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

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

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

A PROOFS OF FORMAL RESULTS

A.1 PROOF OF ERROR GUARANTEES

We recall the notion of distance defined in the main paper, inspired by the $\mathcal{H}$-divergence in the domain adaptation literature [Kifer et al., 2004, Ben-David et al., 2010],

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

Let us prove the first part of Theorem 4.3:

Lemma A.1. *For a novelty detection problem as in Definition 3.1, let $h \in \mathcal{H}$ and denote $\alpha(h) = \mathbb{E}_{P_{\mathcal{T}}}[h(\mathbf{x})]$, while $\beta(h) = \mathbb{E}_{P_{\mathcal{S}}}[h(\mathbf{x})]$. Define,*

$$
\begin{aligned}
\bar{R}^{l_{01}}_{\mathcal{T}}(h) = {} & [\alpha - \alpha(h)] + \\
& (1-\alpha)\left[\beta(h) + d_{\mathcal{H},\beta(h)}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right)\right] .
\end{aligned}
$$

Then we have that $R^{l_{01}}_{\mathcal{T}}(h) \leq \bar{R}^{l_{01}}_{\mathcal{T}}(h)$.

Proof. We decompose the error as follows:

$$
\begin{aligned}
R^{l_{01}}_{\mathcal{T}}(h) &= (1-\alpha) \cdot \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T},0}}[h(\mathbf{x})] + \alpha \cdot \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T},1}}[1 - h(\mathbf{x})] \\
&= (1-\alpha) \cdot \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T},0}}[h(\mathbf{x})] + \alpha \cdot \left(1 - \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T},1}}[h(\mathbf{x})]\right) \\
&= \alpha - \mathbb{E}_{\mathbf{x} \sim (1-\alpha)P_{\mathcal{T},0} + \alpha P_{\mathcal{T},1}}[h(\mathbf{x})] + 2 \cdot (1-\alpha) \cdot \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T},0}}[h(\mathbf{x})] \\
&= \alpha - \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T}}}[h(\mathbf{x})] + 2 \cdot (1-\alpha) \cdot \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T},0}}[h(\mathbf{x})] \\
&= \alpha - \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T}}}[h(\mathbf{x})] + 2 \cdot (1-\alpha) \cdot \left[\mathbb{E}_{\mathbf{x} \sim P_{\mathcal{S}}}[h(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T},0}}[h(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{S}}}[h(\mathbf{x})]\right] \\
&\leq \alpha - \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{T}}}[h(\mathbf{x})] + 2 \cdot (1-\alpha) \cdot \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{S}}}[h(\mathbf{x})] + 2 \cdot (1-\alpha) \cdot \left|P_{\mathcal{T},0}\left(h(\mathbf{x}) = 1\right) - P_{\mathcal{S}}\left(h(\mathbf{x}) = 1\right)\right| \\
&= \alpha - \alpha(h) + 2 \cdot (1-\alpha)\beta(h) + 2 \cdot (1-\alpha) \cdot \left|P_{\mathcal{T},0}\left(h(\mathbf{x}) = 1\right) - P_{\mathcal{S}}\left(h(\mathbf{x}) = 1\right)\right| \\
&\leq \alpha - \alpha(h) + 2 \cdot (1-\alpha)\left[\beta(h) + d_{\mathcal{H},\beta(h)}(P_{\mathcal{S}} \| P_{\mathcal{T},0})\right] = \bar{R}^{l_{01}}_{\mathcal{T}}(h).
\end{aligned} \tag{1}
$$

□

With this inequality in hand, we can now prove Proposition 4.1.

Proposition. *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.* [1] *Scarcity-of-Unicorns (Assumption 4.2) holds with $\beta, \varepsilon_{shift}$ set to* 0.

Proof. Let $\mathcal{B}$ denote the measurable subsets w.r.t both $P_{\mathcal{S}}$ and $P_{\mathcal{T},0}$ and define,

$$d_{1,\beta}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right) = \sup_{B \in \mathcal{B} : P_{\mathcal{S}}(B) \leq \beta} 2 \Big| P_{\mathcal{S}}(B) - P_{\mathcal{T}}(B) \Big| .$$

Taking for any $g \in \mathcal{H}$ the subset of inputs where it equals 1, $I(g) \subseteq \mathcal{X}$, and $I(\mathcal{H}) = \{I(g) \ : \ g \in \mathcal{H}\}$, we see that $I(\mathcal{H}) \subseteq \mathcal{B}$ and hence we have

$$d_{\mathcal{H},\beta}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right) \leq d_{1,\beta}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right), \tag{2}$$

for any $\beta \geq 0$. Under separability we have that $d_{1,0}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right) = 0$, since if $P_{\mathcal{T},0}(\tilde{B}) > 0$ for some $\tilde{B} \in \mathcal{B}$ then we must also have $P_{\mathcal{S}}(\tilde{B}) > 0$ and then $\tilde{B} \notin \{B \in \mathcal{B} : P_{\mathcal{S}}(B) \leq 0\}$. This means that for any $\tilde{B} \in \{B \in \mathcal{B} : P_{\mathcal{S}}(B) \leq 0\}$ we must have $P_{\mathcal{T},0}(\tilde{B}) = 0$ and hence $d_{1,0}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right) = 0$. The claim is proved by combining this with Equation (2), to obtain $d_{\mathcal{H},\beta}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right) \leq d_{1,\beta}\left(P_{\mathcal{S}} \| P_{\mathcal{T},0}\right) = 0$, and since the divergence is non-negative it must equal 0, meaning Scarcity-of-Unicorns holds with $\beta, \varepsilon_{\text{shift}}$ set to 0. □

Next let us restate the proposed learning rule

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

We derive generalization bounds for solutions to the empirical version of this problem. Recall the Rademacher complexity of $\mathcal{H}$ with respect to n samples from distribution P is denoted by $R_{n,P}(\mathcal{H}) = \mathbb{E}_{P^n}\left[\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}[\sup_{h \in \mathcal{H}} \sum_i \sigma_i h(\mathbf{x}_i)]\right]$, the following statement gives the statistical guarantee we require for our result.

Lemma A.2. *Let $\mathcal{H}$ be a hypothesis class with Rademacher complexities $R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})$ and $R_{n_{\mathcal{T}},P_{\mathcal{T}}}(\mathcal{H})$ respectively, and $\hat{h}$ a solution to the empirical estimate of Equation* (5)*, with $\beta \geq \beta(h^*) + \frac{R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{S}}}}$. Then with probability at least $1 - 4\delta$ we have that simultaneously,*

$$\alpha(\hat{h}) \geq \alpha(h^*) - R_{n_{\mathcal{T}},P_{\mathcal{T}}}(\mathcal{H}) - \sqrt{\frac{2\ln(1/\delta)}{n_{\mathcal{T}}}}, \tag{3}$$

$$\beta(\hat{h}) \leq \beta + \frac{R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{S}}}}. \tag{4}$$

Proof. From standard Rademacher bounds on the risk of classifiers in a hypothesis class (e.g. Bartlett and Mendelson [2002, Theorem 5]), we have that with probability $1 - 2\delta$:

$$\left|\hat{\beta}(h) - \beta(h)\right| \leq \frac{R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{S}}}} \quad \forall h \in \mathcal{H}.$$

Therefore Equation (4) holds since $\hat{\beta}(\hat{h}) \leq \beta$. Also, from the lower bound on β assumed in our lemma statement, all classifiers with False Positive Rate smaller than $\beta(h^*)$ will be in the feasible set of Equation (5). This follows from the above inequality since for all $h \in \mathcal{H}$ where $\beta(h) < \beta(h^*)$ it holds that

$$\hat{\beta}(h) \leq \beta(h) + \frac{R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{S}}}}$$

$$\Rightarrow \hat{\beta}(h) \leq \beta(h^*) + \frac{R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{S}}}}.$$

[1] 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.

Specifically, this means that with probability at least $1-2\delta$, h^* is a feasible solution to Equation (5) and taking $\hat{h}$ that is optimal for Equation (5), we can gather that $\hat{\alpha}(\hat{h}) \geq \hat{\alpha}(h^*)$. For the second part of the proof, we use the same inequality as before to obtain that with probability at least $1-2\delta$,

$$|\hat{\alpha}(h) - \alpha(h)| \leq \frac{R_{n_{\mathcal{T}},P_{\mathcal{T}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{T}}}} \quad \forall h \in \mathcal{H}. \tag{5}$$

Then we take a union bound to conclude that with probability at least $1-4\delta$,

$$\begin{aligned} \alpha(h^*) - \alpha(\hat{h}) &= \alpha(h^*) - \hat{\alpha}(h^*) + \hat{\alpha}(h^*) - \hat{\alpha}(h) + \hat{\alpha}(h) - \alpha(\hat{h}) \\ &\leq \alpha(h^*) - \hat{\alpha}(h^*) + \hat{\alpha}(h) - \alpha(\hat{h}) \\ &\leq R_{n_{\mathcal{T}},P_{\mathcal{T}}}(\mathcal{H}) + \sqrt{\frac{2\ln(1/\delta)}{n_{\mathcal{T}}}}. \end{aligned}$$

The first inequality follows from our previous conclusion that $\hat{\alpha}(\hat{h}) \geq \hat{\alpha}(h^*)$ and the second from Equation (5). □

With the concentration properties in hand, recall that we assume $d_{\mathcal{H},\beta}(P_{\mathcal{S}} \| P_{\mathcal{T},0}) \leq \varepsilon_{\text{shift}}$ for some fixed $\beta, \varepsilon_{\text{shift}} \geq 0$, and let us combine this with the previous claims to bound the error as required for the second part of Theorem 4.3.

Lemma A.3. *Let $h^* \in \mathcal{H}$ be a minimizer of $R_{\mathcal{T}}^{l_{01}}(h)$ and assume $\hat{h}$ solves Equation* (5) *with $\beta \geq \beta(h^*) + \frac{R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{S}}}}$, then with probability at least $1-4\delta$ it holds that*

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

Proof. Let us assume that the inequalities in Lemma A.2 hold, which occurs with probability at least $1-4\delta$. We write down the gap in risks between the hypotheses $\hat{h}$ and h^*, while using these inequalities:

$$\begin{aligned} R_{\mathcal{T}}^{l_{01}}(\hat{h}) - R_{\mathcal{T}}^{l_{01}}(h^*) &= \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T}}}\left[h^*(\mathbf{x}) - \hat{h}(\mathbf{x})\right] + 2\cdot(1-\alpha)\cdot\mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},0}}\left[\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right] \\ &= \alpha(h^*) - \alpha(\hat{h}) + 2\cdot(1-\alpha)\cdot\mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},0}}\left[\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right] \\ &= \alpha(h^*) - \alpha(\hat{h}) + 2\cdot(1-\alpha)\cdot\Big[\mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},0}}\left[\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}\left[\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right] \\ &\qquad\qquad + \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}\left[\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right]\Big] \\ &= \alpha(h^*) - \alpha(\hat{h}) + 2\cdot(1-\alpha)\cdot\Big[\mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},0}}\left[\hat{h}(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}\left[\hat{h}(\mathbf{x})\right] \\ &\qquad\qquad + \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}[h^*(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},0}}[h^*(\mathbf{x})] \\ &\qquad\qquad + \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}\left[\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right]\Big] \\ &\leq R_{n_t,P_{\mathcal{T}}}(\mathcal{H}) + \sqrt{\frac{2\ln(1/\delta)}{n_{\mathcal{T}}}} + 2\cdot(1-\alpha)\cdot\Big[\mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},0}}\left[\hat{h}(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}\left[\hat{h}(\mathbf{x})\right] \\ &\qquad\qquad + \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}[h^*(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{T},0}}[h^*(\mathbf{x})] \\ &\qquad\qquad + \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}\left[\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right]\Big] \\ &\leq R_{n_t,P_{\mathcal{T}}}(\mathcal{H}) + \sqrt{\frac{2\ln(1/\delta)}{n_{\mathcal{T}}}} + 2\cdot(1-\alpha)\cdot\left[2\epsilon_{\text{shift}} + \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{S}}}\left[\hat{h}(\mathbf{x}) - h^*(\mathbf{x})\right]\right]\Big] \\ &\leq R_{n_t,P_{\mathcal{T}}}(\mathcal{H}) + \sqrt{\frac{2\ln(1/\delta)}{n_{\mathcal{T}}}} + 2\cdot(1-\alpha)\cdot\left[2\epsilon_{\text{shift}} + \beta - \beta(h^*) + \frac{R_{n_{\mathcal{S}},P_{\mathcal{S}}}(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathcal{S}}}}]\right]. \end{aligned}$$

The first and third inequalities are obtained by applying Lemma A.2, the second holds due to Assumption 4.2. It is easy to see that the above expression lower bounds the one in our claim since $\alpha \in [0, 1]$ and hence our proof is concluded. □

The theorem in the main paper follows directly from the statements we proved above.

Proof of Theorem 4.3. The first part of the theorem follows directly from Lemma A.1, while the second is a direct consequence of Lemma A.3. □

Possible Extension of Results. We note that one clear gap in our results is that they apply to l_{01} instead of other surrogate losses that we use in practice. This is also a gap in the work of Ben-David et al. [2010] on domain adaptation and it is a result of using the $d_{\mathcal{H}}$ divergence. Hence a possible path to generalize our result is to use other divergences in the proof of Lemma A.1, e.g. like that used in Mansour et al. [2009] to extend the results of Ben-David et al. [2010]. The other component for proving Theorem 4.3, namely the proof of Lemma A.2, does not depend explicitly on l_{01} and can be extended using standard arguments on Rademacher complexity.

A.2 SUFFICIENT AND NECESSARY CONDITIONS FOR LEARNING AND MIXTURE PROPORTION ESTIMATION

We complete proofs of claims made in the main paper with simple proofs for the necessity and sufficiency of assumptions in detecting classes under distribution shift. Our first claim was impossibility of learning when no distributional assumptions are made.

Proposition. *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 $R_{\mathcal{T}}^{l_{01}}(h^*) = 0$, while $\mathbb{E}_{S_{\mathcal{S}},S_{\mathcal{T}}}\left[R_{\mathcal{T}}^{l_{01}}(\mathcal{A}(S_{\mathcal{S}}, S_{\mathcal{T}}))\right] \geq 0.5$.*

Proof. Let $\alpha = 0.5$, and P, Q, D distributions such that for some hypothesis class $\mathcal{H}$ there is $h^* \in \mathcal{H}$ for which $\mathbb{E}_Q[h^*(\mathbf{x})] = 0, \mathbb{E}_D[h^*(\mathbf{x})] = 1$. Consider two problems where in one $P_{\mathcal{S}} = P, P_{\mathcal{T},0} = Q, P_{\mathcal{T},1} = D$, and in the other $\tilde{P}_{\mathcal{S}} = P, \tilde{P}_{\mathcal{T},0} = D, \tilde{P}_{\mathcal{T},1} = Q$. That is, the roles of D and Q are switched between the two problems. Notice that the target distributions $P_{\mathcal{T}}$ and $\tilde{P}_{\mathcal{T}}$ are the same in both problems since $P_{\mathcal{T}} = 0.5P_{\mathcal{T},0} + 0.5P_{\mathcal{T},1} = 0.5\tilde{P}_{\mathcal{T},0} + 0.5\tilde{P}_{\mathcal{T},1} = \tilde{P}_{\mathcal{T}}$. Hence training data for a learning algorithm $\mathcal{A}$ are drawn from the same distribution. Yet if we denote the risk w.r.t to the second problem by $\tilde{R}_{\mathcal{T}}^{l_{01}} : \mathcal{H} \to [0, 1]$ then for any $h \in \mathcal{H}$ if $R_{\mathcal{T}}^{l_{01}}(h) = \varepsilon$ it holds that $\tilde{R}_{\mathcal{T}}^{l_{01}} = 1 - \varepsilon$. Hence a learning algorithm $\mathcal{A}$ that achieves expected error smaller than 0.5 on one problem will incur expected error larger than 0.5 in the other, which proves the statement. □

Note that Blanchard et al. [2010] prove that irreducibility is required for identification of α and for learning under the SCAR assumption (i.e. $P_{\mathcal{S}} = P_{\mathcal{T},0}$). Irreducibility states that $\max_{\gamma \geq 0}\{P_{\mathcal{T},1} = \gamma P_{\mathcal{S}} + (1-\gamma)Q \ : \ Q \in \Delta\} = 0$, where Δ is the set of all distributions over the measurable set $\mathcal{X}$. While the statement we prove above is much simpler and says we cannot learn unless something is known about the target distribution, it does not follow from their proof. When irreducibility does not hold, then there is also no $h^* \in \mathcal{H}$ with $R_{\mathcal{T}}^{l_{01}}(h^*) = 0$ since $P_{\mathcal{T},1}$ is a mixture with a non-zero component of $P_{\mathcal{S}}$, and thus cannot be perfectly separated from $P_{\mathcal{T},0} = P_{\mathcal{S}}$. Our statement demands the existence of h^* that achieves loss 0, and thus the conditions for the statments are different.

The last remaining claim we made in the paper and has not been proven above, is that under separability and given perfect knowledge of $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ the mixture proportion α can be recovered.

Lemma A.4. *Assume the novel class detection problem satisfies* ***(No-Overlap)****: there exists a subset $B_{sep} \subset \mathcal{B}$ such that $P_{\mathcal{S}}(B_{sep}) = 1$, $P_{\mathcal{T},0}(B_{sep}) = 1$ and $P_{\mathcal{T},1}(B_{sep}) = 0$, Then α is identifiable.*

Proof. Define the set of distributions over $\mathcal{X}$ that fully overlaps with $P_{\mathcal{S}}$, that is $\mathcal{P}(P_{\mathcal{S}}) = \{P \in \Delta : P(B) > 0 \Rightarrow P_{\mathcal{S}}(B) > 0 \ \forall B \in \mathcal{B}\}$ where $\mathcal{B}$ is the set of all measurable subsets of $\mathcal{X}$. Let us define the following principle for approximating α:

$$\hat{\alpha} = \arg\min_{\gamma \in [0,1]} \{P_{\mathcal{T}} = (1-\gamma)P + \beta Q \ : \ P \in \mathcal{P}(P_{\mathcal{S}}) \text{ and } Q \text{ a distribution}\}. \tag{6}$$

Because given the ground truth distributions $P_{\mathcal{T},0}, P_{\mathcal{T},1}$, we know that $P_{\mathcal{T}} = (1-\alpha)P_{\mathcal{T},0} + \alpha P_{\mathcal{T},1}$, we have that $P_{\mathcal{T}}(X_{sep}) = (1-\alpha)P_{\mathcal{T},0}(X_{sep}) = 1-\alpha$. Clearly, taking $\gamma = \alpha$, $P = P_0$ and $Q = P_1$ gives a feasible solution to the right hand side of Equation (6). Now assume that there exists some feasible solution with $\gamma < \alpha$, $P \in \mathcal{P}(P_{\mathcal{S}})$ and a distribution Q. Then $P_{\mathcal{T}}(X_{sep}) \geq (1-\gamma)P(X_{sep}) = 1-\gamma > 1-\alpha$, which contradicts our conclusion that $P_{\mathcal{T}}(X_{sep}) = 1-\alpha$ must hold. Hence α is identifiable and given by the solution to Equation (6). □

Having proven all claims made in the main paper, we turn to a short supplementary discussion on the divergence $d_{\mathcal{H},\beta}(P\|Q)$ we used in our assumptions and corresponds to the frequency that rare events in P take in distribution Q.

A.3 FURTHER DISCUSSION ON $d_{\mathcal{H},\beta}(P\|Q)$

In the domain adaptation literature [Ben-David et al., 2010, Kifer et al., 2004], the $\mathcal{H}$-divergence defined as

$$d_{\mathcal{H}}(P, Q) = 2 \sup_{g \in \mathcal{H}} |P\left[I(g)\right] - Q\left[I(g)\right]|,$$

is used for two reasons. As in our use of $d_{\mathcal{H},\beta}(P\|Q)$, the term $d_{\mathcal{H}}(P_{\mathcal{S}}, P_{\mathcal{T}})$ is included in an upper bound on error w.r.t a target distribution. While $d_{\mathcal{H}}(P_{\mathcal{S}}, P_{\mathcal{T}})$ can be estimated from data, and therefore one can optimize the resulting upper bound w.r.t $\mathcal{H}$, this is not true in our case. Unfortunately calculation of $d_{\mathcal{H},\beta}(P_{\mathcal{S}}, P_{\mathcal{T},0})$ requires a sample from $P_{\mathcal{T},0}$, and to obtain an upper bound we require an assumption about the magnitude of the divergence. The second reason that the $\mathcal{H}$-divergence is used in domain adaptation is that it provides a much tighter bound than the one based on standard divergences between distributions, e.g. in our case it is an alternative to $d_{1,\beta}\left(P\|Q\right) = \sup_{B \in \mathcal{B}: P(B) \leq \beta} 2\Big|P(B) - Q(B)\Big|$, taken w.r.t measurable subsets $\mathcal{B}$ under the two distributions. This indeed tightens our bounds by weakening the assumption required in Assumption 4.2, though it has no practical implication on the algorithm we use.

It is worth noting that if we obtain samples from distributions P and Q then $d_{\mathcal{H},\beta}(P\|Q)$ can be estimated efficiently by solving a rate-constrained classification problem. This can be helpful in case we wish to reason about $d_{\mathcal{H},\beta}(P_{\mathcal{S}}\|P_{\mathcal{T},0})$ in a data-driven manner. For instance, say $P_{\mathcal{S}}$ is a distribution over EHRs in one hospital, and we have a dataset from another hospital with corresponding distribution Q where we do not think that novel groups have emerged. If we are willing to assume that in our target distribution $P_{\mathcal{T}} = \alpha P_{\mathcal{T},1} + (1-\alpha)P_{\mathcal{T},0}$, it holds that $d_{\mathcal{H},\beta}(P_{\mathcal{S}}\|P_{\mathcal{T},0})$ does not exceed $d_{\mathcal{H},\beta}(P_{\mathcal{S}}\|Q)$, then we can get an upper bound on the divergence we are interested in by estimating $d_{\mathcal{H},\beta}(P_{\mathcal{S}}\|Q)$ from data. The following lemma tells us this can be done by solving a rate-constrained Empirical Risk Minimization problem.

Lemma A.5. *Let S_P, S_Q be i.i.d sampled datasets of size n from P, Q respectively, $\mathcal{H}$ a symmetric hypothesis class (i.e. that $1 - h \in \mathcal{H}$ for any $h \in \mathcal{H}$), and $d_{\mathcal{H},\beta}(\hat{P}\|\hat{Q})$ the empirical estimate of $d_{\mathcal{H},\beta}(P\|Q)$ (i.e. where we replace P, Q with empirical distributions defined by a uniform distribution over the examples in the datasets). Then we have that:*

$$d_{\mathcal{H},\beta}(S_P\|S_Q) = 2\left(1 - \min_{h \in \mathcal{H}: n^{-1}\sum_{\mathbf{x} \in S_P} h(\mathbf{x}) \leq \beta} \left[\frac{1}{n}\sum_{\mathbf{x}: h(\mathbf{x})=1} I[\mathbf{x} \in S_P] + \frac{1}{n}\sum_{\mathbf{x}: h(\mathbf{x})=0} I[\mathbf{x} \in S_Q]\right]\right)$$

Proof. Denoting by $\hat{P}, \hat{Q}$ the empirical distributions corresponding to S_P and S_Q, we will follow the proof of Ben-David et al. [2010, Lemma 2] to show that for any $h \in \mathcal{H}$,

$$\hat{Q}[I(h)] - \hat{P}[I(h)] = 1 - \left[\frac{1}{n}\sum_{\mathbf{x}: h(\mathbf{x})=1} I[\mathbf{x} \in S_P] + \frac{1}{n}\sum_{\mathbf{x}: h(\mathbf{x})=0} I[\mathbf{x} \in S_Q]\right]. \tag{7}$$

Once this is shown, we get the result in the statement by maximizing w.r.t $h \in \mathcal{H} : \hat{\beta}(h) \leq \beta$, since $\hat{\beta}(h) = n^{`-1}\sum_{\mathbf{x} \in S_P} h(\mathbf{x})$ by definition. The absolute value on the left hand side of the above equation, which appears in the definition of $d_{\mathcal{H},\beta}(P\|Q)$ is obtained from the symmetry of $\mathcal{H}$. Now for completeness let us give the proof of the required equality. We start by taking,

$$1 = \frac{1}{2n}\sum_{\mathbf{x}: h(\mathbf{x})=0} I[\mathbf{x} \in S_P] + I[\mathbf{x} \in S_Q] + \frac{1}{2n}\sum_{\mathbf{x}: h(\mathbf{x})=1} I[\mathbf{x} \in S_P] + I[\mathbf{x} \in S_Q],$$

and plugging-in to the right hand side of Equation (7) we get:

$$\begin{aligned}
1-&\left[\frac{1}{n}\sum_{\mathbf{x}:h(\mathbf{x})=1} I[\mathbf{x}\in S_P]+\frac{1}{n}\sum_{\mathbf{x}:h(\mathbf{x})=0} I[\mathbf{x}\in S_Q]\right]\\
=&\frac{1}{2n}\sum_{\mathbf{x}:h(\mathbf{x})=0} I[\mathbf{x}\in S_P]-I[\mathbf{x}\in S_Q]+\frac{1}{2n}\sum_{\mathbf{x}:h(\mathbf{x})=1} I[\mathbf{x}\in S_Q]-I[\mathbf{x}\in S_P]\\
=&\frac{1}{2}(1-\hat{P}[I(h)]-1+\hat{Q}[I(h)])+\frac{1}{2}\left(\hat{Q}[I(h)]-\hat{P}[I(h)]\right)\\
=&\hat{Q}[I(h)]-\hat{P}[I(h)].
\end{aligned}$$

□

The lemma tells us that the divergence can be estimated with rate-constrained optimization, and using similar techniques to the ones used in other constrained learning works [Donini et al., 2018, Chamon et al., 2022] and in Theorem 4.3, we can obtain generalization bounds for estimation of $d_{\mathcal{H},\beta}(P\|Q)$ from a finite sample.

B ADDITIONAL DETAILS ON EXPERIMENTAL RESULTS

In this section we provide additional details on our experiments. We start with details about relative AU-ROC and Av.-Precision in cases where CoNoC is not the best performing method. On MIMIC-III CoNoC obtains a mean relative AU-ROC (denoted by AU-ROC/AU-ROC$_{\text{best}}$) of 0.997 even in the 3 rounds where it is not the best method. For comparison, the propensity baseline achieves a mean relative AU-ROC of 0.953 in rounds where it is not the best method. Similarly with relative Av.-Precision, the propensity baseline achieves a mean value of 0.844 in rounds where it is not the best performing method, while CoNoC has mean 0.964 under the respective rounds. For Tabula-Muris, the relative AU-ROC upon not being best performing is similar for all methods, but in relative Av.-Precision CoNoC only loses one round and it is comparable to the best performing method as it achieves 0.995 relative AU-ROC. The losses for other methods are by a far more significant margin, as implied in Table 1.

The rest of this section begins by describing the way we generate distribution shifts, continue to implementation details, and finally provide a few additional analyses.

Generation of distribution shifts. As explained in Section 6, from the collection of available labels in the dataset $\mathcal{Y}$, one label is taken as the novel subgroup y_{novel}. Then for each label $y\in\mathcal{Y}\setminus\{y_{\text{novel}}\}$, denoting by $I_y=\{i:y_i=1\}$ the examples with label y, we draw a number γ_y uniformly from $[0.1,1]$ and put (randomly drawn) $\gamma_y\cdot|I_y|$ of the examples with label y in $S_{\mathcal{S}}$. The other $(1-\gamma_y)\cdot|S_{\mathcal{T}}|$ go in $S_{\mathcal{T}}$. In MIMIC-III, the labels are phenotypes and each example (corresponding to a patient admission) can be assigned with more than one label. In this case we iterate over the different labels in some order and create the shift for each one as described above, but I_y will not contain indices where patients were assigned with a label that came before it in the iterative process. Before the iterative process begins we also keep away all the examples belonging to the novel class $\{\mathbf{x}\}_{i\in I_{\text{novel}}}$ where $I_{\text{novel}}=\{i:y_{\text{novel}}=1\}$ and put them in $S_{\mathcal{T}}$.

Finally, we also draw validation sets $V_{\mathcal{S}},V_{\mathcal{T}}$ out of $S_{\mathcal{S}}$ and $S_{\mathcal{T}}$ respectively, to be used for model selection and validation as described later in the next part.

Details on implementation and model selection. In our experiments we use a multilayer perceptron with 2 hidden layers for Tabula-Muris (feature dimension 2866, number of hidden units at each layer 64), following [Cao et al., 2021], and a linear model for MIMIC-III (features dimension is 714) used as one of the methods in [Harutyunyan et al., 2019]. We note that the computational complexity of the algorithm depends on the implementation of the constrained optimization step (line 4 in Algorithm 1), results on some methods are given in Chamon et al. [2022], Cotter et al. [2019] and to obtain the computational complexity of CoNoC we should multiply the running time by L, which is the size of $\boldsymbol{\alpha}$. It is likely that this runtime can be reduced significantly by more efficient search methods for α, we keep exploration of implementation improvements for CoNoC to future work.

The Domain Discriminator baseline, h_{disc}, is trained by minimizing the log-loss. For MIMIC-III we use the cross validated Logistic Regression method from sklearn [Pedregosa et al., 2011], while for Tabula Muris we train with Adam [Kingma and Ba, 2015] for 150 epochs and select the weights at the end of the epoch where the model achieves highest accuracy (on classification of $V_{\mathcal{S}}$ vs. $V_{\mathcal{T}}$) over a held-out validation set. The propensity-weighted baseline is trained in the same manner, except we use the following weighted loss from Bekker et al. [2019], Gerych et al. [2022]:

$$R_{\mathcal{S},e}^{\log}(h) = n_{\mathcal{S}}^{-1} \sum_{\mathbf{x}\in S_{\mathcal{S}}} e(\mathbf{x})^{-1} l_{\log}(h(\mathbf{x}), 0) + (1 - e(\mathbf{x})^{-1}) l_{\log}(h(\mathbf{x}), 1) \\ + n_{\mathcal{T}}^{-1} \sum_{\mathbf{x}\in S_{\mathcal{T}}} l_{\log}(h(\mathbf{x}), 1). \tag{8}$$

Here $e(\mathbf{x})$ is the propensity score, which we obtain from the output of the Domain Discriminator model, h_{disc}. Namely, it is the probability assigned by the model used in h_{disc} that the example $\mathbf{x}$ is from $P_{\mathcal{S}}$. We calibrate the Domain Discriminator model over the validation set using Platt scaling [Platt et al., 1999] before retrieving $e(\mathbf{x})$, this improves the propensity score estimation and also downstream performance on the learning task. Finally, model selection for this baseline is the same as for the h_{disc}, except we use the weighted accuracy (i.e. Equation (8) with l_{0-1} instead of unweighted accuracy).

In both datasets, CoNoC is trained by alternating steps of Adam for the model parameters, and gradient descent for the Lagrange multiplier. Model selection for CoNoC is done by selecting weights at the end of the epoch where the recall, $\hat{\alpha}(h) = |V_{\mathcal{T}}|^{-1} \sum_{\mathbf{x}\in V_{\mathcal{T}}} h(\mathbf{x})$, is highest and False Positive Rate, $|V_{\mathcal{S}}|^{-1} \sum_{\mathbf{x}\in V_{\mathcal{S}}} h(\mathbf{x})$, is smaller than $\beta = 0.01$. We train models with several values of α and choose the final model using this criterion.

For additional details on the implementation of methods, please advise our code, to be released here upon publication.

Mixture Proportion Estimation As mentioned in Section 6, the outputs of h_{disc} and the propensity weighted risk minimizer are not good binary classifiers in case we simply set their decision threshold at probability 0.5 for $y = 1$. Instead we need to adjust this threshold with a Mixture Proportion Estimation. We use methods from Elkan and Noto [2008], Li and Liu [2003], denoted by EN and $FPR < 0.1$ respectively. The first estimator is designed under the assumption that $P_{\mathcal{S}} = P_{\mathcal{T},0}$, while the second one is included since it follows the model selection principle we use in our method of thresholding the FPR. To report the MPE for CoNoC we simply use $\hat{\alpha}(h)$, the fraction of positive labels predicted on the validation set from the target distribution.

For the same reasons mentioned in Section 6, the metric we use for evaluation is a relative metric. We denote the estimated mixture proportion by $\hat{\alpha}$, the true proportion by α, and use a quantity we call Relative Absolute Mixture Proportion Error (RAMPE), $|1 - \frac{\hat{\alpha}}{\alpha}|$. E.g., if the novel class comprises 4% of the population, and our approximation is 1%, the RAMPE is 0.75.

As seen in Figure 1, the combination of the estimator from Elkan and Noto [2008] and the domain discriminator give the best performance for MIMIC-III (note that the domain discriminator is worst in terms of AU-ROC and Av.-Precision according to Table 1 of the main paper). However, this estimator is very inaccurate for the Tabula Muris dataset. Occurences of such large errors may be expected, as the estimator is designed under the assumption that $P_{\mathcal{S}} = P_{\mathcal{T},0}$. Hence, while it may happen to provide a reasonable estimate at times, it can have very large errors at others. These results suggest that in terms of estimating the mixture proportion, no single combination of baseline algorithm and MPE technique is preferred for both datasets, while CoNoC performs comparably to using the FPR, which avoids the very large errors that estimators based on the SCAR assumption can incur.

Effect of distribution shift on performance of CoNoC vs. baselines. Continuing our motivating example from Section 3.1, Figure 2 shows how the methods compare under the same categories, but when there is no distribution shift, i.e. $P_{\mathcal{S}} = P_{\mathcal{T},0}$. As expected, the methods learn weights that are very close to one another, but differ in their bias terms. Hence in terms of detection abilities for the novel class they are equivalent. That is, under an appropriate setting of the decision threshold, both models will detect the novel class.

Details on values of α and raw AU-ROC values. We include raw AU-ROC and AU-PRC values for all repetitions of our experiments. Table 1 has gives the details for the Tabula-Muris experiments. We observe that in repetitions where α is large, the gap between CoNoC and baselines is somewhat smaller. This is intuitive, since in case the addition of the novel category makes up most of the distribution shift between $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$, we arrive at a case that is somewhat similar to our synthetic example in Figure 2 where CoNoC coincides with a domain discriminator. However, in all repetitions CoNoC has either the best AU-ROC or AU-PRC, and in 5 out of 8 runs it is best on both metrics. We note that for very small classes (i.e. smaller than 0.002), all methods perform poorly and we do not include such novel categories in our experiments.

In MIMIC-III we use one phenotype as the novel category and draw different distribution shifts on each repetition. Hence the size of the novel category does not vary much, and it is $\alpha = 0.075 \pm 0.002$. The raw values of the AU-ROC and AU-PRC can still change quite a lot, since different shifts entail different detection abilities for all the methods. Hence Table 2 gives

Algorithm	RAMPE: $\lvert 1-\hat{\alpha}/\alpha \rvert$		
	MPE method	MIMIC-III	Tabula Muris
Domain Disc.	EN	$\mathbf{0.28} \pm 0.18$	6.60 ± 5.52
	$FPR < 0.1$	0.55 ± 0.12	0.72 ± 0.86
Propensity	EN	0.62 ± 0.14	6.58 ± 5.45
	$FPR < 0.1$	0.54 ± 0.12	$\mathbf{0.50} \pm 0.57$
CoNoC		0.44 ± 0.11	0.76 ± 0.99

Figure 1: Average Relative Absolute Mixture Proportion Error ($\lvert 1-\hat{\alpha}/\alpha \rvert$) for evaluated methods, where $\hat{\alpha}$ is the estimated proportion and α is the true one. The estimator derived in Elkan and Noto [2008] under assumptions that do not hold in our setting of distribution shift, demonstrates unstable performance while thresholding FPR values seems to offer comparable performance on all methods.

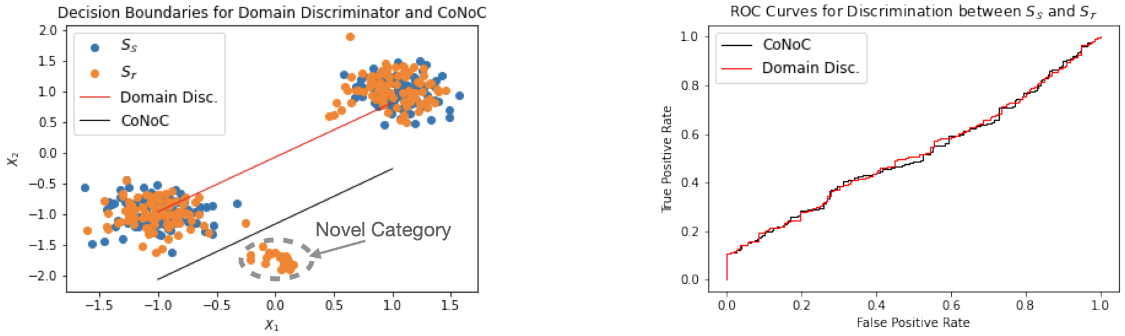


Figure 2: **(Left)** Toy example from Section 3.1, but without distribution shift. The learned models mostly differ in their bias terms, hence under an appropriate choice of the decision threshold (e.g. via the results of Elkan and Noto [2008]) both can detect the novel category successfully. Hence CoNoC performs on-par with unconstrained approaches. **(Right)** The ROC-Curves of the two classifiers coincide, emphasizing their equivalence in terms of ability to detect the novel category.

the details results for these runs. We observe that the performance of all methods changes in unison according to the drawn shifts (as explained earlier, some shifts entail more difficult problems than others), but in relative performance CoNoC performs best on most repetitions.

Effect of β on performance of CoNoC. We use the Tabula-Muris dataset to examine the effect of choosing different values of β in our procedure. To do that we take the training history from running Algorithm 1, and change the model selection in the last step of the algorithm to have a different value of β. Hence by varying the values of β we choose different models. Appendix B shows the relative AU-ROC and Av.-Precision as we vary β between low and high values. It is important to note that the problem here is separable in the sense that training classifiers with true label for the novel class achieves AU-ROC values around the range of 0.97 to 0.99. Hence low values of β are expected to produce favorable results, as may be confirmed by the figure. As we move towards larger values of β the metrics become more noisy and also comparable to the baseline (we do not show the propensity estimation baseline since it has inferior performance in this dataset). It also worth mentioning that increasing β only affects the performance of CoNoC, hence the change in relative performance is only due to variation in performance of our method. Our conclusion is that while the method is robust to the choice of β, large deviations from the ideal selection $\beta(h^*)$ will result in degraded performance.

References

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, Pieter Robberechts, and Jesse Davis. Beyond the selected completely at random assumption for learning from

Trial index	1	2	3	4	5	6	7	8
AU-ROC Domain Disc.	0.904	0.891	0.774	0.784	0.741	0.931	0.995	0.932
AU-ROC Propoensity	0.885	0.884	**0.790**	0.730	**0.834**	0.905	0.986	0.919
AU-ROC CoNoC	**0.914**	**0.947**	0.755	**0.860**	0.791	**0.958**	**0.996**	**0.978**
AU-PRC Domain Disc.	0.656	0.268	0.324	0.217	0.450	0.835	0.994	**0.858**
AU-PRC Propensity	0.234	0.088	0.152	0.058	0.211	0.572	0.900	0.458
AU-PRC CoNoC	**0.764**	**0.505**	**0.410**	**0.400**	**0.570**	**0.906**	**0.995**	0.854
α	0.0106	0.0181	0.0667	0.0459	0.0385	0.1527	0.2367	0.0517

Table 1: Raw AU-PRC and AU-ROC values and size of novel category, α, for all runs on the Tabula-Muris dataset. CoNoC performs best both in terms of AU-PRC for 5 out of 8 runs, other runs do not have a distinct winning method, though CoNoC performs best either in terms of AU-ROC or AU-PRC on all runs.

AU-ROC	1	2	3	4	5	6	7	8
Domain Disc.	0.757	0.825	0.837	0.753	0.812	0.841	0.763	0.740
Propoensity	0.771	0.832	0.840	**0.789**	0.850	0.821	0.795	0.777
CoNoC	**0.837**	**0.857**	**0.848**	0.787	**0.858**	**0.867**	**0.829**	**0.845**

AU-PRC	1	2	3	4	5	6	7	8
Domain Disc.	0.275	0.412	0.389	0.262	0.348	0.374	0.295	0.291
Propensity	0.299	0.401	0.389	**0.314**	0.395	0.367	0.340	0.324
CoNoC	**0.402**	**0.461**	**0.433**	0.295	**0.422**	**0.432**	**0.401**	**0.396**

AU-ROC	9	10	11	12	13	14	15
Domain Disc.	0.798	0.813	0.760	**0.819**	0.741	0.831	0.782
Propensity	0.819	**0.856**	0.803	0.755	0.767	0.838	0.798
CoNoC	**0.857**	0.855	**0.840**	0.817	**0.838**	**0.857**	**0.829**

AU-PRC	9	10	11	12	13	14	15
Domain Disc.	0.351	0.350	0.257	0.322	0.260	0.420	0.246
Propensity	0.402	0.411	0.314	0.270	0.297	0.424	0.261
CoNoC	**0.465**	**0.476**	**0.382**	**0.319**	**0.396**	**0.462**	**0.325**

Table 2: Raw AU-PRC and AU-ROC values for all runs on the MIMIC-III dataset. CoNoC performs best both in terms of AU-PRC for 5 out of 8 runs, other runs do not have a distinct winning method, though CoNoC performs best either in terms of AU-ROC or AU-PRC on all runs.

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.

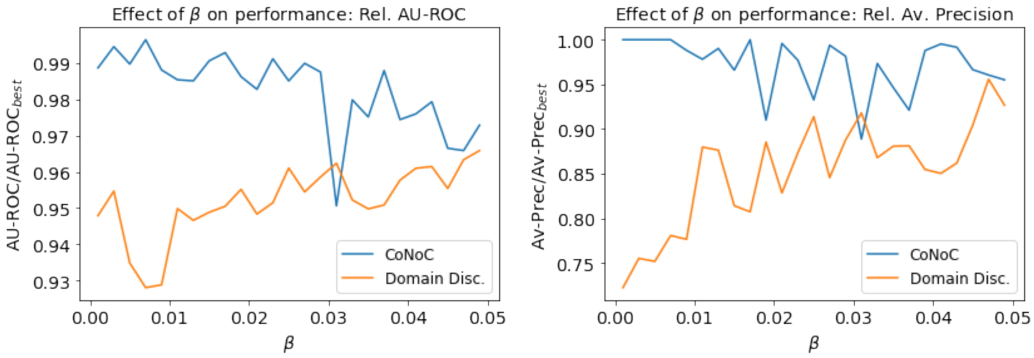
Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept learners for few-shot learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=eJIJF3-LoZO`.

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

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, 2019.

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.

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.



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.

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.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

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

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL `http://www.cs.mcgill.ca/%7Ecolt2009/papers/003.pdf#page=1`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.