]. We provide a short review of these problems below, and after defining our specific setting formally in Section 3, we will review assumptions made in more closely related work in Section 3.2. Our own assumptions are described in Section 4 along with their theoretical guarantees. Then in Section 5 and Section 6 we present our method CoNoC and its experimental evaluation.

2 RELATED WORK

Our results apply to a generalized form of Novel Category Detection. Let us discuss this setting and other problems related to ours.

Novel Category Discovery, Open World Learning and PU Learning. In Open World Learning [Parmar et al., 2023] (and specifically Open Set Domain Adaptation [Panareda Busto and Gall, 2017]) and Novel Category Detection [Liu et al., 2018] the learner is given labelled data from known categories, and unlabelled data from both known and novel categories. These problems bear similarity to learning from Positive and Unlabelled data (PU-learning) [Bekker and Davis, 2020], where we are given data that are labelled as positive, which in our terminology means "from a previously observed category", or unlabelled (i.e. from both observed and novel categories). The task is then to learn a model that classifies categories, both known and novel. The main difference between these settings and our work is that prior work provides guarantees for cases where the base distribution does not shift [2]. We thus refer to our generalized setting as *OOD* Novel Category Detection.

OOD Detection. Identifying anomalous instances that are not members of previously seen classes, or out-of-support for the distribution of observed data, is a well-studied problem in ML. The main difference from our setting is that in OOD detection the learner does not observe any examples from the target distribution (e.g. COVID-19 patients and shifted baseline population in our healthcare example) at training time, and the baseline distribution is assumed to be fixed. Classic approaches for this problem include One-Class Support Vector Machines [Schölkopf et al., 2001] and Kernel Density Estimation [Parzen, 1962], they have modern counterparts suited for flexible models such as neural networks [Chalapathy et al., 2018, Nachman and Shih, 2020]. We refer the reader to [Ruff et al., 2021] for a comprehensive survey. Since in OOD detection no data from the target distribution is observed, theoretical guarantees such as PAC-learning generalization bounds are limited. Recent work shows that it is often impossible to provide such guarantees on OOD detection, unless we make restrictive assumptions relating our hypothesis class (i.e. architecture), and the target distribution [Fang et al., 2022]. In our setting, we instead allow the learner to access target data. This is a reasonable assumption in any setting where our novelty detector can adapt to newly observed data, and it lets us alleviate assumptions on the hypothesis class and to provide guarantees under shifts in the baseline distribution.

Constrained Learning and Fairness. Our method draws on developments in learning with data-dependent constraints, and specifically rate-constraints (e.g. constraining the amount of examples that are labelled positive) (e.g [Eban

[2] we elaborate on existing results in Section 3.2

et al., 2017, Donini et al., 2018]). Many of these methods were motivated by applications in fairness [Agarwal et al., 2018, Woodworth et al., 2017, Donini et al., 2018], yet general frameworks for learning with data dependent constraints are useful for many other tasks and there is growing interest in them [Donti et al., 2021, Chamon et al., 2022].

We now turn to provide a formal definition of the OOD novel category detection problem, and an intuition to our proposed solution.

3 PROBLEM SETTING

In *OOD Novel Category Detection* we seek to detect a novel category (also called novel class, or subgroup) within a dataset that contains both known and novel categories. Crucially, *the distribution of known categories can shift*.

Consider a dataset $S_{\mathcal{S}} = \{\mathbf{x}_i\}_{i=1}^{N_{\mathcal{S}}}$ collected under a certain protocol, we formally treat this as an i.i.d sample from some *source distribution* $P_{\mathcal{S}}$. For instance, in our healthcare example, data collected in the months preceding the pandemic. At a later time or under different conditions, we collect more data $S_{\mathcal{T}} = \{\mathbf{x}_i\}_{i=1}^{N_{\mathcal{T}}}$ which contains a novel category that we would like to detect, of proportion $\alpha \in [0, 1]$ in the population. The category is unlabelled, i.e. we are not given any examples that are labelled as novelties. We treat this category as a sample from a *novelty distribution* $P_{\mathcal{T},1}$, and call its proportion in the new data, α, the *mixture proportion*. The rest of the data in $S_{\mathcal{T}}$ is sampled from a nominal distribution $P_{\mathcal{T},0}$, which we think of as a shifted version of $P_{\mathcal{S}}$. In summary, $S_{\mathcal{T}}$ is an i.i.d sampled dataset from $P_{\mathcal{T}} = (1 - \alpha)P_{\mathcal{T},0} + \alpha P_{\mathcal{T},1}$. Our task is as follows.

Definition 3.1 (OOD Novel Category Detection)**.** The tuple $\langle P_{\mathcal{S}}, P_{\mathcal{T},0}, P_{\mathcal{T},1}, \alpha, n_{\mathcal{S}}, n_{\mathcal{T}}\rangle$ defines an OOD novel category detection problem where $S_{\mathcal{S}}, S_{\mathcal{T}}$ are datasets of $n_{\mathcal{S}}, n_{\mathcal{T}}$ examples sampled i.i.d from $P_{\mathcal{S}}, P_{\mathcal{T}}$ respectively, where $P_{\mathcal{T}} = (1 - \alpha)P_{\mathcal{T},0} + \alpha P_{\mathcal{T},1}$. For a hypothesis class of binary classifiers $\mathcal{H}$, Let $h^* \in \mathcal{H}$ be the minimizer of the expected $0 - 1$ risk over the target distribution:

$$R^{l_{01}}_{\mathcal{T}}(h) = (1 - \alpha) \cdot \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},0}}[h(\mathbf{x})] + \alpha \cdot \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},1}}[1 - h(\mathbf{x})]. \quad (1)$$

An algorithm $\mathcal{A} : \mathcal{X}^{n_{\mathcal{S}}} \times \mathcal{X}^{n_{\mathcal{T}}} \to \mathcal{H}$ is a learner for the novel class detection problem if for every $\varepsilon, \delta > 0$ it satisfies $R^{l_{01}}_{\mathcal{T}}(\mathcal{A}(S_{\mathcal{S}}, S_{\mathcal{T}})) \leq R^{l_{01}}_{\mathcal{T}}(h^*) + \varepsilon$ with probability at least $1 - \delta$ whenever $\min\{n_{\mathcal{S}}, n_{\mathcal{T}}\} \geq m_{\mathcal{H}}(\varepsilon^{-1}, \delta^{-1})$ for a function $m_{\mathcal{H}} : [0, 1]^2 \to \mathbb{N}$.

Further, in this problem $P_{\mathcal{S}}$ and $P_{\mathcal{T},0}$ may contain different mixture proportions of the same latent subpopulations (Duchi et al. [2022], Sagawa et al. [2020a]). This allows us to tackle challenging scenarios like the healthcare scenario described earlier where the types of patients visiting the hospital changes over time including the introduction of new COVID-19 related patient subgroups. Later we will specify the precise distribution shifts that we treat. Denoting the distributions corresponding to subpopulations by $\{G_i\}_{i=1}^K$ for some $K \in \mathbb{N}$, and the probability simplex over $[K]$ by Δ^{K-1},

$$P_{\mathcal{S}} = \sum_{i=1}^{K} \gamma_i G_i,\ P_{\mathcal{T},0} = \sum_{i=1}^{K} \hat{\gamma}_i G_i,\ \boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}} \in \Delta^{K-1}. \quad (2)$$

3.1 MOTIVATING EXAMPLE

To motivate our solution consider a simple case of latent subpopulation shift, as in Equation (2), plotted in Figure 1 (Left). There are two latent subpopulations that make up known categories in $S_{\mathcal{S}}$, and the novel category is marked with a dashed circle. Let us examine how a method that does not handle distribution shift works in this example.

Detection without distribution shift. Formally, our problem can be cast in the framework of learning from Positive and Unlabelled data (PU-learning). Most work under this framework relies on the Selected-Completely-At-Random assumption (SCAR) [Elkan and Noto, 2008], that is $P_{\mathcal{S}} = P_{\mathcal{T},0}$. Besides being very restrictive, it turns out that many of the approaches based on SCAR can fail when it breaks. Common algorithms for PU-learning are based on a classifier trained to distinguish the domains $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ (Domain Discriminator). Intuitively, this approach is effective since examples from the novel class turn out to be "farthest" from the decision threshold. Then adjustment of the decision threshold according to a successful Mixture Proportion Estimation (MPE), i.e. an estimate of α, should enable us to classify novelties [Elkan and Noto, 2008, du Plessis et al., 2014, Garg et al., 2021]. The example in Figure 1 shows this approach can run into problems when there is distribution shift between $P_{\mathcal{S}}$ and $P_{\mathcal{T},0}$. A domain discriminator trained with logistic regression is biased towards separating the categories that are observed in the source data $S_{\mathcal{S}}$. This is due to their varying mixture coefficients between $P_{\mathcal{S}}$ and $P_{\mathcal{T},0}$, and it results in the examples from the novel category not being farthest from the decision boundary.

Why CoNoC solves this. However, a linear classifier *can* separate the novel category, and the method we propose in this work is able to recover it as can be seen for the classifier labelled CoNoC in Figure 1 (Left). In a nutshell, CoNoC seeks to maximize the number of points in $S_{\mathcal{T}}$ that are detected as novelties, while keeping the number of points in $S_{\mathcal{S}}$ that are wrongly detected as novelties below a certain threshold. Figure 1 (Right) illustrates why this approach is expected to work by plotting the Receiver-Operator Curve (ROC) for two models, the domain discriminator and the one trained with CoNoC. The domain discriminator does better (in terms of aggregate classification metrics such as average loss, or F1-Score) in classifying $S_{\mathcal{S}}$ vs. $S_{\mathcal{T}}$, and note that larger distribution shifts further improve its discriminative

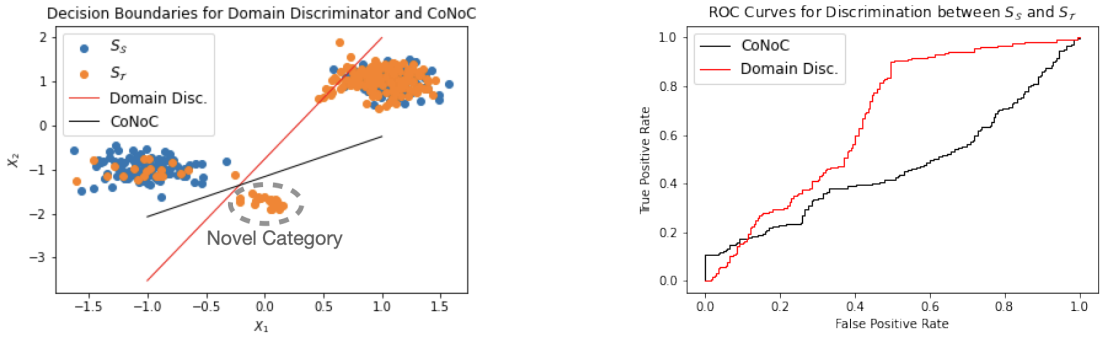

Figure 1: **(Left)** Toy example where a classifier learned with CoNoC is favorable over a domain discriminator in detecting a novel category. A domain discriminator is trained to reduce overall loss and hence it is biased towards labelling the upper right cluster with label 1 (i.e. as a novelty). **(Right)** ROC-Curves for a domain discriminator and a model trained with CoNoC for the data in the left panel, where $S_{\mathcal{T}}$ are labeled positive and $S_{\mathcal{S}}$ negative. An optimal classifier for the novel category is suboptimal w.r.t aggregate performance metrics (e.g. AU-ROC), but has higher TPR when constrained to a small FPR, illustrating why our constrained learning approach can recover novel categories successfully.

ability. This is clear from the figure, as its ROC curve dominates the other one for most values of the False Positive Rate (FPR). However, an *optimal novel category detector* (in our case this coincides with CoNoC) has better True Positive Rate (TPR) for small FPR values, as observed in Figure 1 (Right). Intuitively, this model sees a sharp increase in the TPR for low FPR values due to correct classifications of the novel category. Hence our suggested approach should prefer the novel category detector over the domain discriminator. But when is this approach guaranteed to detect the novel category? What are the required assumptions, sample size, and how should we set the bound on the FPR? In the following sections we provide answers to these questions and an implementation of the proposed principle.

Concluding this example, we note that other solutions can be devised for the specific dataset we considered. For instance clustering, or training a domain discriminator from a larger hypothesis class. Yet these solutions do not extend gracefully to more general settings. For instance, it is unlikely that in every dataset of interest, clustering high dimensional data retrieves the accurate subgroups that undergo shift. Expressive hypothesis classes are also not a reliable solution, as they introduce biases of their own. For example, it is well-known that large overparameterized models tend to perform poorly on small subgroups [Hashimoto et al., 2018, Sagawa et al., 2020b, Menon et al., 2021, Wald et al., 2022]. In our setting these may correspond to the novel category which comprises a small part of $S_{\mathcal{T}}$, thus detection of the novel category may be poor.

3.2 NECESSARY AND SUFFICIENT ASSUMPTIONS FOR LEARNING

Moving towards a principled approach for OOD Novel Category Detection, our first challenge is that we do not have access to samples from $P_{\mathcal{T},0}$ and $P_{\mathcal{T},1}$, hence $R^{l_{01}}_{\mathcal{T}}(h)$ cannot be estimated from data. It is easy to show that without any distributional assumptions, guarantees on the performance of a learning algorithm cannot be derived. We state this below and give the proof in Appendix A.2.

Proposition 3.1. *Let $\mathcal{A}$ be a learning algorithm for the task of OOD novel category detection. There are distributions $P_{\mathcal{S}}, P_{\mathcal{T},0}, P_{\mathcal{T},1}$ such that $\exists h^* \in \mathcal{H}$ for which $R^{l_{01}}_{\mathcal{T}}(h^*) = 0$, while $\mathbb{E}_{S_{\mathcal{S}},S_{\mathcal{T}}}\left[R^{l_{01}}_{\mathcal{T}}(\mathcal{A}(S_{\mathcal{S}}, S_{\mathcal{T}}))\right] \geq 0.5$.*

Since it is impossible to guarantee better-than-chance performance for a learning algorithm, several distributional assumptions have been formulated in the literature under which learning is possible.

No distribution shift scenario When $P_{\mathcal{S}} = P_{\mathcal{T},0}$, assumptions like irreducibility, which says that $P_{\mathcal{T},1}$ cannot be written as a mixture of $P_{\mathcal{S}}$ and another distribution, enable identification of α and learning [Blanchard et al., 2010]. Stricter assumptions can help devise more efficient algorithms, e.g. [Scott, 2015, Garg et al., 2021], but they are insufficient once we consider distribution shifts.

Known subpopulations and invariance of order More recent works [Garg et al., 2022, Shanmugam and Pierson, 2021, Jain et al., 2020], consider the subpopulation shift scenario of Equation (2) where *the subgroups are known*, or the learner is given a sample from each G_i. Once subgroups are known, some variations on the assumptions for the no-distribution-shift scenario enable learning, and methods such as reweighting and resampling can counteract the effects of the shift to obtain learning algorithms. In contrast, in this work we ask what can be done in cases where the subgroups are unknown to the learner. Another type of assumption that has been explored is "invariance of order" [Kato et al., 2018, He et al., 2018], which can roughly be

summarized as $P_{\mathcal{S}}(\mathbf{x}) > P_{\mathcal{T}}(\mathbf{x}) \Rightarrow P_{\mathcal{T},0}(\mathbf{x}) > P_{\mathcal{T},1}(\mathbf{x})$. Meaning examples that are more likely in the source distribution are less likely to be novelties. This type of assumption is unsuitable for our goals, as it entails an asymmetry between $P_{\mathcal{S}}$ and $P_{\mathcal{T},0}$. For instance, it is reasonable to expect that detection of a novel subgroup of patients is possible regardless of whether it has been introduced in hospital A (corresponds to $P_{\mathcal{S}}$), or hospital B (resp. $P_{\mathcal{T},0}$). This type of symmetry is denied by the assumption on orderings.

Separability The closest assumption to ours is separability, which says that the support of $P_{\mathcal{S}}$ must be disjoint from that of $P_{\mathcal{T},1}$ and fully overlap with that of $P_{\mathcal{T},0}$. Showing that the mixture proportion can be recovered, given perfect knowledge of $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ is rather straightforward (see Appendix A.2), and learning with infinitely large samples can also be done. Bekker et al. [2019] propose using the propensity score, $P_{\mathcal{S}}(\mathbf{x})/\left(P_{\mathcal{S}}(\mathbf{x}) + P_{\mathcal{T},0}(\mathbf{x})\right)$ to augment and reweigh $S_{\mathcal{S}}$, forming a debiased risk minimization problem. [3] Gerych et al. [2022] show that under separability, the propensity score can be identified from data. They do not provide finite sample guarantees, and the method requires solving a challenging density ratio approximation problem. Our contributions include a relaxed version of the separability assumption, which leads to finite sample generalization bounds and a learning rule that is markedly different from approaches based on estimating importance scores.

4 AN ASSUMPTION ON THE RATE OF RARE EVENTS AND AN ERROR BOUND

We now turn to develop our algorithm and derive its statistical guarantees. Our first step is to define a divergence that measures the extent to which rare events in a distribution P are likely under distribution Q. Given a threshold $\beta > 0$, used to describe an event being "rare", we consider the following divergence. [4]

Definition 4.1. For distributions P, Q over domain $\mathcal{X}$, a hypothesis class $\mathcal{H}$ and $\beta > 0$, we define for each $g \in \mathcal{H}$ the set it characterizes $I(g) = \{\mathbf{x} | g(\mathbf{x}) = 1\}$ and denote,

$$d_{\mathcal{H},\beta}\left(P \| Q\right) = \sup_{g \in \mathcal{H}: P[I(g)] \leq \beta} 2\Big|P\left[I(g)\right] - Q\left[I(g)\right]\Big|. \tag{3}$$

The divergence is similar to the well-known $\mathcal{H}$-divergence from the domain adaptation literature [Ben-David et al., 2010, Kifer et al., 2004], but has an additional rate constraint where $g(\mathbf{x})$ may only make a fraction β of positive predictions under P. We use this divergence to state our distributional assumption in what follows, in Appendix A.3 we also give a short discussion on properties of this divergence.

The Scarcity-of-Unicorns Assumption. Intuitively, if rare events (or "unicorns") under our source distribution $P_{\mathcal{S}}$ are common under $P_{\mathcal{T},0}$, it is impossible to tell whether such events are novelties (i.e. were sampled from $P_{\mathcal{T},1}$) or not. Therefore a bound on the rate of such rare events seems like a reasonable assumption to form the basis of our learning algorithm. In practice, users will have to set a parameter $\beta \geq 0$ that approximates the False Positive Rate (FPR) of an ideal classifier for the new category (which we denote by $\beta(h^*)$). For instance, if we expect to find distinct novel patterns in images, we may set $\beta = 0$. An alternative, more involved scenario, may arise when we observe features such as vitals and lab results of patients, where a novel subpopulation can have some small overlap with previous data. Then regulators and domain experts may define appropriate values for this overlap that warrant further examination. The probability of these false positive events under the shifted distribution $P_{\mathcal{T},0}$ appears as an additional error $\varepsilon_{\text{shift}}$ in our bound (that is presented in Theorem 4.3), and our main assumption is that this error is bounded.

Assumption 4.2. For a known value $\beta \geq 0$ and $\varepsilon_{\text{shift}} \in [0, 1)$ it holds that $d_{\mathcal{H},\beta}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right) \leq \varepsilon_{\text{shift}}$.

The error $\varepsilon_{\text{shift}}$ is incurred due to distribution shift, and it can be reduced for setting a smaller value for β. However, if β is too low we cannot detect instances of the novel category. Our theoretical result provides guidance on how to scale β with the sample size and complexity of $\mathcal{H}$, however in general we must reason about $\beta(h^*)$ using domain knowledge, and this will be reflected by the term $\beta - \beta(h^*)$ in our error bound. We discuss potential data-driven methods to reason about β in the appendix, and close this part by emphasizing an important special case of Assumption 4.2. Namely the separability assumption, common in PU-learning literature (e.g. [Bekker and Davis, 2020, Gerych et al., 2022]).

Proposition 4.1. *Assume separability holds, which postulates that $P_{\mathcal{T},0}(B) > 0 \Rightarrow P_{\mathcal{S}}(B) > 0$ for any measurable subset B w.r.t both distributions.* [5] *Scarcity-of-Unicorns (Assumption 4.2) holds with $\beta, \varepsilon_{\text{shift}}$ set to 0.*

4.1 A CONSTRAINED LEARNING RULE AND ITS GENERALIZATION PROPERTIES

We are now in place to present our learning rule and its statistical guarantee. The following theorem, that we prove

[3] This expression for the score assumes a uniform prior on being sampled from the source vs. target distribution. Generally, the score is the probability that $\mathbf{x}$ was sampled from $P_{\mathcal{S}}$.

[4] It is worth noting that this notion of distance, taken w.r.t measurable subsets $\mathcal{B}$ under the two distributions instead of the hypothesis class $\mathcal{H}$, that is $d_{1,\beta}\left(P \| Q\right) = \sup_{B \in \mathcal{B}: P(B) \leq \beta} 2\Big|P(B) - Q(B)\Big|$, upper bounds $d_{\mathcal{H},\beta}$ and is perhaps more intuitive to reason about.

[5] separability also assumes $\exists h^* \in \mathcal{H}$ such that $R^{l_{01}}_{\mathcal{T}}(h^*) = 0$, but to prove Proposition 4.1 we do not require this.

in Appendix A.1, summarizes our proposal and result. We use the Rademacher complexity [Bartlett and Mendelson, 2002], denoted by $R_{n,P}(\mathcal{H})$ for a distribution P and sample size n, as a measure for the expressiveness of $\mathcal{H}$, yet other standard notions can be used.

Theorem 4.3. *Let $\langle P_{\mathcal{S}}, P_{\mathcal{T},0}, P_{\mathcal{T},1}, \alpha, n_{\mathcal{S}}, n_{\mathcal{T}} \rangle$ define an OOD novel category detection problem (see Definition 3.1) and $h^* \in \mathcal{H}$ the minimizer of $R^{l_{01}}_{\mathcal{H}}$. The following statements hold:*

- *Let $\beta(h) = \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}[h(\mathbf{x})], \alpha(h) = \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T}}}[h(\mathbf{x})]$ be the False Positive Rate (FPR) and recall of a hypothesis $h \in \mathcal{H}$ w.r.t the task of classifying source and target data. The target risk on detecting the novel category can be bounded by*

$$R^{l_{01}}_{\mathcal{T}}(h) \leq [\alpha - \alpha(h)] + (1-\alpha)\left[\beta(h) + d_{\mathcal{H},\beta(h)}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right)\right]. \tag{4}$$

- *Let $\delta > 0$ and assume our problem satisfies Assumption 4.2 with parameters $\beta \geq \beta(h^*) + \frac{R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{S}}}}$ and $\varepsilon_{shift} \geq 0$. Consider $\hat{h} = \mathcal{A}(S_{\mathcal{S}}, S_{\mathcal{T}})$ that solves the empirical learning rule,*

$$\max_{h\in\mathcal{H}} \hat{\alpha}(h) \quad \text{s.t. } \hat{\beta}(h) \leq \beta, \tag{5}$$

where $\hat{\alpha}(h), \hat{\beta}(h)$ are empirical estimates of $\alpha(h), \beta(h)$ from $S_{\mathcal{T}}, S_{\mathcal{S}}$ respectively. We have with probability at least $1 - 4\delta$ that

$$R^{l_{01}}_{\mathcal{T}}(\hat{h}) \leq R^{l_{01}}_{\mathcal{T}}(h^*) + 4\varepsilon_{shift} + 2(\beta - \beta(h^*)) + R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H}) + R_{n_{\mathcal{T}},P_{\mathcal{T}}}(\mathcal{H}) + \sqrt{2\ln(1/\delta)}\left[n_{\mathcal{S}}^{-\frac{1}{2}} + n_{\mathcal{T}}^{-\frac{1}{2}}\right]. \tag{6}$$

Let us break down the statement and draw conclusions. The proposed learning rule in Equation (5) optimizes an upper bound on the error, where the upper bound is drawn in the first part of the theorem (Equation (4)). Unfortunately, the upper bound in Equation (4) cannot be estimated from data, since a sample from $P_{\mathcal{T},0}$ is required to estimate $d_{\mathcal{H},\beta(h)}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right)$. This is where Assumption 4.2 comes in and lets us replace the divergence term, under the condition that $\beta(h)$ is small enough. Finally, we draw a generalization bound on the error of the learned classifier in Equation (6).

Takeaways from Theorem 4.3 Focusing on separable problems (see Proposition 4.1), we may discard the terms that depend on $\varepsilon_{\text{shift}}$ and β from the generalization bound in Equation (6).[6] Then we gather that the algorithm $\mathcal{A}$ which solves Equation (5) is a learning algorithm for the problem, as prescribed in Definition 3.1, so long that $\mathcal{H}$ is learnable under the standard terminology of learning theory [Shalev-Shwartz and Ben-David, 2014]. Note that previously proposed approaches solve more general problems such as clustering [Jain et al., 2020] or density ratio approximation [Gerych et al., 2022], and hence do not provide this type of learnability guarantee. Following the principle that one should not solve a more general problem than required [Vapnik, 2006], we opt for direct optimization of an upper-bound on the error, sidestepping such intermediate steps. We also conclude that upon using our proposed learning rule, the value of β in the constraint of Equation (5) should be set above 0 even when separability holds (i.e. $\beta(h^*) = 0$). It should scale with the complexity of $\mathcal{H}$ and inversely with $n_{\mathcal{S}}$. When separability does not hold, we incur an additional irreducible error proportional to $\varepsilon_{\text{shift}}$.

[6] That is if we set β according to the separability assumption, approaching 0 with growing $n_{\mathcal{S}}$. Otherwise the error $\beta - \beta(h^*)$ does not approach 0, reflecting how well we approximate $\beta(h^*)$.

5 CONOC: A CONSTRAINED LEARNING METHOD FOR OOD NOVEL CATEGORY DETECTION

Most computationally efficient gradient methods and classical ML theory results on statistical efficiency apply to standard risk minimization problems, hence applying them to our problem of solving Equation (5) is not straightforward. Fortunately, recent literature on fairness and constrained learning presents effective tools and beautiful theory to tackle this type of problem [Eban et al., 2017, Chamon et al., 2022, Cotter et al., 2019b,a, Donini et al., 2018, Agarwal et al., 2018, Woodworth et al., 2017]. In this section we adapt these methods and insights to tackle our novelty detection problem and arrive at a constrained learning approach, that we call CoNoC (**Co**nstrained **No**vel **C**ategory detection).

In terms of formal guarantees, constrained learning methods offer attractive bounds on the optimization error for solving equation 5, whereas Theorem 4.3 provides statistical guarantees. By directly plugging in optimization error terms to the bound of Theorem 4.3, error bounds on the complete procedure can be derived. Since combining these results does not require any novel insight or technique, we dedicate the rest of this section to present parts of the method we use in practice which depart from the algorithms discussed in the works above. Our implementation of a constrained learning optimization algorithm uses a simple primal-dual optimization approach with alternating gradient steps where one player controls the model parameters, and the other controls a Lagrange multiplier for the rate constraint. Many further improvements and variations are possible, and we refer the interested reader to Cotter et al. [2019b,a], Woodworth et al. [2017], Chamon et al. [2022], Agarwal et al. [2018] for details on a variety of optimization algorithms and their guarantees.

5.1 DETECTING NOVEL CATEGORIES IN PRACTICE

Empirically, we find that estimating the solution to Equation (5) directly with Lagrangian Optimization delivers poor results. Intuitively, this happens sinec for a loss function $l : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$, maximizing $\hat{\alpha}^l(h) = \sum_{\mathbf{x} \in S_\mathcal{T}} l(h(\mathbf{x}_i), 1)$ fits noisy labels to mixed data. That is, the dataset $S_\mathcal{T}$ contains both novel and non-novel points, trying to fit as many of them with $y = 1$ (i.e. labelling them as novelties) results in overfitting. On the other hand, minimizing $\hat{\beta}^l(h)$ fits correct labels since examples from $S_\mathcal{S}$ do not belong to the novel category, and we observe that this inhibits overfitting. Hence we find that constraining $\hat{\alpha}^{l_{01}}(h)$ while minimizing $\hat{\beta}^l(h)$ works much better than a direct implementation of Equation (5) which constrains $\hat{\beta}^{l_{01}}(h)$ while maximizing $\hat{\alpha}$. As we explain shortly, this will be performed with various values to constrain $\hat{\alpha}^{l_{01}}(h)$.

To obtain solutions for an optimization problem of the form

$$\min_{h \in \mathcal{H}} \hat{\beta}(h) \quad \text{s.t. } \hat{\alpha}(h) \geq \tilde{\alpha}, \tag{7}$$

where $\tilde{\alpha} > 0$ is some threshold on the empirical recall, define the Lagrangian

$$\mathcal{L}_{\hat{\alpha}}(h, \lambda, S_\mathcal{S}, S_\mathcal{T}) = n_\mathcal{S}^{-1} \sum_{\mathbf{x} \in S_\mathcal{S}} l_{\log}(h(\mathbf{x}), 0) + \lambda \cdot \left[n_\mathcal{T}^{-1} \sum_{\mathbf{x} \in S_\mathcal{T}} l_\sigma(h(\mathbf{x})) - \hat{\alpha} \right].$$

We replace the $0-1$ loss over $S_\mathcal{S}$ with a surrogate log-loss (denoted by $l_{\log}$), and the loss in the constraints with a sigmoid (resp. l_σ) which past work found to be effective for differentiable approximation of the indicator function in several problems, including rate-constrained optimization [Chamon et al., 2022, Goh et al., 2016, Maddison et al., 2017, Jang et al., 2017]. We optimize this Lagrangian with alternating gradient steps over the parameters of h and λ.

In summary, beyond the Lagrangian optimization procedure, our proposal for a practical algorithm includes two important components. One is a line search on the value of α, where in practice we simply solve problems with several values $\tilde{\alpha}$ that constrain $\hat{\alpha}(h)$ in Equation (7). The second component is model selection using a validation set. For each learned model we approximate its error rate on $P_\mathcal{S}$, and its recall w.r.t $P_\mathcal{T}$ (treating the target data as positively labelled) using a validation set. We then select the hypothesis h that achieves highest empirical recall ($\hat{\alpha}(h)$) whose empirical error on $P_\mathcal{S}$, $\hat{\beta}(h)$, does not exceed the user-provided value β. Hence our model selection is dictated by Equation (5) which is the overall objective of our algorithm.[7] The procedure is summarized in Algorithm 1.

Algorithm 1 CoNoC: Constrained Learning for OOD Novel Category Detection

1: **Input:** datasets $S_\mathcal{S}, S_\mathcal{T}$, hypothesis class $\mathcal{H}$, target FPR $\beta > 0$ and search range $\boldsymbol{\alpha} \in [0, 1]^L$.
2: Draw validation set $V_\mathcal{S}, V_\mathcal{T}$ from $S_\mathcal{S}, S_\mathcal{T}$ respectively
3: **for** $\alpha \in \boldsymbol{\alpha}$ **do**
4: Train model h_α to solve Equation (5) using primal-dual optimization.
5: Calculate approx. FPR $\hat{\beta}(h_\alpha) = \frac{1}{|V_\mathcal{S}|} \sum_{\mathbf{x} \in V_\mathcal{S}} h_\alpha(\mathbf{x})$, and recall $\hat{\alpha}(h_\alpha) = \frac{1}{|V_\mathcal{T}|} \sum_{\mathbf{x} \in V_\mathcal{T}} h_\alpha(\mathbf{x})$.
6: **end for**
7: return $\arg\max_{h_\alpha : \alpha \in \boldsymbol{\alpha}, \hat{\beta}(h_\alpha) < \beta} \hat{\alpha}(h_\alpha)$

Let us turn to evaluate the performance of our method.

6 EXPERIMENTS

We evaluate CoNoC in two real-world large and high-dimensional datasets.

Experimental Setting. For each dataset we have features $S = \{\mathbf{x}_i\}_{i=1}^N$ that are available to the learner and labels $\{y_i\}_{i=1}^N$ that are not. These labels are used to set up the novel categories and distribution shifts in our experiments. The procedure for each experiment is as follows; From a set of possible labels $\mathcal{Y}$, we choose $y_{\text{novel}} \in \mathcal{Y}$, and collect all examples that belong to the group $\mathcal{I} = \{i : y_i = y_{\text{novel}}\}$ into a dataset $S_{\text{novel}} = \{\mathbf{x}_i\}_{i \in \mathcal{I}}$. This divides our data into disjoint subsets S_{novel} containing the novel category and $S_{\text{seen}} = S \setminus S_{\text{novel}}$ containing the rest of the examples. We further split S_{seen} into disjoint subsets $S_\mathcal{S}$ and $S_{\mathcal{T},0}$, where we create a sub-population shift (see Equation (2)) between these two subsets by randomly drawing the prevalence of each subgroup in $\mathcal{Y} \setminus \{y_{\text{novel}}\}$ (see further details in the Appendix B).[8] Then each algorithm is run with the datasets $S_\mathcal{S}$ and $S_\mathcal{T} = S_{\mathcal{T},0} \cup S_{\text{novel}}$ as inputs (this means the true mixture proportion is $\alpha = |S_{\text{novel}}|/(|S_{\mathcal{T},0}| + |S_{\text{novel}}|)$). We repeat this procedure, creating a different subpopulation shift each time.

Baselines and evaluation metrics. We compare CoNoC with the algorithm proposed in [Gerych et al., 2022] based on propensity weighting [Bekker et al., 2019] (the idea is to estimate the density ratio of $P_\mathcal{S}$ and $P_\mathcal{T}$ and use it as importance weights, see Appendix B for details). We choose to present results for this method since it outperforms other relevant baselines (e.g. a clustering based approach

[7] Note that in principle, if we consider h^* that solves Equation (5), and h^{dual} that solves Equation (7) with $\hat{\alpha}$ set to $\alpha(h^*)$ then we can show h^{dual} is also be optimal for Equation (5). Hence our procedure is indeed an approximate solution to Equation (5)

[8] In MIMIC-III, each example has multiple labels, hence notation here is slightly abused. This is also detailed in the appendix.

Algorithm	AU-ROC/AU-ROC$_{\text{best}}$ (wins | reps.)		AU-PRC/AU-PRC$_{\text{best}}$ (wins | reps.)	
	MIMIC-III	Tabula Muris	MIMIC-III	Tabula Muris
Domain Disc.	0.940 ± 0.035 (1 | 15)	0.954 ± 0.035 (0 | 8)	0.797 ± 0.098 (1 | 15)	0.815 ± 0.173 (1 | 8)
Propensity	0.959 ± 0.028 (2 | 15)	0.953 ± 0.046 (2 | 8)	0.854 ± 0.064 (1 | 15)	0.428 ± 0.235 (0 | 8)
CoNoC	$\mathbf{0.999} \pm 0.001$ **(12 | 15)**	$\mathbf{0.988} \pm 0.020$ **(6 | 8)**	$\mathbf{0.995} \pm 0.015$ **(13 | 15)**	$\mathbf{0.999} \pm 0.001$ **(7 | 8)**

Table 1: Average Relative Area Under the Receiver-Operator Curve, AU-ROC/AU-ROC$_{\text{best}}$, where at each repetition AU-ROC$_{\text{best}}$ is taken as the area for the best method and AU-ROC is that of the evaluated method. Relative performance to best method is reported instead of raw AU-ROC since performance under different drawn distribution shifts varies. We also present the Relative Average Precision in the same manner, to summarize the Precision-Recall curve.

of Jain et al. [2020]), and since other methods for biased PU-learning Kato et al. [2018], He et al. [2018] are based on assumptions that do not hold in our setting. Our second baseline is a domain discriminator, trained to distinguish between $S_{\mathcal{S}}$ and $S_{\mathcal{T}}$, which forms the basis for many PU-learning techniques, e.g. [Elkan and Noto, 2008, du Plessis et al., 2014, Garg et al., 2021].

To calculate metrics such as accuracy, precision and recall, we need to obtain binary predictions of whether examples belong to y_{novel} or not. In both baselines, this requires an approximation of α (an MPE), that should be incorporated into the classifier (e.g. by setting the appropriate decision threshold, see Bekker and Davis [2020, Sections 5.3, 6]). Since most MPE methods are designed under the assumption that $P_{\mathcal{S}} = P_{\mathcal{T},0}$ and there is no single method that is designed to perform well under a variety of distribution shifts, we evaluate methods with metrics for predictive ability that are independent of the decision threshold. In Appendix B we include MPE results, with two different techniques [Elkan and Noto, 2008, Li and Liu, 2003] for the baselines (CoNoC does not require MPE, since we simply use the raw outputs for classification). Another point we take into account in choosing evaluation metrics is that for each repetition of the experiment, a different distribution shift is drawn. Therefore the ability of models to distinguish y_{novel} from the rest of the data can vary between repetitions. In this case, comparison of raw metrics becomes less informative and relative metrics between the different methods are more appropriate. Taking together the above considerations, Table 1 includes the following metrics.

We use Area Under the Receiver-Operator Curve (AU-ROC), and the Average Precision (Av.-Precision) as summaries of the ROC and Precision-Recall curves respectively, where the classification task is detection of y_{novel} vs. $\mathcal{Y} \setminus \{y_{\text{novel}}\}$. At each round we take the AU-ROC for the best performing method, denoted by AU-ROC$_{\text{best}}$, and for each method calculate AU-ROC/AU-ROC$_{\text{best}}$ to get a relative measure of performance (respectively for Av.-Precision). We also include the number of rounds where each method turned out to perform best. The absolute AU-ROC values for each repetition of the experiments are detailed in Appendix B.

Datasets. In the Tabula Muris single cell dataset [Consortium, 2020], the categories $\mathcal{Y}$ are cell types and features $\mathcal{X}$ are gene expressions. Then the shift between $S_{\mathcal{S}}$ and $S_{\mathcal{T},0}$ is due to differing proportions of the observed cell types. This follows the experimental setting in Garg et al. [2022], with the crucial difference that in our setting the learner does not observe cell types in $S_{\mathcal{S}}$. In the benchmark dataset devised by Harutyunyan et al. [2019] for MIMIC-III [Johnson et al., 2016], categories correspond to phenotypes (e.g., kidney disease, pneumonia, liver disease) and features are high-dimensional extracted statistics from time-series data, such as vitals and lab measurements, recorded over ICU stays (see Harutyunyan et al. [2019, Tables 2,3] for list of phenotypes and features considered). The proportion of novel categories α within $S_{\mathcal{T}}$ in these experiments is between 0.005 and 0.06, more details on these values and the effect of α on performance are in Appendix B.

6.1 RESULTS

Table 1 shows that CoNoC performs favorably with respect to the baselines in terms of relative AU-ROC and Av.-Precision on both datasets. It is the best performing method in the vast majority of repeated experiments and a further examination of the results shows that when it is not, the gap in performance is very small (see Appendix B for more details).

Takeaways from experiments. The results above demonstrate the effectiveness of constrained learning approaches in detecting novelties under distribution shift. The main choice to be made when using CoNoC is the value of β, denoting our approximation to the false positive rate of the optimal hypothesis. In our experiments, the setting of

$\beta = 0.01$ turned out to be good enough to obtain favorable performance with respect to baselines, though we observe that it is not necessarily the optimal choice. For instance in the Tabula-Muris dataset, examining the results under a lower setting of β reveals an improvement in performance (details are given in Appendix B). We attribute this to our conservative over-estimation of $\beta(h^*)$, as training an oracle classifier with true labels of y_{novel} gives a near perfect predictor in terms of test accuracy (i.e. the problem is approximately separable). On the other hand, in MIMIC-III an oracle classifier does not achieve near-perfect accuracy, and decreasing β does not improve results. This may be expected, as subgroups of patients, such as those with a certain phenotype are diagnosed using additional features that are not available to the learner. Our conclusion is that while most reasonable choices of β with a sufficiently small value lead to favorable performance w.r.t baselines, reasoning about the expected $\beta(h^*)$ with domain knowledge can further improve the performance of CoNoC.

We note that in many works on robustness to distribution shifts, benchmark tasks are designed to fail standard methods such as Empirical Risk Minimization (e.g. the Waterbirds and CelebA examples in [Sagawa et al., 2020a], or Colored MNIST in [Arjovsky et al., 2019]). In contrast, our setting randomly assigns prevalence of human-annotated subgroups (following Equation (2)), hence the shifts are *not* specially designed to create extreme and adversarial scenarios. This suggests that accounting for distribution shifts with our method might be beneficial in many cases, and the results are not limited to carefully designed examples. With that being said, in Appendix B we demonstrate by further experiments on MIMIC-III and a synthetic example that when there is no distribution shift between $P_{\mathcal{S}}$ and $P_{\mathcal{T},0}$, CoNoC does not improve over the baselines. Let us turn to conclude our work with a broad overview of the results and potential ways forward.

7 DISCUSSION, LIMITATIONS AND FUTURE WORK

We proposed a constrained learning approach for OOD novel category detection, based on a distributional assumption that bounds the shift in probability of rare events. A potential use of our method is in ML safety, where by detecting novel groups that were not part of our historical data, we may alert practitioners to issues that require further analysis. This complements methods that detect other types of safety issues such as error cases [d'Eon et al., 2022, Eyuboglu et al., 2022, Singla et al., 2021], under-performing subgroups [Subbaswamy et al., 2021], and OOD-detection methods that provide alerts on single examples instead of classes [Ruff et al., 2021].

Our formal framework is based on PU-learning and our method on advances in rate-constrained optimization. Early literature on the PU-learning problem (without distribution shift) recognizes that constrained optimization may be a useful approach, yet forgoes this path since it seems like a challenging optimization problem [Liu et al., 2002] (mixture proportion estimation based on trade-offs between recall and FPR has been explored more extensively [Blanchard et al., 2010, Scott, 2015, Jain et al., 2016a,b]). Later it has been shown that unconstrained risk minimization techniques may be devised to solve PU-learning problems under the SCAR assumption [Elkan and Noto, 2008, du Plessis et al., 2014], which seems to make constrained optimization unnecessary. Our work claims that without the SCAR assumption, a constrained learning approach can be beneficial. Importantly, we show that for our constrained learning rule, formal guarantees can be derived in settings where to the best of our knowledge, learnability in the sense of Definition 3.1 has not been shown.

Our approach has some limitations. The choice of hyperparameter β should be done carefully and requires reasoning about properties of the groups we hope to detect. Theorem 4.3 provides guidance in cases where we have a good approximation of $\beta(h^*)$, for instance when we are willing to assume that separability (approximately) holds, which is a reasonable assumption in many applications. In experiments, the performance of our method is still favorable w.r.t baselines when β is not fine-tuned. This is encouraging, yet it does not prove that such insights generalize to all real-world scenarios. Other aspects of Algorithm 1 can likely be improved, such as replacing line search over $\boldsymbol{\alpha}$ with other approaches for hyperparameter tuning, and experimenting with more sophisticated constrained optimization algorithms than the alternating primal-dual steps we use in our implementation.

Assumption 4.2 on the frequency of rare events is rather non-restrictive and is likely to hold in several cases of interest. On the other hand, its generality also means it is not tailored towards other types of distribution shifts. For instance, recent works on PU-learning make structural assumptions on the distribution shift [Garg et al., 2022, Shanmugam and Pierson, 2021] that are very different from ours and can be useful. Combining different types of assumptions into a rich framework for novelty detection under distribution shift is an exciting avenue for future research. Extensions to settings such as time-series and multiple data sources is also an exciting future direction. Recent works on invariance and stability under distribution shifts offer structural frameworks that would be interesting to explore in the context of novelty detection [Peters et al., 2016, Arjovsky et al., 2019, Subbaswamy et al., 2019, 2021, Puli et al., 2022, Wald et al., 2021]. We hope that this paper encourages further work on novelty detection in changing environments with guarantees on their performance.

Acknowledgements

We wish to thank Adarsh Subbaswamy and Amir Feder for discussions in early stages of the paper, and to Shravan Chaudhari for comments on the final version.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4): 719–760, 2020.

Jessa Bekker, Pieter Robberechts, and Jesse Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 71–85. Springer, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.

Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*, 2022.

The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583(7817):590–595, 2020.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pages 1397–1405. PMLR, 2019a.

Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.*, 20(172):1–59, 2019b.

Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981, 2022.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

Priya L Donti, David Rolnick, and J Zico Kolter. Dc3: A learning method for optimization with hard constraints. *arXiv preprint arXiv:2104.12225*, 2021.

Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 2022.

Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Ryan Rifkin, and Gal Elidan. Scalable learning of non-decomposable objectives. In *Artificial intelligence and statistics*, pages 832–840. PMLR, 2017.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural*

Information Processing Systems, 2022. URL https://openreview.net/forum?id=sde_7ZzGXOE.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

Saurabh Garg, Yifan Wu, Alexander J Smola, Sivaraman Balakrishnan, and Zachary Lipton. Mixture proportion estimation and pu learning: A modern approach. *Advances in Neural Information Processing Systems*, 34: 8532–8544, 2021.

Saurabh Garg, Sivaraman Balakrishnan, and Zachary C. Lipton. Domain adaptation under open set label shift, 2022. URL https://arxiv.org/abs/2207.13048.

Walter Gerych, Thomas Hartvigsen, Luke Buquicchio, Emmanuel Agu, and Elke Rundensteiner. Recovering the propensity score from biased positive unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6694–6702, 2022.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8401–8409, 2019.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hashimoto18a.html.

Fengxiang He, Tongliang Liu, Geoffrey I. Webb, and Dacheng Tao. Instance-dependent PU learning by bayesian optimal relabeling. *CoRR*, abs/1808.02180, 2018. URL http://arxiv.org/abs/1808.02180.

Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016a.

Shantanu Jain, Martha White, Michael W Trosset, and Predrag Radivojac. Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*, 2016b.

Shantanu Jain, Justin Delano, Himanshu Sharma, and Predrag Radivojac. Class prior estimation with biased positives and unlabeled examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4255–4263, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International conference on learning representations*, 2018.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, volume 4, pages 180–191. Toronto, Canada, 2004.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI*, volume 3, pages 587–592. Citeseer, 2003.

Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. Partially supervised classification of text documents. In *ICML*, volume 2, pages 387–394. Sydney, NSW, 2002.

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

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2021.

Benjamin Nachman and David Shih. Anomaly detection with density estimation. *Physical Review D*, 101(7): 075042, 2020.

Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. Open-world machine learning: applications, challenges, and opportunities. *ACM Computing Surveys*, 55(10):1–37, 2023.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Aahlad Manas Puli, Lily H Zhang, Eric Karl Oermann, and Rajesh Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=12RoR2o32T.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2008.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=ryxGuJrFvS.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/sagawa20a.html.

Suchi Saria and Adarsh Subbaswamy. Tutorial: safe and reliable machine learning. *arXiv preprint arXiv:1904.07204*, 2019.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Clayton Scott. A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 838–846, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/scott15.html.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Divya Shanmugam and Emma Pierson. Quantifying inequality in underreported medical conditions, 2021. URL https://arxiv.org/abs/2110.04133.

Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding failures of deep networks via robust feature extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12853–12862, 2021.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International Conference on Artificial Intelligence and Statistics*, pages 2611–2619. PMLR, 2021.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in neural information processing systems*, 34: 2215–2227, 2021.

Yoav Wald, Gal Yona, Uri Shalit, and Yair Carmon. Malign overfitting: Interpolation can provably preclude invariance. *arXiv preprint arXiv:2211.15724*, 2022.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953. PMLR, 07–10 Jul 2017. URL `https://proceedings.mlr.press/v65/woodworth17a.html`.

Hu Xu, Bing Liu, Lei Shu, and P Yu. Open-world learning and application to product classification. In *The World Wide Web Conference*, pages 3413–3419, 2019.