Domain Adaptation Under Open Set Label Shift

Saurabh Garg ¹  Sivaraman Balakrishnan ¹  Zachary C. Lipton ¹

Abstract

We introduce the problem of domain adaptation under Open Set Label Shift (OSLS) where the label distribution can change arbitrarily and a new class may arrive during deployment, but the class-conditional distributions \( p(x|y) \) are domain-invariant. The learner’s goals here are two-fold: (a) estimate the target label distribution, including the novel class; and (b) learn a target classifier. First, we establish necessary and sufficient conditions for identifying these quantities. Second, we propose practical methods for both tasks. Unlike typical Open Set Domain Adaptation (OSDA) problems, which tend to be ill-posed and amenable only to heuristics, OSLS offers a well-posed problem amenable to more principled machinery. Experiments across numerous semi-synthetic benchmarks on vision, language, and medical datasets demonstrate that our methods consistently outperform OSDA baselines, achieving 10–25% improvements in target domain accuracy. Finally, we analyze the proposed methods, establishing finite-sample convergence to the true label marginal and convergence to an optimal classifier for linear models in a Gaussian setup.

1. Introduction

Suppose that we wished to deploy a machine learning system to recognize diagnoses based on their clinical manifestations. If the distribution of data were static over time, then we could rely on the standard machinery of statistical prediction. However, disease prevalences are constantly changing, violating the assumption of independent and identically distributed (iid) data. In such scenarios, we might reasonably apply the label shift assumption, where prevalences can change but clinical manifestations cannot. When only the relative proportion of previously seen diseases can change, principled methods can detect and correcting for label shift on the fly (Saerens et al., 2002; Zhang et al., 2013; Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2021; Garg et al., 2020). But what if a new disease, like COVID-19, were to arrive suddenly?

Traditional label shift adaptation techniques break when faced with a previously unseen class. A distinct literature on Open Set Domain Adaptation (OSDA) seeks to handle such cases (Panareda Busto and Gall, 2017; Baktashmotlagh et al., 2019; Tan et al., 2019; You et al., 2019; Saito et al., 2018; 2020; Fu et al., 2020). Given access to labeled source data and unlabeled target data, the goal in OSDA is to adapt classifiers in general settings where previous classes can shift in prevalence (and even appearance), and novel classes must be separated out from those previously seen. Most work on OSDA is driven by the creation of and progress on benchmark datasets (e.g., DomainNet, OfficeHome). Existing OSDA methods are heuristic in nature, addressing settings where the right answers seem intuitive but are not identified mathematically. However, absent assumptions on: (i) the nature of distribution shift among source classes and (ii) the relation between source classes and novel class, standard impossibility results for domain adaptation condemn us to guesswork (Ben-David et al., 2010).

In this work, we introduce domain adaptation under Open Set Label Shift (OSLS), a coherent instantiation of OSDA that applies the label shift assumption but allows for a new class to show up in the target distribution. Formally, the label distribution may shift between source and target \( p_s(y) \neq p_t(y) \), but the class-conditional distributions among previously seen classes may not (i.e., \( \forall y \in \{1, 2, \ldots, k\}, p_s(x|y) = p_t(x|y) \)). Moreover, a new class \( y = k + 1 \) may arrive in the target period. Notably, OSLS subsumes label shift (Saerens et al., 2002; Storkey, 2009) (when \( p_t(y = k + 1) = 0 \)) and learning from Positive and Unlabeled (PU) data (Letouzey et al., 2000; Elkan and Noto, 2008) (when \( k = 1 \)). As with label shift and PU learning, our goals are two-fold. Here, we must (i) estimate the target label distribution \( p_t(y) \) (including the novel class); (ii) train a \((k+1)\)-way target-domain classifier.

First, we characterize when the parameters of interest are identified (Sec. 2). Namely, we define a (necessary) weak positivity condition, which states that there exists a subset of each label’s support that has zero probability mass un-
We conduct extensive semi-synthetic experiments adapting thus motivated, we propose the PULSE framework, which
outperforms methods that naively solve k PU problems on
believing distributions, and model architectures, PULSE consis-
seven benchmark datasets spanning vision (CIFAR10, CI-
and CVIR techniques proposed in Garg et al. (2021b).

Our methods build on recent advances in label shift and
PU learning problems (Sec. 2). However, we
study to elucidate efficacy of CVIR over other consistent
objectives, nnPU (Kiryo et al., 2017) and uPU (Du Plessis
et al., 2015).

2. Open Set Label Shift

For a vector \( v \in \mathbb{R}^d \), we use \( v_j \) to denote its \( j \)-th entry, and for an event \( E \), we let \( I[E] \) denote the binary indicator of the event. Let \( \mathcal{X} \) be the input space and \( \mathcal{Y} = \{1, 2, \ldots, k+1\} \) be the output space for classification. Let \( P_s \) and \( P_t \) be the source and target distributions and let \( p_s \) and \( p_t \) denote the corresponding probability density (or mass) functions. \( \ell(z, y) \) denotes loss incurred by predicting \( z \) when the true label is \( y \). As in standard unsupervised domain adaptation, we are given independently and identically distributed (iid) samples from labeled source data \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \sim P_s^n \) and iid samples from unlabeled target data \( \{x_{n+1}, \ldots, x_{n+m}\} \sim P_t^m \). OSLS setting generalizes the label shift and PU learning setting (refer to Sec. C for preliminaries):

**Definition 1** (Open set label shift). Define \( \mathcal{Y}_i = \mathcal{Y} \) and as \( \mathcal{Y}_s = \mathcal{Y} \setminus \{k+1\} \). Under OSLS, the label distribution among source classes \( \mathcal{Y}_s \) may change but the class conditional \( p(x|y) \) for those classes remain invariant between source and target, and the target domain may contain a novel class, i.e., \( p_s(x|y = j) = p_t(x|y = j) \) \( \forall j \in \mathcal{Y}_s \) and \( p_s(y = k+1) \neq p_t(y = k+1) \).
While it may seem that existence of a subdomain $X_k$ (previously-seen) labels in the source distribution, i.e., for all $y \in \mathcal{Y}_s$, $p_s(y) \geq c$ for some $c > 0$.

Under OSLS, our goal naturally breaks down into two tasks: (i) estimate the target label marginal $p_t(y)$ (ii) train a classifier $f : \mathcal{X} \rightarrow \Delta^k$ to approximate $p_t(y|x)$.

**Identifiability of OSLS** We assume access to population distribution, i.e., $p_s(x, y)$ and $p_t(x)$. To keep the discussion simple, we assume finite input domain $\mathcal{X}$ which can then be relaxed to continuous inputs. We relegate extension to continuous cases and proofs to App. D.

Given access to $p_t(y)$, the class conditional $p_t(x|y = k + 1)$ can be obtained in closed form as $p_t(x) - \sum_{j=1}^{k} p_t(y = j)p_t(x|y = j)$. We can then simply apply Bayes rule to obtain $p_t(y|x)$. Henceforth, we will focus our discussion on identifiability of $p_t(y)$ which implies identifiability of $p_t(y|x)$. In following proposition, we present weak positivity, a necessary condition for $p_t(y)$ to be identifiable.

**Proposition 1 (Necessary conditions).** Assume $p_t(y) > 0$ for all $y \in \mathcal{Y}_t$. Then $p_t(y)$ is identified only if $p_t(x|y = k + 1)$ and $p_s(x|y)$ for all $y \in \mathcal{Y}_s$ satisfy weak positivity, i.e., there exists a subdomain $X_{wp} \subset \mathcal{X}$ such that: (i) $p_t(X_{wp}|y = k + 1) = 0$; and (ii) the matrix $[p_s(x|y)]_{x \in X_{wp}, y \in \mathcal{Y}_s}$ is full column-rank.

Intuitively, Proposition 1 states that if the target marginal doesn’t lie on the vertex of the simplex $\Delta^k$, then their must exist a subdomain $X_{wp}$ where the support of novel class is zero and within $X_{wp}$, $p_t(y)$ for source classes is identifiable. While it may seem that existence of a subdomain $X_{wp}$ is enough, we show that for the OSLS problem, existence doesn’t imply uniqueness (refer to App. D.1 for counter example). Next, we extend weak positivity to two stronger conditions, either of which (alone) implies identifiability.

**Proposition 2 (Sufficient conditions).** The target marginal $p_t(y)$ is identified if for all $y \in \mathcal{Y}_t \setminus \{k + 1\}$, $p_t(x|y = k + 1)$ and $p_s(x|y)$ satisfy either: (i) Strong positivity, i.e., there exists $X_{wp} \subset \mathcal{X}$ such that $p_t(X_{wp}|y = k + 1) = 0$ and the matrix $[p_s(x|y)]_{x \in X_{wp}, y \in \mathcal{Y}_s}$ is full-rank and diagonal; or (ii) Separability, i.e., there exists $X_{sep} \subset \mathcal{X}$, such that $p_t(X_{wp}|y = k + 1) = 0$, $p_s(X_{sep}) = 1$, and the matrix $[p_s(x|y)]_{x \in X_{sep}, y \in \mathcal{Y}_s}$ is full-column-rank.

Informally, strong positivity extends weak positivity by making an additional assumption that the matrix formed by $p(x|y)$ on inputs in $X_{wp}$ is diagonal and the separability assumption extends the weak positivity condition to the full input domain of source classes instead of just $X_{wp}$.

**Reduction of OSLS to $k$ PU Problems** Under the strong positivity condition, the OSLS problem can be broken down into $k$ PU problems as follows: By treating a given source class $y_j \in \mathcal{Y}_s$ as positive and grouping all other classes together as negative we observe that the unlabeled target data is then a mixture of data from the positive and negative classes. This yields a PU learning problem and the corresponding mixture proportion gives the fraction $\alpha_j$ of class $y_j$ among the target data. By iterating this process for all source classes, we can solve for the entire target label marginal $p_t(y)$. Thus, OSLS reduces to $k$ instances of PU learning problem. Formally, note that $p_t(x)$ can be written as:

$$p_t(x) = \sum_{y \in \mathcal{Y}_s} p_t(y = j)p_s(x|y = j) + (1 - p_t(y = j)) \left( \sum_{y \in \mathcal{Y}_s \setminus \{j\}} \frac{p_s(y = i)}{1 - p_t(y = j)} p_s(x|y = i) \right),$$

individually for all $j \in \mathcal{Y}_s$. By repeating this reduction for all classes, we obtain $k$ separate PU learning problems. Hence, a natural choice is to leverage this structure and solve $k$ PU problems to solve the original OSLS problem. We explain this procedure more formally in App. C.1. However, we observe that the naive reduction doesn’t scale to datasets with large number of classes because of error accumulation in each of the $k$ MPEs and $k$ one-versus-other PU classifiers.

**3. The PULSE Framework for OSLS**

Rather than dividing an OSLS instance into $k$ PU problems, we exploit the joint structure of the problem. We note that if only we could apply a label shift correction to source, i.e., re-sample source classes according to their relative proportion in the target data, then we could subsequently consider the unlabeled target data as a mixture of (i) the (rewighted) source distribution (say $p_s'(t)$); and (ii) the novel class distribution ($p_t(x|y = k + 1)$) obtaining a single PU learning problem. Formally, we have

$$p_t(x) = \sum_{j \in \mathcal{Y}_s} p_t(y = j)p_s(x|y = j)$$

$$= \sum_{j \in \mathcal{Y}_s} \frac{p_t(y = j)}{p_s(y = j)} p_s(x, y = j)$$

$$+ p_t(x|y = k + 1)p_s(y = k + 1)$$

$$= (1 - p_t(y = k + 1))p_s'(x) + p_t(y = k + 1)p_t(x|y = k + 1),$$

where $p_s'(x)$ is the label-shift-corrected source distribution, i.e., $p_s'(x) = \sum_{j \in \mathcal{Y}_s} w(j)p_s(x, y = j)$, where $w(j) = (p_t(y = j)\alpha_j)/p_s(y = j)$ for all $j \in \mathcal{Y}_s$. We describe this formally in Sec. E.

To estimate label shift among source classes, we build on recently proposed Best Bin Estimation (BBE) technique from Garg et al. (2021b). We tailor BBE to estimate the relative fraction of previously seen classes in the target distribution by exploiting a k-way source classifier $f_s$ trained on labeled source data. We describe the procedure in App. E (Algorithm 2). After estimating the fraction of source classes in target, we re-sample the source data to mimic samples from
We compare PULSE with several popular OSDA methods, which was shown to outperform alternative approaches. We design experiments with different fraction of novel class prevalence, classes and the novel class. For target marginal, we report $\tau$ and classifier $f'_s$ with importance-reweighted label shift correction. Our approach is summarized in Algorithm 1.

### Algorithm 1 Positive and Unlabeled learning post Label Shift Estimation (PULSE) framework

**Input**: Labeled source data $\{X^S, y^S\}$ and unlabeled target samples $X^T$.

1. Randomly split data into training $\{X^S_1, y^S_1\}, X^T_1$ and hold out partition $\{X^S_2, y^S_2\}, X^T_2$. 
2. Train a classifier $f_s$ on labeled source data $\{X^S, y^S\}$.
3. Estimate label shift $\hat{\gamma}_l(y = j) = \frac{\hat{\mu}_j(y = j)}{\sum_{k \in Y_s} \hat{\mu}_k(y = k)}$ and hence importance ratios $\hat{w}(j)$ among source classes $j \in Y_s$ using Algorithm 2.
4. Re-sample training source data according to label distribution $\hat{\gamma}_l$ to get $\{X^S_1, \hat{y}^S_1\}$ and $\{X^S_2, \hat{y}^S_2\}$.
5. Using Algorithm 3, train a discriminator $f_d$ and estimate novel class fraction $\hat{\mu}_k(y = k + 1)$.
6. Assign $[f_s(x)]_j = (f_d(x)) \frac{\hat{w}(j) \cdot [f_s(x)]}{\sum_{k \in Y_s} \hat{w}(k) \cdot [f_d(x)]}$ for all $j \in Y_s$ and $[f_s(x)]_{k+1} = 1 - f_d(x)$. And for all $j \notin Y_s$, assign $\hat{\mu}_k(y = j) = (1 - \hat{\mu}_k(y = k + 1)) \cdot \hat{\gamma}_l(y = j)$.

**Output**: Marginal estimate $\hat{\mu}_k \in \Delta^k$ and classifier $f_s(x)$.

Distribution $p'_u(x)$. Thus, obtaining a PU learning problem instance, we resort to PU learning techniques to (i) estimate the fraction of novel class $p(y = k + 1)$; and (ii) learn a binary classifier $f_d(x)$ to discriminate between label shift corrected source $p'_u(x)$ and novel class $p_u(x | y = k + 1)$.

For PU learning, we employ Conditional Value Ignoring Risk (CVIR) loss proposed in Garg et al. (2021b) which was shown to outperform alternative approaches. We discuss this procedure in detail in App. E.

Finally, to obtain a $(k + 1)$-way classifier $f_t(x)$ on target we combine discriminator $f_d$ and source classifier $f_s$ with importance-reweighted label shift correction. Our approach is summarized in Algorithm 1.

### 4. Experiments

Here, we present key experimental findings and present more details in Sec. A. We conduct experiments with seven benchmark classification datasets. For each dataset, we simulate an OSLS problem as described in next paragraph. We compare PULSE with several popular OSDA methods: DANCE (Saito et al., 2020), UAN (You et al., 2019), CMU (Fu et al., 2020), STA (Liu et al., 2019), Backprop-ODA (or BODA) (Saito et al., 2018). For alternative baselines, we experiment with (i) domain discriminator classifier for source versus target; and (ii) $k$ PU classifiers. To evaluate classifiers, we report target accuracy on all classes, seen classes and the novel class. For target marginal, we report absolute difference between true and estimated marginal.

#### OSLS Setup

To simulate an OSLS problem, we experiment with different fraction of novel class prevalence, source marginal, and target marginal. We randomly choose classes that constitute the novel target class. Then, we split the training data from each source class randomly into two partitions creating a random label distribution for shared classes among source and target. We then club novel classes to assign them a new class (i.e., $k + 1$). Finally, we throw away target labels to obtain an unsupervised DA problem.

Across different datasets, we observe that PULSE consistently outperforms other methods for the target classification and novel prevalence estimation (Table 1). We observe that with default hyperparameters, popular OSDA methods significantly under perform as compared to PULSE. We show that the primary reasons underlying the poor performance of OSDA methods are (i) the heuristics employed to detect novel classes; and (ii) loss functions incorporated to improve alignment between examples from common classes in source and target. In contrast, our two-stage method PULSE, first estimates the fraction of novel class which then guides the classification of novel class versus previously seen classes avoiding the need to guess $\kappa$.

#### Ablations

We perform ablations with novel class proportion (App. H.6), contrastive pre-training (App. H.7), alternative PU learning plugin methods (App. H.8) and show that PULSE continues to dominate other methods.

### 5. Theoretical Analysis of PULSE Framework

We analyse key steps of our PULSE procedure for target label marginal estimation (Step 3, 5 Algorithm 1) and learning the domain discriminator classifier (Step 5, Algorithm 1). We relegate formal statements and proofs to App. F.

#### Target marginal estimation

Building on BBE results from Garg et al. (2021b), we present finite sample results for target label marginal estimation.

**Theorem 1** (Informal). Assume that for each class $y \in Y_s$, there exists a threshold $c_y$ such that for the classifier $f_s$, if $[f_s(x)]_y > c_y$ for any $x$ then the true label for that sample $x$ is $y$. Then, we have $|\hat{\mu}_t - p_t|_1 \leq O\left(\sqrt{\frac{\log(4k/\delta)}{n}} + \sqrt{\frac{\log(4k/\delta)}{m}}\right)$.

#### Analysis of CVIR in population

While the CVIR loss was proposed in Garg et al. (2021b), no analysis was provided for the iterative gradient descent procedure. In our work, we show that in population on a separable Gaussian dataset, CVIR will recover the optimal classifier.

We consider a binary classification problem where we have access to positive distribution (i.e., $p_p$), unlabeled distribution (i.e., $p_u := \alpha p_p + (1 - \alpha)p_n$), and mixture coefficient $\alpha$. First we introduce some notation. For a classifier $f$ and loss function $\ell$, define $\text{VIR}_\alpha(f) = \inf\{r \in \mathbb{R} : P_{x \sim p_u}(\ell(x, -1; f) \leq r) \geq 1 - \alpha\}$. Intuitively, $\text{VIR}_\alpha(f)$ identifies a threshold $\tau$ to capture bottom $1 - \alpha$ fraction of
Table 1: Comparison of PULSE with other methods. Across all datasets, PULSE outperforms alternatives for both target classification and novel class prevalence estimation. Acc (All) is target accuracy, Acc (Seen) is target accuracy on examples from previously seen classes, and Acc (Novel) is recall for novel examples. MPE (Novel) is absolute error for novel prevalence estimation. Results reported by averaging across 3 seeds. Detailed results for each dataset with all methods in App. H.4.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc (All)</td>
<td>Acc (Seen)</td>
</tr>
<tr>
<td>Source-Only</td>
<td>67.1</td>
<td>87.0</td>
</tr>
<tr>
<td>UAN (You et al., 2019)</td>
<td>15.4</td>
<td>19.7</td>
</tr>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>63.1</td>
<td>66.2</td>
</tr>
<tr>
<td>DANCE (Saito et al., 2020)</td>
<td>70.4</td>
<td>85.5</td>
</tr>
<tr>
<td>STA (Liu et al., 2019)</td>
<td>57.9</td>
<td>69.6</td>
</tr>
<tr>
<td>CMU (Fu et al., 2020)</td>
<td>62.1</td>
<td>77.9</td>
</tr>
<tr>
<td>Domain Disc. (Elkan and Noto, 2008)</td>
<td>47.4</td>
<td>87.0</td>
</tr>
<tr>
<td>k-PU</td>
<td>83.6</td>
<td>79.4</td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>86.1</td>
<td>91.8</td>
</tr>
</tbody>
</table>

the loss \( \ell(x, -1) \) for points \( x \) sampled from \( p_u \). Additionally, define CVIR loss as \( \mathcal{L}(f, w) = \alpha \mathbb{E}_{p_u} [\ell(x, 1; f)] + \mathbb{E}_{p_w} [\ell(x, -1; f)] \) for classifier \( f \) and some weights \( w(x) \in \{0, 1\} \). Recall that given a classifier \( f_t \) at an iterate \( t \), CVIR procedure proceeds as follows:

\[
\begin{align*}
\omega_t(x) &= I[\ell(x, -1; f_t)] < \text{VIR}_t(f_t), \\
f_{t+1} &= f_t - \eta \nabla \mathcal{L}_f(f_t, \omega_t) .
\end{align*}
\]

We assume a data generating setup where \( x \) are drawn from a multivariate Gaussian, i.e., \( x \sim \mathcal{N}(0, I_d) \) and labels are assigned by a random separator \( \theta_{opt} \), i.e., \( y = I[\theta_{opt}^T x > 0] \) for some fixed direction \( \theta_{opt} \in \mathbb{R}^d \).

**Theorem 2 (Informal).** *In the data setup described above, a linear classifier \( f(x; \theta) = I[\theta^T x > 0] \) trained with CVIR procedure as in equations (1)-(2) will converge to an optimal positive versus negative classifier.*

**Empirical investigation in overparameterized models**

While the above analysis highlights consistency of CVIR procedure, it doesn’t capture the observed empirical efficacy of CVIR over alternative methods in overparameterized models. With experiments in the Gaussian setup described above, we investigate efficacy of CVIR in overparameterized models. We observe that early learning phenomena followed by the self-training procedure crucially helps CVIR.

6. Conclusion

In this work, we introduce OSLS a well-posed instantiation of OSDA that subsumes label shift and PU learning into a framework for learning adaptive classifiers. We presented identifiability conditions for OSLS and proposed PULSE, a simple and effective approach to tackle the OSLS problem. Moreover, our extensive experiments demonstrate efficacy of PULSE over popular OSDA alternatives when the OSLS assumptions are met. We would like to highlight the brittle nature of benchmark driven progress in OSDA and hope that our work can help to stimulate more solid foundations and enable systematic progress in this area. In future work, we hope to bridge the gap between the necessary and sufficient identifiability conditions. While we empirically investigate reasons for CVIR’s efficacy in overparameterized models, we aim to extend our theory to overparameterized settings in future. Finally, we hope that our open source code and benchmarks will foster further progress on OSLS.
Domain Adaptation Under Open Set Label Shift

References


J. Bekker and J. Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In Association for the Advancement of Artificial Intelligence (AAAI), 2018.


C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In International Conference Knowledge Discovery and Data Mining (KDD), pages 213–220, 2008.


Domain Adaptation Under Open Set Label Shift


Supplementary Materials

A. Summary of Experiments

**Baselines** We compare PULSE with several popular methods from OSDA literature. While these methods are not specifically proposed for OSLS, they are introduced for the more general OSDA problem. In particular, we make comparisons with DANCE (Saito et al., 2020), UAN (You et al., 2019), CMU (Fu et al., 2020), STA (Liu et al., 2019), Backprop-ODA (or BODA) (Saito et al., 2018). We use the open source implementation available at https://github.com/thuml/PU classifiers. For alternative baselines, we experiment with source classifier directly deployed on the target data which may contain novel class and label shift among source classes (referred to as source-only). We also train a domain discriminator classifier for source versus target (referred to as domain disc.). This is adaptation of PU learning baseline (Elkan and Noto, 2008) which assumes no label shift among source classes. Finally, per the reduction presented in Sec. 2, we train k PU classifiers (referred to as k-PU). We include detailed description of each method in App. H.1.

**Datasets** We conduct experiments with seven benchmark classification datasets across vision, natural language, biology and medicine. For each dataset, we simulate an OSLS problem as described in next paragraph. For vision, we use CIFAR10, CIFAR100 (Krizhevsky and Hinton, 2009) and Entity30 (Santurkar et al., 2021). For language, we experiment with Newsgroups-20 (http://qwone.com/~jason/20Newsgroups/) dataset. Additionally, inspired by applications of OSLS in biology and medicine, we experiment with Tabula Muris (Consortium et al., 2020) (Gene Ontology prediction), Dermnet (skin disease prediction https://dermnetnz.org/), and BreakHis (Spanhol et al., 2015) (tumor cell classification). These datasets span language, image and table modalities. We provide interpretation of OSLS problem for each dataset along with other details in App. H.2.

**OSLS Setup** To simulate an OSLS problem, we experiment with different fraction of novel class prevalence, source label distribution, and target label distribution. We randomly choose classes that constitute the novel target class. After randomly choosing source and novel classes, we first split the training data from each source class randomly into two partitions. This creates a random label distribution for shared classes among source and target. We then club novel classes to assign them a new class (i.e. \( k + 1 \)). Finally, we throw away labels for the target data to obtain an unsupervised DA problem. We repeat the same process on iid hold out data to obtain validation data with no target labels.

**Training and Evaluation** We use Resnet18 (He et al., 2016) for CIFAR10, CIFAR100, and Entity30. For newsgroups, we use a convolutional architecture. For Tabular Muris and MNIST, we use a fully connected MLP. For Dermnet and BreakHis, we use ResNet-50. For all methods, we use the same backbone for discriminator and source classifier. For kPU, we use a separate final layer for each class with the same backbone. We use default hyperparameters for all methods. For OSDA methods, we use default method specific hyperparameters introduced in their works. Since OSDA methods do not estimate the prevalence of novel class explicitly, we use the fraction of examples predicted in class \( k + 1 \) as a surrogate. We train models till the performance on validation source data (labeled) ceases to increase. Unlike OSDA methods, note that we do not use early stopping based on performance on held-out labeled target data. To evaluate classification performance, we report target accuracy on all classes, seen classes and the novel class. For novel class prevalence estimation, we report absolute difference between true and estimated marginal. We open-source our code and by simply changing a single config file, new OSLS setups can be generated and experimented with. We provide precise details about hyperparameters, OSLS setup for each dataset and code in App. H.3.

**Results** Across different datasets, we observe that PULSE consistently outperforms other methods for the target classification and novel prevalence estimation (Table 1). For detection of novel classes (Acc (Novel) column), kPU achieves superior performance as compared to alternative approaches because of its bias to default to \((k + 1)\)th class. This is evident by the sharp decrease in performance on previously seen classes. For each dataset, we plot evolution of performance with training in App. H.4. We observe more stability in performance of PULSE as compared to other methods.

We observe that with default hyperparameters, popular OSDA methods significantly under perform as compared to PULSE. We hypothesize that the primary reasons underlying the poor performance of OSDA methods are (i) the heuristics employed to detect novel classes; and (ii) loss functions incorporated to improve alignment between examples from common classes in source and target. To detect novel classes, a standard heuristic employed popular OSDA methods involves thresholding uncertainty estimates (e.g., prediction entropy, softmax confidence (You et al., 2019; Fu et al., 2020; Saito et al., 2020)) at a predefined threshold \( \kappa \). However, a fixed \( \kappa \), may not for different datasets and different fractions of the novel class. In App. H.5, we ablate by (i) removing loss function terms incorporated with an aim to improve source target alignment; and (ii) vary threshold \( \kappa \) and show improvements in performance of these methods. In contrast, our two-stage method PULSE,
Ablations Different datasets, in our setup span different fraction of novel class prevalence ranging from 0.22 (in CIFAR10) to 0.64 (in Tabula Muris). For each dataset, we perform more ablations on the novel class proportion in App. H.6. For kPU and PULSE, in the main paper, we include results with BBE and CVIR (Garg et al., 2021b). In App. H.8, we perform experiments with alternative PU learning approaches and highlight the superiority of BBE and CVIR over other methods. Moreover, since we have access to unlabeled target data, we experiment with SimCLR (Chen et al., 2020a) pre-training on the mixture of unlabeled source and target dataset. We include setup details and results in App. H.7. While pre-trained backbone architecture improves performance for all methods, PULSE continues to dominate other methods.

B. Related Work

(Closed Set) Domain Adaptation (DA) Under DA, our goal is to adapt a predictor from a source distribution with labeled data to a target distribution from which we observe only unlabeled examples. DA is classically explored under two distribution shift scenarios (Storkey, 2009): (i) Covariate shift (Zhang et al., 2013; Zadrozny, 2004; Cortes et al., 2010; Cortes and Mohri, 2014; Gretton et al., 2009) where \( p(y|x) \) remains invariant among source and target; and (ii) Label shift (Saerens et al., 2002; Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2021; Garg et al., 2020; Zhang et al., 2021) where \( p(x|y) \) is shared across source and target. In these settings most theoretical analysis requires that the target distribution’s support is a subset of the source support (Ben-David et al., 2010). However, recent empirically work in DA (Long et al., 2015; 2017; Sun and Saenko, 2016; Sun et al., 2017; Zhang et al., 2019; 2018; Ganin et al., 2016; Sohn et al., 2020) focuses on settings motivated by benchmark datasets (e.g., WILDS (Sagawa et al., 2021; Koh et al., 2021), Office-31 (Saenko et al., 2010) OfficeHome (Venkateswaran et al., 2017), DomainNet (Peng et al., 2019)) where such overlap assumptions are violated. Instead, they rely on some intuitive notion of semantic equivalence across domains. These problems are not well-specified and in practice, despite careful hyperparameter tuning, these methods often do not improve over standard empirical risk minimization on source data alone for practical, and importantly, previously unseen datasets (Sagawa et al., 2021).

Open Set Domain Adaptation (OSDA) OSDA (Panareda Busto and Gall, 2017; Bendale and Boult, 2015; Scheirer et al., 2013) extends DA to settings where along with distribution shift among previously seen classes, we may observe a novel class in the target data. This setting is also known as universal domain adaptation (You et al., 2019; Saito et al., 2020). Rather than making precise assumptions about the nature of shift between source and target, the OSDA literature is primarily governed by semi-synthetic problems on benchmark DA datasets (e.g. DomainNet, Office-31 and OfficeHome). Numerous OSDA methods have been proposed (Baktashmotlagh et al., 2019; Cao et al., 2019; Tan et al., 2019; Lian et al., 2019; You et al., 2019; Saito et al., 2018; 2020; Fu et al., 2020; Bucci et al., 2020). At a high level, most OSDA methods perform two steps: (i) align source and target representation for previously seen classes; and (ii) train a discrimination to reject novel class from previously seen classes. The second step typically uses novelty detection heuristics to identify novel samples.

PU Learning Positive and Unlabeled (PU) learning is the base case of OSLS. Here, we observe labeled data a single source class and unlabeled target data contains data from both the novel class and the source class. In PU learning, our goals are: (i) Mixture Proportion Estimation (MPE), i.e., determining the fraction of previously seen class in target ; and (ii) PU classification, i.e., learning to discriminate between the novel and the positive (source) class. Several classical methods were proposed for both MPE (Elkan and Noto, 2008; Du Plessis and Sugiyama, 2014; Scott, 2015; Jain et al., 2016; Bekker and Davis, 2018; 2020) and classification (Elkan and Noto, 2008; Du Plessis et al., 2014; 2015). However, classical MPE methods do not scale to high-dimensional settings (Ramaswamy et al., 2016). More recent methods alleviate these issues by operating in classifier output space (Garg et al., 2021b; Ivanov, 2019). For classification, traditional methods fail when deployed with models classes with high capacity due to their capacity of fitting random labels (Zhang et al., 2017). Recent methods (Garg et al., 2021b; Kiryo et al., 2017; Chen et al., 2020b), avoid over-fitting by employing regularization or self-training techniques.

Other related work A separate line of work looks at the problem of Out-Of-Distribution (OOD) detection (Hendrycks and Gimpel, 2017; Geifman and El-Yaniv, 2017; Lakshminarayanan et al., 2016; Jiang et al., 2018; Ovadia et al., 2019; Zhang et al., 2020). Here, they identify novel examples, i.e., samples that lie out of the support of training distribution. The main different between OOD detection and OSDA is that in OOD detection we do not have access to unlabeled data containing a novel class. Recently, Cao et al. (2022) proposed open-world semi-supervised learning, where the task is to not only identify novel classes in target but also to separate out different novel classes in an unsupervised manner.

Our work takes a step back from the hopelessly general OSDA setup, introducing OSLS, a well-posed OSDA setting where
After obtaining an estimate for mixture proportion discuss recent methods for MPE that operate in the classifier output space to avoid curse of dimensionality:

Note that given access to population of positives and unlabeled, determining the fraction of positives in unlabeled examples, PU learning the target domain (Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2021).

One common approach to label shift involves estimating the importance ratios \( p_t(y)/p_u(y) \) by leveraging a blackbox classifier and then employing re-sampling of source data or importance re-weighted ERM on source to obtain a classifier for the target domain (Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2021).

PU learning Under PU learning, we possess labeled source data from a positive class \((p_p)\) and unlabeled target data from \(p_u = \alpha p_p + (1 - \alpha)p_n\), a mixture of positive and negative class \((p_n)\). Our goals naturally break down in to two tasks: (i) MPE, determining the fraction of positives \(p_p\) in \(p_u\) and (ii) PU classification, learning a positive-versus-negative classifier on target.

Note that given access to population of positives and unlabeled, \(\alpha\) can be estimated as \(\min_x p_u(x)/p_p(x)\). Next, we briefly discuss recent methods for MPE that operate in the classifier output space to avoid curse of dimensionality:

(i) **EN**: Given a domain discriminator classifier \(f_d\) trained to discriminate between positive and unlabeled, Elkan and Noto (2008) proposed the following estimator:

\[
\sum_{x \in X_p} f_d(x)/\sum_{x \in X_u} f_d(x) \quad \text{where } X_p \text{ is the set of positive examples and } X_u \text{ is the set of unlabeled examples.}
\]

(ii) **DEDPUL**: Given a domain discriminator classifier \(f_d\), Ivanov (2019) proposed an estimator that leverages density of the data in the output space of the classifier \(f_d\) to directly estimate \(\min p_u(f(x))/p_p(f(x))\).

(iii) **BBE**: BBE (Garg et al., 2021b) identifies a threshold on probability scores assigned by the classifier \(f_d\) such that by estimating the ratio between the fractions of positive and unlabeled points receiving scores above the threshold, we obtain proportion of positives in unlabeled.

After obtaining an estimate for mixture proportion \(\alpha\), following methods can be employed for PU classification:

(i) **Domain Discriminator**: Given positive and unlabeled data, Elkan and Noto (2008) trained a classifier \(f_d\) to discriminate between them. To make a prediction on test point from unlabeled data, we can then use Bayes rule to obtain the following transformation on probabilistic output of the domain discriminator:

\[
f = \alpha \left( \frac{m}{n} \right) \left( \frac{f_d(x)}{1 - f_d(x)} \right), \quad \text{where } n \text{ and } m \text{ are the number of positives and unlabeled examples used to train } f_d \text{ (Elkan and Noto, 2008).}
\]

(ii) **uPU**: Du Plessis et al. (2015) proposed an unbiased loss estimator for positive versus negative training. In particular, since \(p_u = \alpha p_p + (1 - \alpha)p_n\), the loss on negative examples \(E_{p_u}[\ell(f(x); -1)]\) can be estimated as:

\[
E_{p_u}[\ell(f(x); -1)] = \frac{1}{1 - \alpha} \left[ E_{p_u}[\ell(f(x); -1)] - \alpha E_{p_p}[\ell(f(x); -1)] \right].
\]

Thus, a classifier can be trained with the following uPU loss:

\[
\mathcal{L}_{uPU}(f) = \alpha E_{p_p}[\ell(f(x); +1)] + E_{p_u}[\ell(f(x); -1)] - \alpha E_{p_u}[\ell(f(x); -1)].
\]

(iii) **nnPU**: While unbiased losses exist that estimate the PvN loss given PU data and the mixture proportion \(\alpha\), this unbiasedness only holds before the loss is optimized, and becomes ineffective with powerful deep learning models capable of memorization. Kiryo et al. (2017) proposed the following non-negative regularization for unbiased PU learning:

\[
\mathcal{L}_{nnPU}(f) = \alpha E_{p_p}[\ell(f(x); +1)] + \max \{ E_{p_u}[\ell(f(x); -1)] - \alpha E_{p_u}[\ell(f(x); -1)], 0 \}.
\]
We now introduce conditions for OSLS, under which the solution is identifiable. For ease, we re-state Proposition 1 and Proposition 2.

Proposition 1 (Necessary conditions). Assume \( p_t(y) > 0 \) for all \( y \in \mathcal{Y}_t \). Then \( p_t(y) \) is identified only if \( p_t(x|y = k + 1) \) and \( p_s(x|y) \) for all \( y \in \mathcal{Y}_s \) satisfy weak positivity, i.e., there must exists a subdomain \( X_{up} \subset X \) such that: (i) \( p_t(X_{up}|y = k + 1) = 0 \); and (ii) the matrix \( [p_s(x|y)]_{x \in X_{up}, y \in \mathcal{Y}_s} \) is full column-rank.

Proof. We prove this by contradiction. Assume that there exists a unique solution \( p_t(y) \). We will obtain contradiction when both (i) and (ii) don’t hold.
First, assume for no subset $X_{wp} \subseteq X$, we have $[p_s(x|y)]_{x \in X_{wp}, y \in Y_s}$ as full-rank. Then in that case, we have vectors $[p_s(x|y = j)]_{x \in X}$ as linearly dependent for $j \in Y_s$, i.e., there exists $[\alpha_j]_{j \in Y_s} \in \mathbb{R}^k$ such that $\sum_j \alpha_j p_s(x|y = j) = 0$ for all $x \in X$. Thus for small enough $\epsilon > 0$, we have infinite solutions of the form $[p_t(y = j) - \epsilon \cdot \alpha_j]_{j \in Y_s}$.

Hence, there exists $X_{wp} \subseteq X$ for which we have $[p_s(x|y)]_{x \in X_{wp}, y \in Y_s}$ as full-rank. Without loss of generality, we assume that $|X_{wp}| = k$. Assume that $p_t(X_{wp}|y = k + 1) > 0$, i.e., $[p_t(x|y = k + 1)]_{x \in X_{wp}}$ has $l < k$ zero entries. We will now construct another solution for the label marginal $p_t$. For simplicity we denote $A = [p_s(x|y)]_{x \in X_{wp}, y \in Y_s}$. Consider the vector $v(\gamma) = [p_t(x) - (p_t(y = k + 1) - \gamma)p_t(x|y = k + 1)]_{x \in X_{wp}}$ for some $\gamma > 0$. Intuitively, when $\gamma = 0$, we have $u = A^{-1}v(0)$ where $u = [p_t(y)|y \in Y_s]$, i.e., we recover the true label marginal corresponding to source classes.

However, since the solution is not at vertex, there exists a small enough $\gamma > 0$ such that $u' = A^{-1}v(\gamma)$ with $\sum_j u'_j \leq 1$ and $u'_j \geq 0$. Since $A$ is full-rank and $v(\gamma) \neq v(0)$, we have $u' \neq u$. Thus we construct a separate solution with $u'$ as $[p_t(y)|y \in Y_s]$, and $p_t(x) - \sum_{j \in Y_s} u'_j p_s(x|y = j)$ as $p_t(x|y = k + 1)$. Hence, when there exists $X_{wp} \subseteq X$ for which we have $[p_s(x|y)]_{x \in X_{wp}, y \in Y_s}$ as full-rank, for uniqueness we obtain a contradiction on the assumption $p_t(X_{wp}|y = k + 1) > 0$. 

We now make some comments on the assumption $p_t(y) > 0$ for all $y \in Y_1$ in Proposition 1. Since, $p_t(y)$ needs to satisfy simplex constraints, if the solution is at a vertex of simplex, then OSLS problem may not require weak positivity. For example, there exists contrived scenarios where $p_t(x|y = j) = p_s(x|y = k)$ for all $j \in Y_s$ and $p_t(x|y = k + 1) \neq p_s(x|y = j)$ for all $j \in Y_s$. Then when $p_t(x) = p_t(x|y = k + 1)$, we can uniquely identify the OSLS solution even when weak positivity assumption is not satisfied.

**Proposition 2** (Sufficient conditions). The target marginal $p_t(y)$ is identified if for all $y \in Y \setminus \{k + 1\}$, $p_t(x|y = k + 1)$ and $p_s(x|y)$ satisfy either: (i) Strong positivity, i.e., there exists $X_{sp} \subset X$ such that $p_t(X_{sp}|y = k + 1) = 0$ and the matrix $[p_s(x|y)]_{x \in X_{sp}, y \in Y_s}$ is full-rank and diagonal; or (ii) Separability, i.e., there exists $X_{sep} \subset X$, such that $p_t(X_{sep}|y = k + 1) = 0$, $p_s(X_{sep}) = 1$, and the matrix $[p_s(x|y)]_{x \in X_{sep}, y \in Y_s}$ is full column-rank.

**Proof.** For each condition, we will prove identifiability by constructing the unique solution.

Under strong positivity, for all $j \in Y_s$ there exists $x \in X_{sp}$ such that $p_t(x|y = k) = 0$ for all $k \in Y \setminus \{j\}$. Set $\alpha_j = \min_{x \in X_{sp}, p_s(x|y = j) > 0} \frac{p_t(x)}{p_s(x|y = j)}$, for all $j \in Y_s$. For $x \in X_{wp}$ such that $p_t(x|y = k) = 0$ for all $k \in Y \setminus \{j\}$, we get $\frac{p_t(x)}{p_s(x|y = j)} = p_t(y = j)$ and for all $x' \neq x$, we have $\frac{p_t(x')}{p_s(x'|y = j)} \geq p_t(y = j)$. Thus, we get $\alpha_j = p_t(y = j)$. Finally, we get $\alpha_{k+1} = 1 - \sum_{j \in Y_s} \alpha_j$. Plugging in values of the label marginal, we can obtain $p_t(x|y = k + 1) = p_t(x) - \sum_{y \in Y_s} p_t(y = j)p_s(x|y = j)$.

Under separability, we can obtain the label marginal $p_t$ for source classes by simply considering the set $X_{sep}$. Denote $A = [p_s(x|y)]_{x \in X_{sep}, y \in Y_s}$ and $v = [p_t(x)]_{x \in X_{sep}}$. Then, since $A$ is full-column-rank by assumption, we can define $u = (A^T A)^{-1}A^T v$. For all $x \in X_{sep}$, we have $p_t(x) = \sum_{y \in Y_s} p_t(y)p_s(x|y)$ and hence, $u = [p_t(y)|y \in Y_s]$. Having obtained $[p_t(y)|y \in Y_s]$, we recover $p_t(y = k + 1) = 1 - \sum_{j \in Y_s} p_t(y = j)$ and $p_t(x|y = k + 1) = p_t(x) - \sum_{j \in Y_s} p_t(y = j)p_s(x|y = j)$.

While the two conditions in Proposition 2 overlap, they cover independent set of OSLS problems. Informally, strong positivity extends weak positivity by making an additional assumption that the matrix formed by $p(x|y)$ on inputs in $X_{wp}$ is diagonal and the separability assumption extends the weak positivity condition to the full input domain of source classes instead of just $X_{wp}$. Both of these conditions identify a support region of $X$ which purely belongs to source classes where we can either independently estimate the proportion of each source classes (i.e., under strong positivity) or jointly estimate the proportion (i.e., under separability).

**Extension to continuous inputs** To extend our identifiability conditions for continuous distributions, the linear independence conditions on the matrix $[p_s(x|y)]_{x \in X_{sep}, y \in Y_s}$ has the undesirable property of being sensitive to changes on sets of measure zero. In particular, by changing a collection of linearly dependent distributions on a set of measure zero, we can make them linearly independent. Hence, we may introduce stronger notions of linear independence as in Lemma 1 of Garg et al. (2020). We discuss this in App. D.2.

**D.1. Examples illustrating importance of weak positivity condition** In this section, we present two examples, one, to show that weak positivity isn’t sufficient for identifiability. Second, we present another example where we show that conditions in Proposition 2 are not necessary for identifiability.
Example 1  Assume $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5\}$ and $\mathcal{Y}_t = \{1, 2, 3\}$. Suppose the $p_t(x|y) = 1$, $p_t(x|y) = 2$, and $p_t(x)$ are given as:

|       | $p_t(x|y = 1)$ | $p_t(x|y = 2)$ | $p_t(x)$   |
|-------|---------------|---------------|------------|
| $x_1$ | 0.4           | 0.56          | 0.356      |
| $x_2$ | 0.3           | 0.3           | 0.207      |
| $x_3$ | 0.2           | 0.1           | 0.09       |
| $x_4$ | 0.1           | 0.04          | 0.042      |
| $x_5$ | 0.0           | 0.0           | 0.305      |

Here, there exists two separate $p_t(x|y = 3)$ and $p_t(y)$ that are consistent with the given $p_t(x|y = 1)$, $p_t(x|y = 2)$, and $p_t(x)$ and both the solutions satisfy weak positivity for two different $X_{wp}$ and $X'_{wp}$.

In particular, notice that $p_t(x|y = 3) = [0.17, 0.0675, 0.0, 0.0, 0.7625]^T$ and $p_t(y) = [0.3, 0.3, 0.4]$ gives us the first solution. $p_t(x|y = 3) = [0.0, 0.0, 0.0645, 0.0096, 0.9839]^T$ and $p_t(y) = [0.19, 0.5, 0.31]$ gives us another solution. For solution 1, $X_{wp} = \{x_3, x_4\}$ and for solution 2, $X'_{wp} = \{x_1, x_2\}$. To check consistency of each solution notice that $\sum_{x \in \mathcal{X}} p_t(y = i)p_t(x|y = i) = p_t(x)$ for each $x \in \mathcal{X}$.

In the above example, the key is to show that absent knowledge of which $x$’s constitute the set $X_{wp}$, we might be able to obtain multiple different solutions, each with different $X_{wp}$ and both $p_t(y)$, $p_t(x|y = k + 1)$ satisfying the given information and simplex constraints.

Next, we will show that in certain scenarios weak positivity is enough for identifiability.

Example 2  Assume $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ and $\mathcal{Y}_t = \{1, 2, 3\}$. Suppose the $p_t(x|y = 1)$, $p_t(x|y = 2)$, and $p_t(x)$ are given as,

|       | $p_t(x|y = 1)$ | $p_t(x|y = 2)$ | $p_t(x)$   |
|-------|---------------|---------------|------------|
| $x_1$ | 0.5           | 0.2           | 0.24       |
| $x_2$ | 0.3           | 0.4           | 0.2        |
| $x_3$ | 0.1           | 0.35          | 0.35       |
| $x_4$ | 0.1           | 0.05          | 0.21       |

Here, out of all $4C_2$ possibilities for $X_{wp}$, only one possibility yields a solution that satisfies weak positivity and simplex constraints. In particular, the solution is given by $p_t(x|y = 3) = [0.0, 0.0, 0.6, 0.4]^T$ and $p_t(y) = [0.4, 0.2, 0.4]$ with $X_{wp} = \{x_1, x_2\}$.

In this example, we show that conditions in Proposition 2 are not necessary to ensure identifiability. For discrete domains, this example also highlights that we can check identifiability in exponential time for any OSLS problem given $p_t(x)$ and $p_s(x|y)$ for all $y \in \mathcal{Y}_s$.

D.2. Extending identifiability conditions to continuous distributions
To extend our identifiability conditions for continuous distributions, the linear independence conditions on the matrix $[p_t(x|y)]_{x \in X_{wp}, y \in \mathcal{Y}_t}$ has the undesirable property of being sensitive to changes on sets of measure zero. In particular, by changing a collection of linearly dependent distributions on a set of measure zero, we can make them linearly independent. As a consequence, we may impose a stronger notion of independence, i.e., the set of distributions $\{p(x|y) : y = 1, \ldots, k\}$ are such that there does not exist $v \neq 0$ for which $\int_X \sum_y p(x|y)v_y dx = 0$, where $X = X_{wp}$ for necessary condition and $X = X_{sp}$ for sufficiency. We refer this condition as strict linear independence.

E. PULSE Framework
We begin with presenting our framework for OSLS problem under strong positivity condition. First, we explain the structure of OSLS that we leverage in PULSE framework and then elaborate design decisions we make to exploit the identified structure.
Rather than simply dividing each OSLS instance into $k$ PU problems, we exploit the joint structure of the problem. To begin, we note that if only we could apply a label shift correction to source, i.e., re-sample source classes according to their relative proportion in the target data, then we could subsequently consider the unlabeled target data as a mixture of (i) the (reweighted) source distribution; and (ii) the novel class distribution (i.e., $p_t(x|y=k+1)$). Formally, we have

$$p_t(x) = \sum_{j \in \mathcal{Y}_t} p_t(y=j) p_t(x|y=j) = \sum_{j \in \mathcal{Y}_s} p_t(y=j) \frac{w(j) p_s(x, y=j)}{p_t(y=j)} + p_t(x|y=k+1) p_t(y=k+1)$$

$$= (1 - p_t(x|y=k+1)) p_s'(x) + p_t(x|y=k+1) p_t(y=k+1), \quad (9)$$

where $p_s'(x)$ is the label-shift-corrected source distribution, i.e., $p_s'(x) = \sum_{j \in \mathcal{Y}_s} w(j) p_s(x, y=j)$, where $w(j) = (p_t(y=j) / \sum_{k} p_t(y=k)) / p_t(y=j)$ for all $j \in \mathcal{Y}_s$. Intuitively, $p_t'(j) = p_t(y=j) / \sum_{k} p_t(y=k)$ is re-normalized label distribution in target among source classes and $w(j)$'s are the importance weights. Hence, after applying a label shift correction to the source distribution $p_s'(x)$, we have reduced the OSLS problem to a single PU learning problem, where $p_s'(x)$ plays the part of the positive distribution and $p_t(x|y=k+1)$ acts as negative distribution with mixture coefficients $1 - p_t(x|y=k+1)$ and $p_t(x|y=k+1)$ respectively. We now discuss our methods to estimate the importance ratios $w(y)$ and to tackle the PU learning instance obtained from OSLS.

While traditional methods for estimating label shift breakdown in high dimensional settings (Zhang et al., 2013), recent methods exploit black-box classifiers to avoid the curse of dimensionality (Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2021). However, these recent techniques require overlapping label distributions, and a direct application would require demarcation of samples from $p_s'(x)$ in target, creating a cyclic dependency. Instead, we build on recently proposed Best Bin Estimation (BBE) technique from Garg et al. (2021b). Given positive and unlabeled data, BBE estimates the fraction of positives in unlabeled by leveraging a black-box classifier. In particular, BBE identifies a threshold on probability scores assigned by the classifier such that by estimating the ratio between the fractions of positive and unlabeled points receiving scores above the threshold, we obtain proportion of positives in unlabeled. We tailor BBE to estimate the probability scores assigned by the classifier such that by estimating the ratio between the fractions of positive and unlabeled.

Extending BBE algorithm to estimate target marginal among previously seen classes In a PU learning problem, given positive and unlabeled data, BBE estimates the fraction of positives in unlabeled by leveraging a black-box classifier. In particular, BBE identifies a threshold on probability scores assigned by the classifier such that by estimating the ratio
between the fractions of positive and unlabeled points receiving scores above the threshold, we obtain proportion of positives in unlabeled. We tailor BBE to estimate the relative fraction of previously seen classes in the target distribution by exploiting a k-way source classifier \( f_s \), trained on labeled source data. We describe the procedure in Algorithm 2.

For given probability density function \( p \) and a scalar output function \( f \), define a function \( q(z) = \int_{A_z} p(x) dx \), where \( A_z = \{ x \in \mathcal{X} : f(x) \geq z \} \) for all \( z \in [0, 1] \). Intuitively, \( q(z) \) captures the cumulative density of points in a top bin, the proportion of input domain that is assigned a value larger than \( z \) by the function \( f \) in the transformed space. We define an empirical estimator \( \hat{q}(z) \) given a set \( X = \{ x_1, x_2, \ldots, x_n \} \) sampled iid from \( p(x) \). Let \( Z = f(X) \). Define \( \hat{q}(z) = \sum_{i=1}^n I[z_i \geq z] / n \).

Our procedure proceeds as follows. Given a held-out dataset of source \( \{X^s_1, y^s_1\} \) and unlabeled target samples \( X^t_1 \), we push all examples through the source classifier \( f \) to obtain \( k \) dimensional outputs. For all \( j \in \mathcal{Y}_s \), we repeat the following: Obtain \( Z^s_j = f_j(X^s_1[|d|]) \) and \( Z^u = f_j(X^t_1) \). Next, with \( Z^s \) and \( Z^u \), we estimate \( \hat{q}_s \) and \( \hat{q}_t \). Finally, we estimate \( \hat{\theta}_1 \) as the ratio \( \hat{q}_t(\hat{c}) / \hat{q}_s(\hat{c}) \) at \( \hat{c} \) that minimizes the upper confidence bound at a pre-specified level \( \delta \) and a fixed parameter \( \gamma \in (0, 1) \). Our method is summarized in Algorithm 2. Throughout all the experiments, we fix \( \delta \) at 0.1 and \( \gamma \) at 0.01.

### Algorithm 2 Extending Best Bin Estimation (BBE) for Step 3 in Algorithm 1

**Input**: Validation source \( \{X^s_2, y^s_2\} \) and unlabeled target samples \( X^t_2 \). Source classifier \( f : \mathcal{X} \to \Delta^{k-1} \). Hyperparameter \( 0 < \delta, \gamma < 1 \).

1. \( \hat{p}_i \leftarrow \text{zeros(size} = |\mathcal{Y}_s|) \)
2. for \( j \in \mathcal{Y}_s \) do
3. \( \text{id}_j \leftarrow \text{where} (y^s_2 = j) \)
4. \( Z_s, Z_t \leftarrow [f(X^s_1[|d|])]_j, [f(X^t_1)]_j \)
5. \( \hat{q}_s(z), \hat{q}_t(z) \leftarrow \frac{\sum_{i\in\text{id}_j} I[z_i \geq z]}{|\text{id}_j|}, \frac{\sum_{i\in\text{id}_j} I[z_i \geq z]}{|X^t_1|} \) for all \( z \in [0, 1] \).
6. \( \hat{c}_j \leftarrow \arg \min_{c \in [0, 1]} \left( \frac{\hat{q}_t(c)}{\hat{q}_s(c)} + \frac{1 + \gamma}{\hat{q}_s(c)} \left( \sqrt{\frac{\log(4/\delta)}{2|\text{id}_j|}} + \sqrt{\frac{\log(4/\delta)}{2|X^t_1|}} \right) \right) \)
7. \( \hat{\theta}_1 \leftarrow \text{zeros(size} = |\mathcal{Y}_s|) \)
8. end for

**Output**: Normalized target marginal among source classes \( \hat{p}_t \leftarrow \hat{p}_1 / |\hat{\theta}_1| \).

### Extending CVIR to train discriminator \( f_d \) and estimate novel class prevalence

After estimating the fraction of source classes in target (i.e., \( p'_d(j) = p_u(y=j) \sum_{y=k} p_u(y=k) \) for all \( j \in \mathcal{Y}_s \)), we re-sample the source data according to \( p'_d(y) \) to mimic samples from distribution \( p'_u(y) \). Thus, obtaining a PU learning problem instance, we resort to PU learning techniques to (i) estimate the fraction of novel class \( p_t(y = k + 1) \); and (ii) learn a binary classifier \( f_d(x) \) to discriminate between label shift corrected source \( p'_d(x) \) and novel class \( p_t(x) \). Assume that sigmoid output \( f_d(x) \) indicates predicted probability of an example \( x \) belonging to label shifted novel \( p'_d(x) \).

Given an estimate of the fraction of novel class \( \hat{p}_t(y = k + 1) \), CVIR objective creates a provisional set of novel examples \( X_t^N \) by removing \((1 - \hat{p}_t(y = k + 1)) \) fraction of examples from \( X^t_1 \) that incur highest loss when predicted as novel class on each training epoch. Next, we update our discriminator \( f_d \) by minimizing loss on label shift corrected source \( X^d_t \) and provisional novel examples \( X^N_t \). This step is aimed to remove any incentive to overfit to the examples from \( p'_d(x) \). Consequently, we employ the iterative procedure that alternates between estimating the prevalence of novel class \( \hat{p}_t(y = k + 1) \) (with BBE) and minimizing the CVIR loss with estimated fraction of novel class. Algorithm 3 summarizes our approach which is used in Step 3 of Algorithm 1.

Note that we need to warm start with simple domain discrimination training, since in the initial stages mixture proportion estimate is often close to 1 rejecting all the unlabeled examples. In Garg et al. (2021b), it was shown that the procedure is not sensitive to the choice of number of warm start epochs and in a few cases with large datasets, we can even get away without warm start (i.e., \( W = 0 \)) without hurting the performance. In our work, we notice that given an estimate \( \hat{\theta} \) of prevalence of novel class, we can use unbiased PU error (4) on validation data as a surrogate to identify warm start epochs for domain discriminator training. In particular, we train the domain discriminator classifier for a large number of epochs, say \( E(\gg W) \), and then choose the discriminator, i.e., warm start epoch \( W \) at which \( f_d \) achieves minimum unbiased validation loss.

Finally, to obtain a \((k + 1)\)-way classifier \( f_i(x) \) on target we combine discriminator \( f_d \) and source classifier \( f_s \) with
importance-reweighted label shift correction. In particular, for all \( j \in \mathcal{Y}_n \), \( [f_t(x)]_j = (f_d(x)) \frac{w(j) \cdot [f_r(x)]_j}{\sum_{k \in \mathcal{Y}_n} w(k) \cdot [f_r(x)]_k} \) and \( [f_l(x)]_j = 1 - f_d(x) \). Similarly, to obtain target marginal \( p_t \), we re-scale the label shift estimate among previously seen classes with estimate of prevalence of novel examples, i.e., for all \( j \in \mathcal{Y}_n \), assign \( \tilde{p}_t(y = j) = (1 - \tilde{p}_t(y = k + 1)) \cdot \tilde{p}_t(y = j) \).

Overall, our approach proceeds as follows (Algorithm 1): First, we estimate the label shift among previously seen classes. Then we employ importance re-weighting of source data to formulate a single PU learning problem between source and target to estimate fraction of novel class \( \tilde{p}_t(y = k + 1) \) and to learn a discriminator \( f_d \) for the novel class. Combining discriminator and label shift corrected source classifier we get \((k+1)\)-way target classifier.

**Algorithm 3** Alternating between CVIR and BBE for Step 5 in Algorithm 1

**input**: Re-sampled training source data \( \tilde{X}_S^T \), validation source data \( \tilde{X}_S^V \). Training target data \( X_T^T \) and validation data \( X_T^V \).

Hyperparameter \( W, B, \delta, \gamma \).

1: Initialize a training model \( f_\theta \) and an stochastic optimization algorithm \( \mathcal{A} \).
2: \( X_N^T \leftarrow X_T^T \).

(/ Warm start with domain discrimination training)

3: for \( i \leftarrow 1 \) to \( B \) do

4: Shuffle \( (X_N^T, X_N^V) \) into \( B \) mini-batches. With \( (X_N^T[i], X_N^V[i]) \) we denote \( i^{\text{th}} \) mini-batch.
5: for \( i \leftarrow 1 \) to \( B \) do

6: Set the gradient \( \nabla_{\theta} \left[ \tilde{L}^+(f_\theta; \tilde{X}_N^T[i]) + \tilde{L}^-(f_\theta; X_N^V[i]) \right] \) and update \( \theta \) with algorithm \( \mathcal{A} \).

7: end for
8: end for
9: \( \hat{\alpha} \leftarrow \text{BBE}(\tilde{X}_S^S, \tilde{X}_S^T, f_\theta) \) \quad [\text{Algorithm 4}]
10: Rank samples \( x \in X_T^T \) according to their loss values \( \ell(f_\theta(x), -1) \).
11: \( X_N^T \leftarrow \{ \tilde{X}_T^T \}^{1 - \hat{\alpha}} \) where \( \{ \tilde{X}_T^T \}^{1 - \hat{\alpha}} \) denote the lowest ranked \( 1 - \hat{\alpha} \) fraction of samples.
12: while training error \( \hat{L}^+(f_\theta; \tilde{X}_S^S) + \hat{L}^-(f_\theta; X_N^T) \) is not converged do

13: Train model \( f_\theta \) for one epoch on \( (\tilde{X}_S^S, X_N^T) \) as in Lines 4-7.
14: \( \hat{\alpha} \leftarrow \text{BBE}(\tilde{X}_S^S, \tilde{X}_S^T, f_\theta) \) \quad [\text{Algorithm 4}]
15: Rank samples \( x \in X_T^T \) according to their loss values \( \ell(f_\theta(x), 1) \).
16: \( X_N^T \leftarrow \{ X_T^T \}^{1 - \hat{\alpha}} \) where \( \{ X_T^T \}^{1 - \hat{\alpha}} \) denote the lowest ranked \( 1 - \hat{\alpha} \) fraction of samples.
17: end while

**output**: Trained discriminator \( f_d \leftarrow f_\theta \) and novel class fraction \( \tilde{p}_t(y = k + 1) \leftarrow 1 - \hat{\alpha} \).

**Algorithm 4** Best Bin Estimation (BBE)

**input**: Re-sampled source data \( \tilde{X}^S \) and target samples \( X^T \). Discriminator classifier \( \hat{f} : \mathcal{X} \rightarrow [0, 1] \). Hyperparameter \( 0 < \delta, \gamma < 1 \).

1: \( Z_s, Z_t \leftarrow f(\tilde{X}^S), f(X^T) \).
2: \( \hat{q}_t(z), \hat{q}_s(z) \leftarrow \frac{\sum_i z_i \cdot [z_i > z]}{|X^s|}, \frac{\sum_i z_t \cdot [z_t > z]}{|X^t|} \) for all \( z \in [0, 1] \).
3: Estimate \( \hat{c} \leftarrow \arg \min_{c \in [0, 1]} \left( \hat{q}_t(c) + \frac{1 + \gamma}{\frac{1}{2} |X^t|} \left( \sqrt{\frac{\log(4/\delta)}{2 |X^t|}} + \sqrt{\frac{\log(4/\delta)}{2 |X^t|}} \right) \right) \).

**output**: \( \hat{\alpha} \leftarrow \hat{q}_s(c) \hat{q}_t(c) \)

E.1. PULSE under separability

Our ideas for PULSE framework can be extended to separability condition since (9) continues to hold. In particular, when OSLS satisfies the separability assumption, we may hope to jointly estimate the label shift among previously seen classes with label shift estimation techniques (Lipton et al., 2018; Alexandari et al., 2021) and learn a domain discriminator classifier. This may be achieved by estimating label shift among examples rejected by domain discriminator classifier as belonging to previously seen classes. However, in our initial experiments, we observe that techniques proposed under strong positivity were empirically stable and outperform methods developed under separability. This is intuitive for many benchmark datasets where it may be more natural to expect that for each class there exists a subdomain that only belongs to that class than
assuming separability only between novel class samples and examples from source classes.

F. Proofs for analysis of OSLS framework

In this section, we provide missing formal statements and proofs for theorems in Sec. 5. This mainly includes analysing key steps of our PULSE procedure for target label marginal estimation (Step 3, 5 Algorithm 1) and learning the domain discriminator classifier (Step 5, Algorithm 1).

F.1. Formal statement and proof of Theorem 1

When the data satisfies strong positivity, we observe that source classifiers often exhibit a threshold \( c_y \) on softmax output of each class \( y \in \mathcal{Y}_s \) above which the top bin (i.e., \([c_y, 1]\)) contains mostly examples from that class \( y \). We give empirical evidence to this claim in App. F.1. Then, we show that the existence of (nearly) pure top bin for each class \( y \) is sufficient for Step 3 in Algorithm 1 to produce (nearly) consistent estimates.

Before introducing the formal statement, we introduce some additional notation. Given probability density function \( p \) and a source classifier \( f : \mathcal{X} \rightarrow \Delta^{k-1} \), define a function \( q(z, j) = \int_{A(z, j)} p(x)dx \) where \( A(z, j) = \{x \in \mathcal{X} : [f(x)]_j \geq z\} \) for all \( z \in [0, 1] \). Intuitively, \( q(z, j) \) captures the cumulative density of points in a top bin for class \( j \), i.e., the proportion of input domain that is assigned a value larger than \( z \) by the function \( f \) at the index \( j \) in the transformed space. We define an empirical estimator \( \hat{q}(z, j) \) given a set \( X = \{x_1, x_2, \ldots, x_n\} \) sampled iid from \( p(x) \). Let \( Z = \{f(X)\}_j \). Define \( \hat{q}(z, j) = \sum_{i=1}^n 1[z_i \geq z]/n \).

For each \( p_s \) and \( p_t \), we define \( q_s \) and \( q_t \) respectively. Moreover, for each class \( j \in \mathcal{Y}_s \), we define \( q_{t, j} \) corresponding to \( p_{t, j} := p_t(x|y = j) \) and \( q_{t, -j} \) corresponding to \( p_{t, -j} := \frac{\sum_{i \in \mathcal{X}, y=i} p_t(x|y=i)}{\sum_{i \in \mathcal{Y}_s, j} p_t(x|y=j)} \). Assume that we have \( n_s \) source examples and \( m_t \) target examples. Now building on BBE results from Garg et al. (2021b), we present finite sample results for target label marginal estimation:

**Theorem 3** (Formal statement of Theorem 1). Define \( c^*_y = \arg\min_{c \in [0, 1]} (q_{t, -j}(c, j)/q_{t, j}(c, j)) \), for all \( j \in \mathcal{Y}_s \). Assume \( \min(n, m) \geq \max_{j \in \mathcal{Y}_s} \left(\frac{2\log(4k/\delta)}{q_{t, j}(c^*_j, j)}\right) \). Then, for every \( \delta > 0 \), \( \hat{p}_t \) (in Algorithm 2 with \( \delta \) as \( \delta/k \)) satisfies with probability at least \( 1 - \delta \), we have:

\[
|\hat{p}_t - p_t|_1 \leq \sum_{j \in \mathcal{Y}_s} (1 - p_t(y = j)) \left( \frac{q_{t, -j}(c^*_j, j)}{q_{t, j}(c^*_j, j)} \right) + O \left( \sqrt{\frac{k^3 \log(4k/\delta)}{n}} + \sqrt{\frac{k^2 \log(4k/\delta)}{m}} \right).
\]

When the data satisfies strong positivity, we observe that source classifiers often exhibit a threshold \( c_y \) on softmax output of each class \( y \in \mathcal{Y}_s \) above which the top bin (i.e., \([c_y, 1]\)) contains mostly examples from that class \( y \). Formally, as long as there exist a threshold \( c^*_y \in (0, 1) \) such that \( q_{t, j}(c^*_j) \geq c \) and \( q_{t, -j}(c^*_j) = 0 \) for some constant \( c > 0 \) for all \( j \in \mathcal{Y}_s \), we show that our estimator \( \hat{\alpha} \) converges to the true \( \alpha \) with convergence rate \( \min(n, m)^{-1/2} \). The proof technique simply builds on the proof of Theorem 1 in Garg et al. (2021b). First, we state Lemma 1 from Garg et al. (2021b). Next, for completeness we provide the proof for Theorem 3 which extends proof of Theorem 1 (Garg et al., 2021b) for \( k \) classes.

**Lemma 1.** Assume two distributions \( p_s \) and \( q_s \) with their empirical estimators denoted by \( \hat{q}_p \) and \( \hat{q}_q \) respectively. Then for every \( \delta > 0 \), with probability at least \( 1 - \delta \), we have for all \( c \in [0, 1] \)

\[
|\hat{q}_q(c) - q_s(c)|/q_p(c) \leq \frac{1}{q_p(c)} \left( \frac{\log(4/\delta)}{2n_s} + \frac{q_u(c)}{q_p(c)} \sqrt{\frac{\log(4/\delta)}{2n_p}} \right).
\]

**Proof of Theorem 3.** The main idea of the proof is to use the confidence bound derived in Lemma 1 at \( \hat{\alpha} \) and use the fact that \( \hat{\alpha} \) minimizes the upper confidence bound. The proof is split into two parts. First, we derive a lower bound on \( \hat{q}_{t, j}(\hat{c}_j) \) for all \( j \in \mathcal{Y}_s \) and next, we use the obtained lower bound to derive confidence bound on \( \hat{p}_t(y = j) \). With \( \hat{\alpha}_j \), we denote \( \hat{p}_t(y = j) \) for all \( j \in \mathcal{Y}_s \). All the statements in the proof simultaneously hold with probability \( 1 - \delta/k \). We derive the bounds for a single \( j \in \mathcal{Y}_s \) and then use union bound to combine bound for all \( j \in \mathcal{Y}_s \). When it is clearly from context, we denote...
where the last inequality follows from the fact that lower bound as follows:

\[
q_{t,j}(c) := \arg \min_{c \in [0,1]} \frac{\hat{q}_t(c)}{q_{t,j}(c)} + \frac{1}{q_{t,j}(c)} \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + (1 + \gamma) \sqrt{\frac{\log(4k/\delta)}{2np_s(y = j)}} \right)
\]

and

\[
\hat{p}_t(y = j) := \frac{\hat{q}_t(c_j)}{q_{t,j}(c_j)}.
\]

Moreover,

\[
c_j^* := \arg \min_{c \in [0,1]} q_t(c) \quad \text{and} \quad \alpha_j^* := \frac{q_t(c_j^*)}{q_{t,j}(c_j^*)}.
\]

**Part 1:** We establish lower bound on \(\hat{q}_{t,j}(c_j^*)\). Consider \(c_j^* \in [0,1]\) such that \(\hat{q}_{t,j}(c_j^*) = \frac{1}{\hat{q}_{t,j}(c_j) \cdot \hat{q}_t(c_j^*)}\). We will now show that Algorithm 2 will select \(c_j < c_j^*\). For any \(c \in [0,1]\), we have with probability \(1 - \delta/k\),

\[
\hat{q}_{t,j}(c) - \sqrt{\frac{\log(4k/\delta)}{2m}} \leq q_{t,j}(c) \quad \text{and} \quad q_t(c) - \sqrt{\frac{\log(4k/\delta)}{2m}} \leq \hat{q}_t(c).
\]

Since \(\frac{q_t(c_j^*)}{q_{t,j}(c_j^*)} \leq \frac{q_t(c)}{q_{t,j}(c)}\), we have

\[
\hat{q}_t(c) \geq q_{t,j}(c) \frac{q_t(c_j^*)}{q_{t,j}(c_j^*)} - \sqrt{\frac{\log(4k/\delta)}{2m}} \geq \left( \frac{q_t(c_j^*)}{q_{t,j}(c_j^*)} - \sqrt{\frac{\log(4k/\delta)}{2m}} \right) \frac{q_t(c_j^*)}{q_{t,j}(c_j^*)} - \sqrt{\frac{\log(4k/\delta)}{2m}}.
\]

Therefore, at \(c\) we have

\[
\frac{\hat{q}_t(c)}{q_{t,j}(c)} \geq \alpha_j^* - \frac{1}{q_{t,j}(c)} \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \frac{q_t(c_j^*)}{q_{t,j}(c_j^*)} \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right).
\]

Using Lemma 1 at \(c^*\), we have

\[
\frac{\hat{q}_t(c)}{q_{t,j}(c)} \geq \frac{\hat{q}_t(c_j^*)}{q_{t,j}(c_j^*)} - \left( \frac{1}{q_{t,j}(c_j^*)} + \frac{1}{q_{t,j}(c)} \right) \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \frac{q_t(c_j^*)}{q_{t,j}(c_j^*)} \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right)
\]

\[
\geq \frac{\hat{q}_t(c_j^*)}{q_{t,j}(c_j^*)} - \left( \frac{1}{q_{t,j}(c_j^*)} + \frac{1}{q_{t,j}(c)} \right) \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right),
\]

where the last inequality follows from the fact that \(\alpha_j^* = \frac{q_t(c_j^*)}{q_{t,j}(c_j^*)} \leq 1\). Furthermore, the upper confidence bound at \(c\) is lower bound as follows:

\[
\frac{\hat{q}_t(c)}{q_{t,j}(c)} + \frac{1 + \gamma}{q_{t,j}(c)} \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right)
\]

\[
\geq \frac{\hat{q}_t(c_j^*)}{q_{t,j}(c_j^*)} + \left( \frac{1 + \gamma}{q_{t,j}(c_j^*)} - \frac{1}{q_{t,j}(c)} \right) \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right)
\]

\[
= \frac{\hat{q}_t(c_j^*)}{q_{t,j}(c_j^*)} + \left( \gamma - \frac{1}{q_{t,j}(c_j^*)} \right) \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right)
\]

Using (20) at \(c = c^*\), we have the following lower bound on ucb at \(c^*\):

\[
\frac{\hat{q}_t(c^*)}{q_{t,j}(c^*)} + \frac{1 + \gamma}{q_{t,j}(c^*)} \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right)
\]

\[
\geq \frac{\hat{q}_t(c_j^*)}{q_{t,j}(c_j^*)} + \left( \gamma - \frac{1}{q_{t,j}(c_j^*)} \right) \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right).
\]
Moreover from (20), we also have that the lower bound on ucb at \( c \), \( c' \) is strictly greater than the lower bound on ucb at \( c' \). Using definition of \( \hat{c} \), we have

\[
\frac{\hat{q}_t(c^*_j)}{\hat{q}_{t,j}(c^*_j)} + 1 + \gamma \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right)
\]  

\[
\geq \frac{\hat{q}_t(c)}{\hat{q}_{t,j}(c)} + 1 + \gamma \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right),
\]  

and hence

\[
\hat{c} \leq c'.
\]

**Part 2:** We now establish an upper and lower bound on \( \hat{F}_j \). We start with upper confidence bound on \( \hat{F}_j \). By definition of \( \hat{c}_j \), we have

\[
\frac{\hat{q}_t(c^*_j)}{\hat{q}_{t,j}(c^*_j)} + 1 + \gamma \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right)
\]  

\[
\leq \min_{c \in [0, 1]} \left[ \frac{\hat{q}_t(c)}{\hat{q}_{t,j}(c)} + 1 + \gamma \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right) \right] 
\]  

\[
\leq \frac{\hat{q}_t(c^*_j)}{\hat{q}_{t,j}(c^*_j)} + 1 + \gamma \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right).
\]

Using Lemma 1 at \( c^*_j \), we get

\[
\frac{\hat{q}_t(c^*_j)}{\hat{q}_{t,j}(c^*_j)} \leq \frac{q_t(c^*_j)}{q_{t,j}(c^*_j)} + 1 + \frac{1}{q_{t,j}(c^*_j)} \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right)
\]  

\[
= \alpha^*_j + 1 + \gamma \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}}.
\]

Combining (28) and (29), we get

\[
\hat{F}_j = \frac{\hat{q}_t(c)}{\hat{q}_{t,j}(c)} \leq \alpha^*_j + 2 + \gamma \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}}.
\]

Using DKW inequality on \( q_{t,j}(c^*_j) \), we have \( \hat{q}_{t,j}(c^*_j) \geq q_{t,j}(c^*_j) - \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \). Assuming \( n \cdot p_s(y = j) \geq \frac{2\log(4k/\delta)}{\alpha^*_j} \), we get \( \hat{q}_{t,j}(c^*_j) \leq q_{t,j}(c^*_j)/2 \) and hence,

\[
\hat{F}_j \leq \alpha^*_j + 4 + 2\gamma \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}}.
\]

Finally, we now derive a lower bound on \( \hat{F}_j \). From Lemma 1, we have the following inequality at \( \hat{c} \)

\[
\frac{q_t(c)}{q_{t,j}(c)} \leq \frac{\hat{q}_t(c)}{\hat{q}_{t,j}(c)} + 1 + \gamma \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right).
\]

Since \( \alpha^*_j \leq \frac{q_t(c)}{q_{t,j}(c)} \), we have

\[
\alpha^*_j \leq \frac{q_t(c)}{q_{t,j}(c)} \leq \frac{\hat{q}_t(c)}{\hat{q}_{t,j}(c)} + 1 + \gamma \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right).
\]
Using (31), we obtain a very loose upper bound on \( \frac{\hat{q}_i(c)}{\hat{q}_{t,i}(c)} \). Assuming \( \min(n \cdot p_s(y = j), m) \geq \frac{2\log(4k/\delta)}{q_{t,j}(c^*_j)} \), we have

\[
\alpha_j^* \leq \frac{\hat{q}_i(c)}{\hat{q}_{t,i}(c)} + \frac{1}{\hat{q}_{t,i}(c)} \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + (5 + 2\gamma) \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right).
\]

Moreover, as \( \tilde{c} \geq c^* \), we have \( \hat{q}_{t,i}(c^*) \geq \frac{\gamma}{\gamma + \gamma} \hat{q}_{t,i}(c^*_j) \) and hence,

\[
\alpha_j^* - \frac{\gamma + 2}{\gamma \hat{q}_{t,i}(c^*_j)} \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + (5 + 2\gamma) \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right) \leq \frac{\hat{q}_i(c)}{\hat{q}_{t,i}(c)} = \hat{\alpha}_j.
\]

As we assume \( n \cdot p_s(y = j) \geq \frac{2\log(4k/\delta)}{q_{t,j}(c^*_j)} \), we have \( \hat{q}_{t,i}(c^*_j) \leq q_{t,j}(c^*_j)/2 \), which implies the following lower bound on \( \alpha \):

\[
\alpha_j^* \geq \frac{2\gamma + 4}{\gamma q_{t,i}(c^*_j)} \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + (5 + 2\gamma) \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right) \leq \hat{\alpha}_j.
\]

Combining lower bound (36) and upper bound (31), we get

\[
|\hat{\alpha}_j - \alpha_j^*| \leq l_j \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{\frac{\log(4k/\delta)}{2n \cdot p_s(y = j)}} \right),
\]

for some constant \( l_j \). Additionally by our assumption of OSLS problem \( p_s(y = j) > c/k \) for some constant \( c > 0 \), we have

\[
|\hat{\alpha}_j - \alpha_j^*| \leq l'_j \left( \sqrt{\frac{\log(4k/\delta)}{2m}} + \sqrt{k \log(4k/\delta)} \right),
\]

for some constant \( l'_j \).

Combining the above obtained bound for all \( j \in \mathcal{Y}_s \) with union bound, we get with probability at least \( 1 - \delta \),

\[
\sum_{j \in \mathcal{Y}_s} |\hat{\alpha}_j - \alpha_j^*| \leq l'_\max \left( \sqrt{\frac{k^2 \log(4k/\delta)}{2m}} + \sqrt{\frac{k^3 \log(4k/\delta)}{2n}} \right),
\]

where \( l'_\max = \max l'_j \). Now, note that for each \( j \in \mathcal{Y}_s \), we have \( q_t(c) = p_t(y = j) \cdot q_{t,j}(c) + (1 - p_t(y = j)) \cdot q_{t,-j}(c) \). Hence \( \alpha_j^* = p_t(y = j) + (1 - p_t(y = j)) \cdot q_{t,-j}(c) / q_{t,j}(c) \). Plugging this in, we get the desired bound.

\[\tag*{□}\]

**Empirical evidence of the top bin property** We now empirically validate the positive pure top bin property (Fig. 2). We include results with Resnet-18 trained on the CIFAR10 OSLS setup same as our main experiments. We observe that source classifier approximately satisfies the positive pure top bin property for small enough top bin sizes.

### F.2. Formal statement and proof of Theorem 2

In this section, we show that in population on a separable Gaussian dataset, CVIR will recover the optimal classifier. Note that here we consider a binary classification problem similar to the one in Step 5 in Algorithm 1. Since we are primarily interested in analysing the iterative procedure for obtaining domain discriminator classifier, we assume that \( \alpha \) is known.

In population, we have access to positive distribution (i.e., \( p_p \)), unlabeled distribution (i.e., \( p_u := \alpha p_p + (1 - \alpha) p_n \)), and mixture coefficient \( \alpha \). Our goal is to recover the classifier that discriminates \( p_p \) versus \( p_n \). Making a parallel connection to Step 5 of PULSE, positive distribution \( p_p \) here refers to the label shift corrected source distribution \( p'_p \) and \( p_u \) refers to \( p_t = p_t(y = k + 1) p_t(x | y = k + 1) + (1 - p_t(y = k + 1)) p'_t(x) \). Our goal is to recover the classifier that discriminates \( p_p \) versus \( p_n \) (parallel \( p'_p \) versus \( p_t(y = k + 1) \)).
Results with ResNet-18 on CIFAR10 OSLS setup. Details of the setup in App. H.2. As the bin size increases for all classes the purity decreases.

For ease, we re-introduce some notation. For a classifier \( f \) and loss function \( \ell \), define

\[
\text{VIR}_\alpha(f) = \inf\{\tau \in \mathbb{R} : P_{x \sim p_u}(\ell(x, -1; f) \leq \tau) \geq 1 - \alpha\}. \tag{40}
\]

Intuitively, \( \text{VIR}_\alpha(f) \) identifies a threshold \( \tau \) to capture bottom \( 1 - \alpha \) fraction of the loss \( \ell(x, -1) \) for points \( x \) sampled from \( p_u \). Additionally, define CVIR loss as

\[
\mathcal{L}(f, w) = \alpha \mathbb{E}_{p_p}[\ell(x, 1; f)] + \mathbb{E}_{p_u}[w(x)\ell(x, -1; f)], \tag{41}
\]

for classifier \( f \) and some weights \( w(x) \in \{0, 1\} \). Recall that given a classifier \( f_t \) at an iterate \( t \), CVIR procedure proceeds as follows:

\[
w_{t+1}(x) = \mathbb{I}[\ell(x, -1; f_t) \leq \text{VIR}_\alpha(f_t)], \tag{42}
\]

\[
f_{t+1} = f_t - \eta \nabla \mathcal{L}(f_t, w_t). \tag{43}
\]

We assume a data generating setup with where the support of positive and negative data is completely disjoint. We assume that \( x \) are drawn from two half multivariate Gaussian with mean zero and identity covariance, i.e.,

\[
x \sim p_p \iff x = \gamma_0 \theta_{\text{opt}} + z |\theta_{\text{opt}}^T z \geq 0, \text{ where } z \sim \mathcal{N}(0, I_d)
\]

\[
x \sim p_n \iff x = -\gamma_0 \theta_{\text{opt}} + z |\theta_{\text{opt}}^T z < 0, \text{ where } z \sim \mathcal{N}(0, I_d)
\]

Here \( \gamma_0 \) is the margin and \( \theta_{\text{opt}} \in \mathbb{R}^d \) is the true separator. Here, we have access to distribution \( p_p \) and \( p_u = \alpha p_p + (1 - \alpha)p_n \). Assume \( \ell \) as the logistic loss. For simplicity, we will denote \( \mathcal{L}(f_0, w_0) \) with \( \mathcal{L}(\theta_0, w_0) \).

**Theorem 4** (Formal statement of Theorem 2). In the data setup described above, a linear classifier \( f(x; \theta) = \mathbb{I}[\theta^T x > 0] \) initialized at some \( \theta_0 \) such that \( \mathcal{L}(\theta_0, w_0) < \log(2) \), trained with CVIR procedure as in equations (42)-(43) will converge to an optimal positive versus negative classifier.

**Proof of Theorem 4.** The proof uses two key ideas. One, at convergence of the CVIR procedure, the gradient of CVIR loss in (41) converges to zero. Second, for any classifier \( \theta \) not in the direction of \( \theta_{\text{opt}} \), we show that the CVIR gradient in (41) is non-zero.

**Part 1** We first show that the loss function \( \mathcal{L}(\theta, w) \) in (41) is 2-smooth with respect to \( \theta \) for fixed \( w \). Using gradient descent lemma with the decreasing property of loss in (42)-(43), we show that gradient converges to zero eventually. Considering gradient of \( \mathcal{L} \), we have

\[
\nabla_{\theta} \mathcal{L}(\theta, w) = \alpha \mathbb{E}_{p_p}[(f(x; \theta) - 1)x] + \mathbb{E}_{p_u}[w(x)(f(x; \theta) - 0)x]. \tag{44}
\]
Moreover, $\nabla^2 \mathcal{L}$ is given by
\[
\nabla^2 \mathcal{L}(\theta, w) = \alpha \mathbb{E}_p \left[ \nabla f(x; \theta)x x^T \right] + \mathbb{E}_p \left[ w(x) \nabla f(x; \theta)x x^T \right].
\] (45)

Since $\nabla f(x; \theta) \preceq 1$, we have $v^T \nabla^2 \mathcal{L} v \leq 2$ for all unit vector $v \in \mathbb{R}^d$. Now, by gradient descent lemma if $\eta \leq 1/2$, at any step $t$ we have, $\mathcal{L}(\theta_{t+1}, w_t) \leq \mathcal{L}(\theta_t, w_t)$. Moreover, by definition of VIR$_u(\theta)$ in (40) and update (42), we have $\mathcal{L}(\theta_{t+1}, w_{t+1}) \leq \mathcal{L}(\theta_t, w_t)$. Hence, we have $\mathcal{L}(\theta_{t+1}, w_{t+1}) \leq \mathcal{L}(\theta_t, w_t)$. Since, the loss is lower bounded from below at 0, for every $\epsilon > 0$, we have for large enough $t$ (depending on $\epsilon$), $\|\mathcal{L}(\theta_t, w_t)\|_2 \leq \epsilon$, i.e., $\|\mathcal{L}(\theta_t, w_t)\|_2 \to 0$ as $t \to \infty$.

**Part 2** Consider a general scenario when $\gamma > 0$. Denote the input domain of $p_p$ and $p_n$ as $P$ and $N$ respectively. At any step $t$, for all points $x \in \mathcal{X}$ such that $p_n(x) > 0$ and $w_t(x) = 0$, we say that $x$ is rejected from $p_n$. We denote the incorrectly rejected subdomain of $p_n$ from $p_n$ as $N_r$ and the incorrectly accepted subdomain of $p_p$ from $p_n$ as $P_a$. Formally, $N_r = \{x : p_n(x) > 0 \text{ and } w_t(x) = 0\}$ and $P_a = \{x : p_p(x) > 0 \text{ and } w_t(x) = 1\}$. We will show that $p_p(P_a) \to 0$ as $t \to \infty$, and hence, we will recover the optimal classifier where we reject none of $p_n$ incorrectly.

Observe that at any time $t$, for fixed $w_t$ and $\theta = \theta_t$, the gradient of CVIR loss in (41), can be expressed as:
\[
\nabla_\theta \mathcal{L}(\theta, w_t) = \alpha \int_{x \in \mathcal{P} \setminus \mathcal{P}_a} (f(x; \theta) - 1)x \cdot p_p(x) dx + (1 - \alpha) \int_{x \in \mathcal{N} \setminus \mathcal{N}_r} (f(x; \theta) - 0)x \cdot p_n(x) dx
\]
\[
+ \alpha \int_{x \in \mathcal{P}_a} (2f(x; \theta) - 1)x \cdot p_p(x) dx.
\] (46)

Note that for any $x, \theta$, $0 \leq f(x; \theta) \leq 1$. Now consider inner product of individual terms above with $\theta_{opt}$, we get
\[
\langle I, \theta_{opt} \rangle = \int_{x \in \mathcal{P} \setminus \mathcal{P}_a} (f(x; \theta) - 1)x^T \theta_{opt} \cdot p_p(x) dx \leq -\gamma_0 \int_{x \in \mathcal{P} \setminus \mathcal{P}_a} (1 - f(x; \theta)) \cdot p_p(x) dx,
\] (47)
\[
\langle II, \theta_{opt} \rangle = \int_{x \in \mathcal{N} \setminus \mathcal{N}_r} (f(x; \theta) - 0)x^T \theta_{opt} \cdot p_n(x) dx \leq -\gamma_0 \int_{x \in \mathcal{N} \setminus \mathcal{N}_r} (f(x; \theta) - 0) \cdot p_n(x) dx,
\] (48)
\[
\langle III, \theta_{opt} \rangle = \int_{x \in \mathcal{P}_a} (2f(x; \theta) - 1)x^T \theta_{opt} \cdot p_p(x) dx \leq -\gamma_0 \int_{x \in \mathcal{P}_a} (1 - 2f(x; \theta)) \cdot p_p(x) dx.
\] (49)

Note that in this problem VIR$_u(\theta) \leq 0.5$. Then equation (49) follows by the fact that $f(x; \theta) \leq 0.5$ for all $x \in \mathcal{P}_a$ since all points in $\mathcal{P}_a$ are assigned $w(x) = 1$.

From Part 1, for gradient $\|\nabla_\theta \mathcal{L}(\theta_t, w_t)\|_2$ to converge to zero as $t \to \infty$, we must have that LHS in equations (47), (48), and (49) converges to zero individually. Since CVIR loss decreases continuously and $\mathcal{L}(\theta_0, w_0) < \log(2)$, we have that $f(x; \theta_t) \to 1$ for all $x \in \mathcal{P} \setminus \mathcal{P}_a$, $f(x; \theta_t) \to 0$ for all $x \in \mathcal{N} \setminus \mathcal{N}_r$, and $p_p(P_a) \to 0$. Hence the classifier obtained has zero error for positive versus negative classification.

The above analysis can be extended to show convergence to max-margin classifier by using arguments from Soudry et al. (2018). In particular, as $p_p(P_a) \to 0$, we will have $\theta_t / \|\theta_t\|_2$ converging to the max-margin classifier for $p_p$ versus $p_n$, i.e., $\theta_{opt}$. Note that we need an assumption that the initialized model $\theta_0$ is strictly better than a model that randomly guesses or initialized at all zeros. This is to avoid convergence to the local minima of $\theta = 0$ with CVIR training. This assumption is satisfied when the classifier is initialized in a way such that $\langle \theta_0, \theta_{opt} \rangle > 0$. In general, we need a weaker assumption that during training with any randomly initialized classifier, there exists an iterate $t$ during CVIR training such that $\langle \theta_t, \theta_{opt} \rangle > 0$.

**F.3. Extension of Theorem 1**

We also extend the analysis in the proof of Theorem 3 to Step 5 of Algorithm 1 to show convergence of estimate $\hat{p}_i(y = k + 1)$ to true prevalence $p_i(y = k + 1)$. In particular, we show that the estimation error for prevalence of the novel class will primarily depend on sum of two terms: (i) error in approximating the label shift corrected source distribution, i.e., $p'_s(x)$; and (ii) purity of the top bin of the domain discriminator classifier.
Before formally introducing the result, we introduce some notation. Similar to before, given probability density function $p$ and a domain discriminator classifier $f : \mathcal{X} \to \Delta$, define a function $q = \int_{A(z)} p(x) dx$, where $A(z) = \{ x \in \mathcal{X} : f(x) \geq z \}$ for all $z \in [0, 1]$. Intuitively, $q(z)$ captures the cumulative density of points in a top bin, i.e., the proportion of input domain that is assigned a value larger than $z$ by the function $f$ in the transformed space. We denote $p_t(x | y = k + 1)$ with $p_{t,k+1}$. For each pdf $p_t, p_{t,k+1}$, and $p'_t$, we define $q_t, q_{t,k+1}$, and $q'_t$ respectively. Note that since we define an empirical estimator $\hat{q}(z)$ given a set $X = \{ x_1, x_2, \ldots, x_n \}$ sampled iid from $p(x)$. Let $Z = f(X)$. Define \( \hat{q}(z) = \sum_{i=1}^{n} I[ z_i \geq z ] / n \).

Recall that in Step 5 of Algorithm 1, to estimate the proportion of novel class, we have access to re-sampled data from approximate label shift corrected source distribution $\hat{q}'(x)$. Assume that we the size of re-sampled dataset is $n$.

**Theorem 5.** Define $e^\ast = \arg\min_{c \in [0, 1]} \left( q_{t,k+1}(c) / \hat{q}'(c) \right)$. Assume $\min(n, m) \geq \left( \frac{16 \log(4/\delta)}{\min(q_{k+1}'(c^\ast))^2} \right)$. Then, for every $\delta > 0$, $[\hat{p}]_{k+1} := \hat{p}_t(y = k + 1)$ in Step 5 of Algorithm 1 satisfies with probability at least $1 - \delta$, we have:

\[
\left| \left[ \hat{p} \right]_{k+1} - [p_t]_{k+1} \right| \leq \left( 1 - [p_t]_{k+1} \right) \frac{|q'_{k+1}(c^\ast) - \hat{q}'(c^\ast)|}{q_{k+1}'(c^\ast)} + \frac{[p_t]_{k+1} \left( q_{t,k+1}(c^\ast) / \hat{q}'(c^\ast) \right)}{\left( q_{k+1}'(c^\ast) / \hat{q}'(c^\ast) \right)} + O \left( \sqrt{\frac{\log(4/\delta)}{n}} + \sqrt{\frac{\log(4/\delta)}{m}} \right).
\]

**Proof.** We can simply prove this theorem as Corollary of Theorem 1 from Garg et al. (2021b). Note that $q_{k+1}(c^\ast) = \left( 1 - p_t(y = k + 1) \right) \cdot q_t(y = k + 1) \cdot q_{t,k+1}(c^\ast)$. Adding and subtracting $\left( 1 - p_t(y = k + 1) \right) \cdot q_t(y = k + 1)$ and dividing by $\hat{q}'_t$, we get $q_{k+1}(c^\ast) = \left( 1 - p_t(y = k + 1) \right) \cdot \left[ q'_t(c^\ast) - \hat{q}'(c^\ast) \right] + \left( 1 - p_t(y = k + 1) \right) + p_t(y = k + 1) \cdot \frac{q_{t,k+1}(c^\ast)}{q_{k+1}'(c^\ast)}$. Plugging in bound for LHS from Theorem 1 in Garg et al. (2021b), we get the desired result.

**F.4. Extensions of Theorem 2 to general separable datasets**

For general separable datasets, CVIR has undesirable property of getting stuck at local optima where gradient in (49) can be zero by maximizing entropy on the subset $P_o$ which is (incorrectly) not-rejected from $p_{o}$ in CVIR iterations. Intuitively, if the classifier can perfectly separate $P \setminus P_o$ and $N \setminus N_o$ and at the same time maximize the entropy of the region $P_o$, then the classifier trained with CVIR can get stuck in this local minima.

However, we can extend the above analysis with some modifications to the CVIR procedure. Note that when the CVIR classifier maximizes the entropy on $P_o$, it makes an error on points in $P_o$. Since, we have access to the distribution $p_o$, we can add an additional regularized penalty to the CVIR loss that ensures that the converged classifier with CVIR correctly classifies all the points in $p_o$. With a large enough regularization constant for the supervised loss on $p_o$, we can dominate the gradient term in (49) which pushes CVIR classifier to correct decision boundary even on $P_o$ (instead of maximizing entropy).

We leave formal analysis of this conjecture for future work. Since we warm start CVIR training with a positive versus unlabeled classifier, if we obtain an initialization close enough to the true positive versus negative decision boundary, by monotonicity property of CVIR iterations, we may never get stuck in such a local minima even without modifications to loss.

**G. Empirical investigation of CVIR in toy setup**

As noted in our ablation experiments and in Garg et al. (2021b), domain discriminator trained with CVIR outperforms classifiers trained with other consistent objectives (nnPU (Kiryo et al., 2017) and uPU (Du Plessis et al., 2015)). While the analysis in Sec. 5 highlights consistency of CVIR procedure in population, it doesn’t capture the observed empirical efficacy of CVIR over alternative methods in overparameterized models. In the Gaussian setup described in Sec. F.2, we train overparameterized linear models to compare CVIR with other methods (Fig. 3). We fix $d = 1000$ and use $n = 250$ positive and $m = 250$ unlabeled points for training with $\alpha = 0.5$. We set the margin $\gamma$ at 0.05. We compare CVIR with unbiased losses uPU and nnPU. We also make comparison with a naive positive versus unlabeled classifier (referred to as P2U). For CVIR, we experiment with a randomly initialized classifier and initialized with a P2U classifier trained for 200 epochs.

First, we observe that when a classifier is trained to distinguish positive and unlabeled data, early learning happens (Liu et al., 2020; Arora et al., 2019; Garg et al., 2021a), i.e., during the initial phase of learning classifier learns to classify positives in unlabeled correctly as positives achieving high accuracy on validation positive versus negative data. While the
Figure 3: **Comparison of different methods in overparameterized toy setup.** CVIR (random) denotes CVIR with random initialization and CVIR (PvU) denotes warm start with a positive versus negative classifier. Vertical line denotes the epoch at which we switch from PvU to CVIR in CVIR (PvU) training. (a) We observe that CVIR (PvU) improves significantly even over the best early stopped PvU model. As training proceeds, we observe that accuracy of nnPU, uPU and PvU training drops whereas CVIR (random) and CVIR (PvU) maintains superior and stable performance. (b) We observe that warm start training helps CVIR over randomly initialized model to correctly identity positives among unlabeled for rejection.

H. Experimental Details

H.1. Baselines

We compare PULSE with several popular methods from OSDA literature. While these methods are not specifically proposed for OSLS, they are introduced for the more general OSDA problem. In particular, we make comparisons with DANCE (Saito et al., 2020), UAN (You et al., 2019), CMU (Fu et al., 2020), STA (Liu et al., 2019), Backprop-ODA (or BODA) (Saito et al., 2018). We use the open source implementation available at https://github.com/thuml and https://github.com/VisionLearningGroup/DANCE/. Since OSDA methods do not estimate the prevalence of novel class explicitly, we use the fraction of examples predicted in class \( k+1 \) as a surrogate. We next briefly describe the main idea for each method:

*Backprob-ODA*  Saito et al. (2018) proposed backprob ODA to train a \((k + 1)\)-way classifier. In particular, the network is trained to correctly classify source samples and for target samples, the classifier (specifically the last layer) is trained to output 0.5 for the probability of the unknown class. The feature extractor is trained adversarially to move the probability of unknown class away from 0.5 on target examples by utilizing the gradient reversal layer.

*Separate-To-Adapt (STA)*  Liu et al. (2019) trained a network that learns jointly from source and target by learning to separate negative (novel) examples from target. The training is divided into two parts. The first part consists of training a multi-binary \( G_{\mathcal{C}} |_{c=1}^{C} \) classifier on labeled source data for each class and a binary classifier \( G_b \) which generates the weights \( w \) for rejecting target samples in the novel class. The second part consists of feature extractor \( G_f \), a classifier \( G_y \) and domain discriminator \( G_d \) to perform adversarial domain adaptation between source and target data in the source label space. \( G_y \) and \( G_d \) are trained with incorporating weights \( w \) predicted by \( G_b \) in the first stage.

*Calibrated Multiple Uncertainties (CMU)*  Fu et al. (2020) trained a source classifier and a domain discriminator to discriminate the novel class from previously seen classes in target. To train the discriminator network, CMU uses a weighted
We conduct experiments with seven benchmark classification datasets across vision, natural language, biology and medicine. We also train a domain discriminator classifier for source versus target (referred to as k-way source classifier). After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as $0$. We experiment with Newsgroups-20 dataset. Additionally, inspired by applications of OSLS in biology and medicine, we experiment with Entity30 (Santurkar et al., 2021). For Entity30, we randomly select 24 classes as the source classes and a novel class formed by aggregating the data from 6 remaining classes. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as $0.3942$.

**DANCE** Saito et al. (2020) proposed DANCE which combines a self-supervised clustering loss to cluster neighboring target examples and an entropy separation loss to consider alignment with source. Similar to CMU, during test time, DANCE uses thresholded prediction entropy of the source classifier to classify a target example as belonging to the novel class.

**Universal Adaptation Networks (UAN)** You et al. (2019) proposed UAN which also trains a source classifier and a domain discriminator to discriminate the novel class from previously seen classes in target. The objective is similar to CMU where instead of using uncertainty estimates from multiple classifiers, UAN uses prediction confidence of domain discriminator classifier. Similar to CMU, at test time, target data $x$ with $w(x) \leq w_0$ (for some pre-defined threshold $w_0$) is classified as an example from previously seen classes and is given a class prediction with source classifier. Otherwise, the target example is classified as belonging to the novel class.

For alternative baselines, we experiment with source classifier directly deployed on the target data which may contain novel class and label shift among source classes (referred to as source-only). This naive comparison is included to quantify benefits of label shift correction and identifying novel class over a typical $k$-way classifiers.

We also train a domain discriminator classifier for source versus target (referred to as domain disc.). This is an adaptation of PU learning baseline (Elkan and Noto, 2008) which assumes no label shift among source classes. We use simple domain discriminator training to distinguish source versus target. To estimate the fraction of novel examples, we use the EN estimator proposed in Elkan and Noto (2008). For any target input, we make a prediction with the domain discriminator classifier (after re-scaling the sigmoid output with the estimate proportion of novel examples). Any example that is classified as target, we assign it the class $k + 1$. For examples classified as source, we make a prediction for them using the $k$-way source classifier.

Finally, per the reduction presented in Sec. 2, we train $k$ PU classifiers (referred to as $k$-PU). To train each PU learning classifier, we can plug any method discussed in Sec. C. In the main paper, we included results obtained with plugin state-of-the-art PU learning algorithms. In App. H.8, we present ablations with other PU learning methods.

**H.2. Dataset and OSLS Setup Details**

We conduct experiments with seven benchmark classification datasets across vision, natural language, biology and medicine. Our datasets span language, image and table modalities. For each dataset, we simulate an OSLS problem. We experiment with different fraction of novel class prevalence, source label distribution, and target label distribution. We randomly choose classes that constitute the novel target class. After randomly choosing source and novel classes, we first split the training data from each source class randomly into two partitions. This creates a random label distribution for shared classes among source and target. We then club novel classes to assign them a new class (i.e. $k + 1$). Finally, we throw away labels for the target data to obtain an unsupervised DA problem. We repeat the same process on iid hold out data to obtain validation data with no target labels. For main experiments in the paper, we next describe important details for the OSLS setup simulated. All the other details can be found in the code repository.

For vision, we use CIFAR10, CIFAR100 (Krizhevsky and Hinton, 2009) and Entity30 (Santurkar et al., 2021). For language, we experiment with Newsgroups-20 dataset. Additionally, inspired by applications of OSLS in biology and medicine, we experiment with Tabula Muris (Consortium et al., 2020) (Gene Ontology prediction), Dermnet (skin disease prediction), and BreakHis (Spanhol et al., 2015) (tumor cell classification).

**CIFAR10** For CIFAR10, we randomly select 9 classes as the source classes and a novel class formed by the remaining class. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as $0.2152$.

**CIFAR100** For CIFAR100, we randomly select 85 classes as the source classes and a novel class formed by aggregating the data from 15 remaining classes. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as $0.2976$.

**Entity30** Entity30 is a subset of ImageNet (Russakovsky et al., 2015) with 30 super classes. For Entity30, we randomly select 24 classes as the source classes and a novel class formed by aggregating the data from 6 remaining classes. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as $0.3942$. 

**CMU** For CMU, we randomly select 2 classes as the source classes and a novel class formed by aggregating the data from 8 remaining classes. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as $0.8521$.

**ELAN** For ELAN, we randomly select 3 classes as the source classes and a novel class formed by aggregating the data from 9 remaining classes. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as $0.3942$.

**UAN** For UAN, we randomly select 2 classes as the source classes and a novel class formed by aggregating the data from 9 remaining classes. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as $0.3942$.
We train models till the performance on validation source data (labeled) ceases to increase. Unlike OSDA methods, we use Resnet18 (He et al., 2016) for CIFAR10, CIFAR100, and Entity30. For all three datasets, in our main experiments, Adam optimizer with a learning rate of 0.0001 and no weight decay. We train with constant learning rate for 200 epochs. We use SGD training with momentum of 0.9 and no weight decay. We train with constant learning rate for 200 epochs. We use a weight decay of $5 \times 10^{-4}$. For CIFAR100 and CIFAR10, we use batch size of 200. For Entity30, we use a batch size of 32. In App. H.7, we experiment with contrastive pre-training instead of random initialization.

For newsgroups, we use a convolutional architecture\(^5\). We use glove embeddings to initialize the embedding layer. We use Adam optimizer with a learning rate of 0.0001 and no weight decay. We use a batch size of 200. We train with constant learning rate for 120 epochs.

For Tabular Muris, we use the fully connected MLP used in Cao et al. (2021). We use the hyperparameters used in Cao et al. (2021). We use Adam optimizer with a learning rate of 0.0001 and no weight decay. We train with constant learning rate for 40 epochs. We use a batch size of 200.

For Dermnet and BreakHis, we use Resnet-50 pre-trained on Imagenet. We use an initial learning rate of 0.0001 and decay it by 0.96 every epoch. We use SGD training with momentum of 0.9 and weight decay of $5 \times 10^{-4}$. We use a batch size of 32. These are the default hyperparameters used in Alom et al. (2019) and Liao (2016).

For all methods, we use the same backbone for discriminator and source classifier. Additionally, for PULSE and domain disc., we use the exact same set of hyperparameters to train the domain discriminator and source classifier. For kPU, we use a separate final layer for each class with the same backbone. We use the same hyperparameters described above for all three methods. For OSDA methods, we use default method specific hyperparameters introduced in their works. Since we do not have access to labels from the target data, we do not perform hyperparameter tuning but instead use the standard hyperparameters used for training on labeled source data. In future, we may hope to leverage heuristics proposed for accuracy estimation without access to labeled target data (Garg et al., 2022).

We train models till the performance on validation source data (labeled) ceases to increase. Unlike OSDA methods, note that we do not use early stopping based on performance on held-out labeled target data. To evaluate classification performance, we report target accuracy on all classes, seen classes and the novel class. For target marginal, we separately report estimation error for previously seen classes and for the novel class. For the novel class, we report absolute difference between true and estimated marginal. For seen classes, we report average absolute estimation error. We open-source our

\(^1\)https://github.com/mireshghallah/20Newsgroups-
\(^2\)https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
\(^3\)http://www.dermnet.com/dermatology-pictures-skin-disease-pictures
\(^4\)https://github.com/snap-stanford/comet
\(^5\)https://github.com/mireshghallah/20Newsgroups-Pytorch

**Newgroups-20** For Newsgroups20\(^1\), we randomly select 16 classes as the source classes and a novel class formed by aggregating the data from 4 remaining classes. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as 0.3733. This dataset is motivated by scenarios where novel news categories can appear over time but the distribution of articles given a news category might stay relatively unchanged.

**BreakHis** BreakHis\(^2\) contains 8 categories of cell types, 4 types of benign breast tumor and 4 types malignant tumors (breast cancer). Here, we simulate OSLS problem specifically where 6 cell types are observed in the source (3 from each) and a novel class appears in the target with 1 cell type from each category. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as 0.2708.

**Dermnet** Dermnet data contains images of 23 types of skin diseases taken from Dermnet NZ\(^3\). We simulate OSLS problem specifically where 18 diseases are observed in the source and a novel class appears in the target with the rest of the 5 diseases. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as 0.3133.

**Tabula Muris** Tabula Muris dataset (Consortium et al., 2020) comprises of different cell types collected across 23 organs of the mouse model organism. We use the data pre-processing scripts provided in (Cao et al., 2021)\(^4\). We just use the training set comprising of 57 classes for our experiments. We simulate OSLS problem specifically where 28 cell types are observed in the source and a novel class appears in the target with the rest of the 29 cell types. After randomly sampling the label marginal for source and target randomly, we get the prevalence for novel class as 0.6366.
code at https://github.com/Neurips2022Anon. By simply changing a single config file, new OSLS setups can be generated and experimented with.

Note that for our main experiments, for vision datasets (i.e., CIFAR10, CIFAR100, and Entity30) and for language dataset, we do not initialize with a (supervised) pre-trained model to avoid overlap of novel classes with the classes in the dataset used for pre-training. For example, labeled Imaginet-1k is typically used for pre-training. However, Imaginet classes overlaps with all three vision datasets employed and hence, we avoid pre-trained initialization. In App. H.7, we experiment with contrastive pre-training on Entity30 and CIFAR100. In contrast, for medical datasets, we leverage Imaginet pre-trained models as there is no overlap between classes in BreakHis and Dermnet with Imaginet.

H.4. Detailed results from main paper

For completeness, we next include results for all datasets. In particular, for each dataset we tabulate (i) overall accuracy on target; (ii) accuracy on seen classes in target; (iii) accuracy on the novel class; (iv) sum of absolute error in estimating target marginal among previously seen classes, i.e., \( \sum_{y \in Y_s} |\hat{p}_t(y) - p_t(y)| \); and (v) absolute error for novel fraction estimation, i.e., \( |\hat{p}_t(y = k + 1) - p_t(y = k + 1)| \). Table 5 presents results on all the datasets. Fig. 4 and Fig. 5 presents epoch-wise results.

H.5. Investigation into OSDA approaches

We observe that with default hyperparameters, popular OSDA methods significantly under perform as compared to PULSE. We hypothesize that the primary reasons underlying the poor performance of OSDA methods are (i) the heuristics employed to detect novel classes; and (ii) loss functions incorporated to improve alignment between examples from common classes in source and target. To detect novel classes, a standard heuristic employed popular OSDA methods involves thresholding uncertainty estimates (e.g., prediction entropy, softmax confidence (You et al., 2019; Fu et al., 2020; Saito et al., 2020)) at a predefined threshold \( \kappa \). However, a fixed \( \kappa \) may not for different datasets and different fractions of the novel class. Here, we ablate by (i) removing loss function terms incorporated with an aim to improve source target alignment; and (ii) vary threshold \( \kappa \) and show improvements in performance of these methods.

For our investigations, we experiment with CIFAR10, with UAN and DANCE methods. For DANCE, we remove the entropy separation loss employed to encourage align target examples with source examples. For UAN, we remove the adversarial domain discriminator training employed to align target examples with source examples. For both the methods, we observe that by removing the corresponding loss function terms we obtain a marginal improvement. For DANCE on CIFAR10, the performance goes up from 70.4 to 72.5 (with the same hyperparameters as the default run). FOR UAN, we observe similar minor improvements, where the performance goes up from 15.4 to 19.6.

Next, we vary the threshold used for detecting the novel examples. By optimally tuning the threshold for CIFAR10 with UAN, we obtain a substantial increase. In particular, the overall target accuracy increases from 19.6 to 33.1. With DANCE on CIFAR10, optimal threshold achieves 75.6 as compared to the default accuracy 70.4. In contrast, our two-stage method PULSE avoids the need to guess \( \kappa \), by first estimating the fraction of novel class which then guides the classification of novel class versus previously seen classes.

H.6. Ablation with novel class fraction

In this section, we ablate on novel class proportion on CIFAR10, CIFAR100 and Newsgroups20. For each dataset we experiment with three settings, each obtained by varying the number of classes from the original data that constitutes the novel classes. We tabulate our results in Table 4.

H.7. Contrastive pre-training on unlabeled data

Here, we experiment with contrastive pre-training to pre-train the backbone networks used for feature extraction. In particular, we initialize the backbone architectures with SimCLR pre-trained weights. We experiment with CIFAR100 and Entity30 datasets. Instead of pre-training on mixture of source and target unlabeled data, we leverage the publicly available pre-trained weights\(^6\). Table 2 summarizes our results. We observe that pre-training improves over random initialization for all the methods with PULSE continuing to outperform other approaches.

\(^6\)For CIFAR100: https://drive.google.com/file/d/1huW-ChBVvKcx7t8HyDaWTQB5Li1Fht9x/view and for Entity30, we use Imaginet pre-trained weights from here: https://github.com/AndrewAtanov/simclr-pytorch.
Table 2: Comparison with different OSLS approaches with pre-trained feature extractor. We use SimCLR pre-training to initialize the feature extractor for all the methods. All methods improve over random initialization (in Table 1). Note that PULSE continues to outperform other approaches.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10 Acc (All)</th>
<th>CIFAR10 MPE (Novel)</th>
<th>Entity30 Acc (All)</th>
<th>Entity30 MPE (Novel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>37.1</td>
<td>0.34</td>
<td>52.1</td>
<td>0.376</td>
</tr>
<tr>
<td>Domain Disc.</td>
<td>49.4</td>
<td>0.041</td>
<td>57.4</td>
<td>0.024</td>
</tr>
<tr>
<td>kPU</td>
<td>37.5</td>
<td>0.297</td>
<td>70.1</td>
<td>0.32</td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>67.3</td>
<td>0.052</td>
<td>72.4</td>
<td>0.002</td>
</tr>
</tbody>
</table>

H.8. Ablation with different PU learning methods

In this section, we experiment with alternative PU learning approaches for PULSE and kPU. In particular, we experiment with the next best alternatives, i.e., nnPU instead of CVIR for classification and DEDPUL instead of BBE for target marginal estimation. We refer to these as kPU (alternative) and PULSE (alternative) in Table 3. We present results on three datasets: CIFAR10, CIFAR100 and Newsgroups20 in the same setting as described in Sec. H.2. We make two key observations: (i) PULSE continues to dominate kPU with alternative choices; (ii) CVIR and BBE significantly outperform alternative choices.

Table 3: Comparison with different PU learning approaches. ‘Alternative’ denotes results with employing nnPU for classification and DEDPUL for target marginal estimation instead of ‘default’ which uses CVIR and BBE.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10 Acc (All)</th>
<th>CIFAR10 MPE (Novel)</th>
<th>CIFAR100 Acc (All)</th>
<th>CIFAR100 MPE (Novel)</th>
<th>Newsgroups20 Acc (All)</th>
<th>Newsgroups20 MPE (Novel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-PU (alternative)</td>
<td>53.4</td>
<td>0.215</td>
<td>12.1</td>
<td>0.298</td>
<td>14.1</td>
<td>0.373</td>
</tr>
<tr>
<td>k-PU (default)</td>
<td>83.6</td>
<td>0.036</td>
<td>36.3</td>
<td>0.298</td>
<td>52.1</td>
<td>0.307</td>
</tr>
<tr>
<td>PULSE (alternative)</td>
<td>80.5</td>
<td>0.05</td>
<td>30.1</td>
<td>0.231</td>
<td>39.8</td>
<td>0.223</td>
</tr>
<tr>
<td>PULSE (default)</td>
<td>86.1</td>
<td>0.008</td>
<td>63.4</td>
<td>0.078</td>
<td>62.2</td>
<td>0.061</td>
</tr>
</tbody>
</table>
Table 4: Comparison with different OSLS approaches for different novel class prevalence. We observe that for on CIFAR100 and Newsgroups20, PULSE maintains superior performance as compared to other approaches. On CIFAR10, as the proportion of novel class increases, the performance of of kPU improves slightly over PULSE for target accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10 ($p_t(k+1) = 0.215$)</th>
<th>CIFAR10 ($p_t(k+1) = 0.406$)</th>
<th>CIFAR10 ($p_t(k+1) = 0.583$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc (All) MPE (Novel)</td>
<td>Acc (All) MPE (Novel)</td>
<td>Acc (All) MPE (Novel)</td>
</tr>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>63.1 0.162</td>
<td>65.5 0.166</td>
<td>48.6 0.265</td>
</tr>
<tr>
<td>Domain Disc.</td>
<td>47.4 0.331</td>
<td>57.5 0.232</td>
<td>68.7 0.144</td>
</tr>
<tr>
<td>kPU</td>
<td>83.6 0.036</td>
<td>87.8 0.010</td>
<td>89.9 0.036</td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>86.1 0.008</td>
<td>87.4 0.009</td>
<td>83.7 0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR100 ($p_t(k+1) = 0.2976$)</th>
<th>CIFAR100 ($p_t(k+1) = 0.4477$)</th>
<th>CIFAR100 ($p_t(k+1) = 0.5676$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc (All) MPE (Novel)</td>
<td>Acc (All) MPE (Novel)</td>
<td>Acc (All) MPE (Novel)</td>
</tr>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>36.1 0.41</td>
<td>41.6 0.075</td>
<td>50.2 0.03</td>
</tr>
<tr>
<td>Domain Disc.</td>
<td>45.8 0.046</td>
<td>52.3 0.092</td>
<td>58.7 0.187</td>
</tr>
<tr>
<td>kPU</td>
<td>36.3 0.298</td>
<td>52.2 0.448</td>
<td>63.9 0.568</td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>63.4 0.078</td>
<td>66.6 0.052</td>
<td>68.2 0.088</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Newsgroups20 ($p_t(k+1) = 0.3733$)</th>
<th>Newsgroups20 ($p_t(k+1) = 0.6452$)</th>
<th>Newsgroups20 ($p_t(k+1) = 0.7688$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc (All) MPE (Novel)</td>
<td>Acc (All) MPE (Novel)</td>
<td>Acc (All) MPE (Novel)</td>
</tr>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>43.4 0.16</td>
<td>25.5 0.645</td>
<td>17.7 0.769</td>
</tr>
<tr>
<td>Domain Disc.</td>
<td>50.9 0.176</td>
<td>44.8 0.085</td>
<td>47.8 0.064</td>
</tr>
<tr>
<td>kPU</td>
<td>52.1 0.373</td>
<td>50.2 0.645</td>
<td>35.5 0.769</td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>62.2 0.061</td>
<td>71.7 0.044</td>
<td>75.73 0.179</td>
</tr>
</tbody>
</table>
Table 5: *Comparison of PULSE with other methods.* Across all datasets, PULSE outperforms alternatives for both target classification and novel class prevalence estimation. Acc (All) is target accuracy, Acc (Seen) is target accuracy on examples from previously seen classes, and Acc (Novel) is recall for novel examples. MPE (Seen) is sum of absolute error for estimating target marginal among previously seen classes and MPE (Novel) is absolute error for novel prevalence estimation. Results reported by averaging across 3 seeds.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR-10</th>
<th></th>
<th></th>
<th>CIFAR-100</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc (All)</td>
<td>Acc (Seen)</td>
<td>Acc (Novel)</td>
<td>MPE (Seen)</td>
<td>MPE (Novel)</td>
<td>MPE (All)</td>
</tr>
<tr>
<td>Source-Only</td>
<td>67.1</td>
<td>87.0</td>
<td>-</td>
<td>46.6</td>
<td>66.4</td>
<td>-</td>
</tr>
<tr>
<td>UAN (You et al., 2019)</td>
<td>15.4</td>
<td>19.7</td>
<td>25.2</td>
<td>1.44</td>
<td>0.214</td>
<td>18.1</td>
</tr>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>63.1</td>
<td>66.2</td>
<td>42.0</td>
<td>0.541</td>
<td>0.162</td>
<td>36.1</td>
</tr>
<tr>
<td>DANCE (Saito et al., 2020)</td>
<td>70.4</td>
<td>85.5</td>
<td>14.5</td>
<td>0.784</td>
<td>0.174</td>
<td>47.3</td>
</tr>
<tr>
<td>STA (Liu et al., 2019)</td>
<td>57.9</td>
<td>69.6</td>
<td>14.9</td>
<td>0.409</td>
<td>0.124</td>
<td>42.6</td>
</tr>
<tr>
<td>CMU (Fu et al., 2020)</td>
<td>62.1</td>
<td>77.9</td>
<td>41.2</td>
<td>0.443</td>
<td>0.183</td>
<td>35.4</td>
</tr>
<tr>
<td>Domain Disc.</td>
<td>47.4</td>
<td>87.0</td>
<td>30.6</td>
<td>-</td>
<td>0.331</td>
<td>45.8</td>
</tr>
<tr>
<td>k-PU</td>
<td>83.6</td>
<td>79.4</td>
<td>-</td>
<td>98.9</td>
<td>0.062</td>
<td>99.1</td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>86.1</td>
<td>91.8</td>
<td>88.4</td>
<td>0.091</td>
<td>0.008</td>
<td>63.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Newsgroup20</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc (All)</td>
<td>Acc (Seen)</td>
<td>Acc (Novel)</td>
<td>MPE (Seen)</td>
<td>MPE (Novel)</td>
<td>MPE (All)</td>
</tr>
<tr>
<td>Source-Only</td>
<td>32.0</td>
<td>53.5</td>
<td>-</td>
<td>39.3</td>
<td>64.4</td>
<td>-</td>
</tr>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>42.22</td>
<td>25.9</td>
<td>67.2</td>
<td>0.367</td>
<td>0.189</td>
<td>43.4</td>
</tr>
<tr>
<td>Domain Disc.</td>
<td>43.2</td>
<td>53.5</td>
<td>68.0</td>
<td>-</td>
<td>0.135</td>
<td>50.9</td>
</tr>
<tr>
<td>k-PU</td>
<td>50.7</td>
<td>22.3</td>
<td>-</td>
<td>94.4</td>
<td>0.99</td>
<td>52.1</td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>58.0</td>
<td>54.3</td>
<td>72.2</td>
<td>0.215</td>
<td>0.054</td>
<td>62.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Tabula Muris</th>
<th></th>
<th></th>
<th>BreakHis</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc (All)</td>
<td>Acc (Seen)</td>
<td>Acc (Novel)</td>
<td>MPE (Seen)</td>
<td>MPE (Novel)</td>
<td>MPE (All)</td>
</tr>
<tr>
<td>Source-Only</td>
<td>33.8</td>
<td>93.3</td>
<td>-</td>
<td>-</td>
<td>70.0</td>
<td>95.8</td>
</tr>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>76.5</td>
<td>59.8</td>
<td>87.0</td>
<td>0.200</td>
<td>0.079</td>
<td>71.5</td>
</tr>
<tr>
<td>Domain Disc.</td>
<td>73.0</td>
<td>93.3</td>
<td>-</td>
<td>94.7</td>
<td>0.071</td>
<td>56.5</td>
</tr>
<tr>
<td>k-PU</td>
<td>85.9</td>
<td>91.6</td>
<td>83.3</td>
<td>0.279</td>
<td>0.307</td>
<td>75.6</td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>87.8</td>
<td>94.6</td>
<td>88.8</td>
<td>0.388</td>
<td>0.058</td>
<td>79.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Dermnet</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc (All)</td>
<td>Acc (Seen)</td>
<td>Acc (Novel)</td>
<td>MPE (Seen)</td>
<td>MPE (Novel)</td>
<td>MPE (All)</td>
</tr>
<tr>
<td>Source-Only</td>
<td>41.4</td>
<td>53.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BODA (Saito et al., 2018)</td>
<td>43.8</td>
<td>31.4</td>
<td>58.4</td>
<td>0.401</td>
<td>0.207</td>
<td></td>
</tr>
<tr>
<td>Domain Disc.</td>
<td>40.6</td>
<td>53.6</td>
<td>82.7</td>
<td>-</td>
<td>-</td>
<td>0.083</td>
</tr>
<tr>
<td>k-PU</td>
<td>46.0</td>
<td>26.0</td>
<td>89.9</td>
<td>1.44</td>
<td>0.313</td>
<td></td>
</tr>
<tr>
<td>PULSE (Ours)</td>
<td>48.9</td>
<td>53.7</td>
<td>57.7</td>
<td>0.41</td>
<td>0.043</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4: **Epoch wise results for target accuracy.** Results aggregated over 3 seeds. PULSE maintains stable and superior performance when compared to alternative methods.
Figure 5: Epoch wise results for novel prevalence estimation. Results aggregated over 3 seeds. PULSE maintains stable and superior performance when compared to alternative methods.