Mixture Proportion Estimation and PU Learning: A Modern Approach

Saurabh Garg [1] **Yifan Wu** [1] **Alex Smola** [2] **Sivaraman Balakrishnan** [1 3] **Zachary C. Lipton** [1]

Abstract

Given only positive examples and unlabeled examples (from both positive and negative classes), we might hope nevertheless to estimate an accurate positive-versus-negative classifier. Formally, this task is broken down into two subtasks: (i) *Mixture Proportion Estimation* (MPE)—determining the fraction of positive examples in the unlabeled data; and (ii) *PU-learning*—given such an estimate, learning the desired positive-versus-negative classifier. Unfortunately, classical methods for both problems break down in high-dimensional settings. Meanwhile, recently proposed heuristics lack theoretical coherence and depend precariously on hyperparameter tuning. In this paper, we propose two simple techniques: *Best Bin Estimation* (BBE) (for MPE); and *Conditional Value Under Optimism* (CVuO), a simple objective for PU-learning. Both methods dominate previous approaches empirically, and for BBE, we establish formal guarantees that hold whenever we can train a model to cleanly separate out a small subset of positive examples. Our final algorithm (TED)^n, alternates between the two procedures, significantly improving both our mixture proportion estimator and classifier.

1. Introduction

When deploying k-way classifiers in the wild, what can we do when confronted with data from a previously unseen class ($y = k + 1$)? Theory dictates that learning under distribution shift is impossible absent assumptions. Yet people exhibit this capability routinely. Faced with new symptoms, doctors can recognize the presence of a previously unseen ailment and attempt to estimate its prevalence. Similarly, naturalists can discover new species, estimate their population, and recognize them reliably going forward.

[1]Machine Learning Department, Carnegie Mellon University [2]Amazon Web Services [3]Department of Statistics and Data Science, Carnegie Mellon University. Correspondence to: Saurabh Garg <sgarg2@andrew.cmu.edu>.

Preliminary work. Under review by the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning. Do not distribute.

To begin making this problem tractable, we might make the label shift assumption (Saerens et al., 2002; Storkey, 2009; Lipton et al., 2018), i.e., that while the class balance $p(y)$ can change, the class conditional distributions $p(x|y)$ do not. Moreover, we might begin by focusing on the base case, where only one class has been seen previously, i.e., $k = 1$. Here, we possess (labeled) positive data from the source distribution, and (unlabeled) data from the target distribution, consisting of both positive and negative instances. This problem has been studied in the literature as *learning from positive and unlabeled data* (De Comité et al., 1999; Letouzey et al., 2000) and has typically been broken down into two subtasks: (i) Mixture Proportion Estimation (MPE) where we estimate α, the fraction of positives among the unlabeled examples; and (ii) PU-learning where this estimate is incorporated into a scheme for learning a Positive-versus-Negative (PvN) binary classifier.

Many methods have been proposed for both MPE (Elkan & Noto, 2008; Ramaswamy et al., 2016; Jain et al., 2016; Bekker & Davis, 2018; Ivanov, 2019) and PU-learning (Du Plessis et al., 2015; Kiryo et al., 2017). However, classical MPE methods break down in high-dimensional settings (Ramaswamy et al., 2016) or yield estimators whose accuracy depends on restrictive conditions (Du Plessis & Sugiyama, 2014a; Scott, 2015). On the other hand, most recent proposals either lack theoretical coherence, rely on heroic assumptions, or depend precariously on tuning hyperparameters that are, by the very problem setting, untunable. For PU learning, Elkan & Noto (2008) suggest training a classifier to distinguish positive from unlabeled data followed by a rescaling procedure. Du Plessis et al. (2015) suggest an unbiased risk estimation framework for PU learning. However, these methods fail badly with large model classes capable of overfitting and thus implementations on high-dimensional datasets rely on hyperparameter tuning and ad-hoc heuristics that do not transport effectively across datasets.

In this paper, we propose (i) Best Bin Estimation (BBE), an effective technique for MPE that produces consistent estimates $\widehat{\alpha}$ under mild assumptions and admits finite-sample statistical guarantees achieving the desired $O(1/\sqrt{n})$ rates; and (ii) learning with the Conditional Value under Optimism (CVuO) objective, which discards the worst $\widehat{\alpha}$ fraction of examples on each training epoch, removing the incentive to

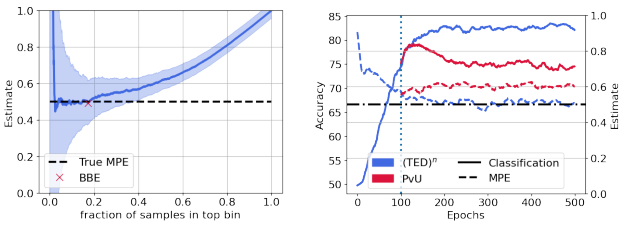


Figure 1. **(left)** BBE and estimate of α with varying fraction of unlabeled examples in the top bin at 200-th epoch of (TED)^n training. **(right)** Accuracy and BBE estimate as training proceeds. Till 100-th epoch (vertical line), we perform PvU training, i.e. warm start for (TED)^n. Post 100-th epoch, we continue with both (TED)^n and PvU training. Clearly as training proceeds with (TED)^n improves both classification accuracy and MPE over PvU training. Results with Resnet-18 on binary-CIFAR.

overfit to the unlabeled positive examples. Both methods are simple to implement, compatible with arbitrary hypothesis classes (including deep networks), and dominate existing methods in our experimental evaluation. Finally, we combine the two in an iterated Transform-Estimate-Discard (TED)^n framework that significantly improves both estimate MSE and classifier error.

We build on label shift methods (Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2019; Garg et al., 2020), that leverage black-box classifiers to reduce dimensionality. While label shift methods rely on classifiers trained to separate previously seen classes, BBE exploits a Positive-versus-Unlabeled (PvU) target classifier, which gives each input a score indicating how likely it is to be a positive sample. In particular, BBE identifies a threshold such that by estimating the ratio between the fractions of positive and unlabeled points receiving scores above the threshold, we obtain an estimate of the mixture proportion α. BBE works because in practice, for many datasets, PvU classifiers, even when uncalibrated, produce outputs with near monotonic calibration diagrams. We show that the existence of a (nearly) pure top bin is sufficient to produce a (nearly) consistent estimate $\widehat{\alpha}$, whose finite-sample convergence rates depend on the fraction of examples in the bin and whose bias depends on the *purity* of the bin. Crucially, we estimate the optimal threshold from data.

We conduct a battery of experiments both to empirically validate our claim that BBE's assumptions are mild and frequently hold in practice. We conduct extensive experiments on semi-synthetic data, adapting a variety of binary classification datasets to the PU learning setup and demonstrating the superior performance of BBE and PU-learning with the CVuO objective. Moreover, we show that (TED)^n, which combines the two in an iterative fashion improves significantly over previous methods across several architectures on a range of image and text datasets.

2. Problem Setup

For an event E, we let $\mathbb{I}[E]$ denote the binary indicator of the event. By $|A|$, we denote the cardinality of set A. Let $\mathcal{X} \in \mathbb{R}^d$ be the input space and $\mathcal{Y} = \{-1, +1\}$ be the output space. Let $\mathrm{P} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be the underlying joint distribution and let p denote its corresponding density.

Let P_p and P_n be the class-conditional distributions for positive and negative class and $p_p(x) = p(x|y = +1)$ and $p_n(x) = p(x|y = -1)$ be the corresponding class-conditional densities. P_u denotes the distribution of the unlabeled data and p_u denotes its density. Let $\alpha \in [0, 1]$ be the fraction of positives among the unlabeled population, i.e., $\mathrm{P}_u = \alpha \mathrm{P}_p + (1 - \alpha)\mathrm{P}_n$. When learning from positive and unlabeled data, we obtain i.i.d. samples from the positive (class-conditional) distribution, which we denote as $X_p = \{x_1, x_2, \ldots, x_{n_p}\} \sim \mathrm{P}_p^{n_p}$ and i.i.d samples from unlabeled distribution as $X_u = \{x_{n_p+1}, x_{n_p+2}, \ldots, x_{n_p+n_u}\} \sim \mathrm{P}_u^{n_u}$.

MPE is the problem of estimating α. Absent assumptions on P_p, P_n and P_u, the mixture proportion α is not identifiable (ref. App. B). In this paper, we propose mild conditions on the PvU classifier that make α identifiable and allows us to derive finite-sample convergence guarantees. With PU learning, the aim is to learn a classifier $f : \mathcal{X} \to [0, 1]$ to approximate $p(y = +1|x)$. We assume that we are given a loss function $\ell : [0, 1] \times \mathcal{Y} \to \mathbb{R}$, such that $\ell(z, y)$ is the loss incurred by predicting z when the true label is y. For a classifier f and a sampled set $X = \{x_1, x_2, \ldots, x_n\}$, we let $\widehat{L}^+(f; X) = \sum_{i=1}^n \ell(f(x_i), +1)/n$ denote the loss when predicting the samples as positive and $\widehat{L}^-(f; X) = \sum_{i=1}^n \ell(f(x_i), -1)/n$ the loss when predicting the samples as negative. For a sample set X each with true label y, we define 0-1 error as $\widehat{\mathcal{E}}^y(f; X) = \sum_{i=1}^n \mathbb{I}[y(f(x_i) - t) \leqslant 0]/n$ for some predefined threshold t . Unless stated otherwise, the threshold is assumed to be 0.5.

3. Mixture Proportion Estimation

In this section, we propose BBE, a new method that leverages a blackbox classifier f to perform MPE and establish convergence guarantees. All proofs are relegated to App. D. To begin, we assume access to a fixed classifier f. For intuition, one may think of f as a PvU classifier. Later, we discuss ways to obtain a suitable classifier from PU data.

We now introduce some additional notation. Assume f transforms an input $x \in \mathcal{X}$ to $z \in [0, 1]$, i.e., $z = f(x)$. For given probability density function p and a classifier f, define a function $q(z) = \int_{A_z} p(x)dx$, where $A_z = \{x \in \mathcal{X} : f(x) \geqslant z\}$ 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 classifier f in the transformed space. We now define an empirical estimator $\widehat{q}(z)$ given a set $X = \{x_1, x_2, \ldots, x_n\}$

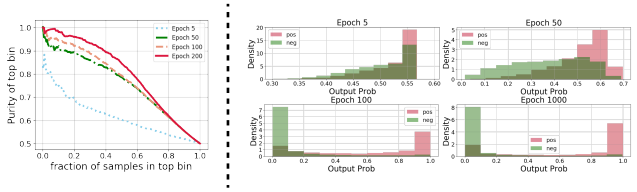


Figure 2. **(left)** Purity and size (in terms of fraction of unlabeled samples) in top bin and **(right)** Distribution of predicted probabilities (of being positive) for unlabeled training data with (TED)^n. Results with ResNet-18 on binary-CIFAR.

sampled iid from $p(x)$. Let $Z = f(X)$. Define $\widehat{q}(z) = \sum_{i=1}^{n} \mathbb{I}\left[z_i \geqslant z\right]/n$. For each pdf p_p, p_n and p_u, we define q_p, q_n and q_u respectively.

Without any assumptions on the underlying distribution and the classifier f, we aim to estimate $\alpha^* = \min_{c\in[0,1]} q_u(c)/q_p(c)$. Under one mild assumption that empirically holds across numerous PU datasets, we show that $\alpha^* = \alpha$. We now explain our procedure. First, given a held-out dataset of positive (X_p) and unlabeled examples (X_u), we push all examples through the classifier f to obtain one-dimensional outputs $Z_p = f(X_p)$ and $Z_u = f(X_u)$. Next, with Z_p and Z_u, we estimate $\widehat{q}_p$ and $\widehat{q}_u$. Finally, we return the ratio $\widehat{q}_u(\widehat{c})/\widehat{q}_p(\widehat{c})$ at $\widehat{c}$ that minimizes the upper confidence bound at a pre-specified level δ (calculated using Lemma 1). BBE is summarized in Algorithm 1. For theoretical guarantees, we add an additional confidence bound term, however in practice, our algorithm always returns the ratio. Refer to App. D.1 for details. We now show that BBE comes with the following guarantee:

Theorem 1. *Let $q_p(c) \geqslant q_n(c)$ for all $c \in [0,1]$. Define $c^* = \arg\min_{c\in[0,1]} q_u(c)/q_p(c)$ and assume $n_p \geqslant \sqrt{2\log(4/\delta)/q_p(c^*)}$. For every $\delta > 0$, the mixture proportion estimator $\widehat{\alpha}$ defined in Algorithm 1 satisfies with probability $1-\delta$: $\alpha^* - c_1 \cdot \left(\sqrt{\frac{\log(4/\delta)}{n_u}} + \sqrt{\frac{\log(4/\delta)}{n_p}}\right) \leqslant \widehat{\alpha}$ and $\widehat{\alpha} \leqslant \alpha^* + \frac{c_2}{q_p(c^*)} \cdot \left(2\sqrt{\frac{\log(4/\delta)}{n_u}} + (1+\alpha^*)\sqrt{\frac{\log(4/\delta)}{n_p}}\right)$, for some constant $c_1, c_2 \geqslant 0$.*

Theorem 1 shows that with high probability, our estimate is close to α^*. As a corollary to Theorem 1 (in App. D), we show that our estimator $\widehat{\alpha}$ converges to the true α with convergence rate $\min(n_p, n_u)^{-1/2}$, as long as there exist a threshold $c_f \in (0,1)$ such that $q_p(c_f) \geqslant \epsilon_p$ and $q_n(c_f) = 0$ for some constant $\epsilon_p > 0$. We refer to this condition as the *pure positive bin* property. Note that in a more general case, our bound in Corollary 1 captures the tradeoff due to the proportion of negative examples in the top bin (bias) versus the proportion of positives in the top bin (variance).

Empirical Validation We now empirically validate the positive pure top bin property (Fig. 2). We observe that as

Algorithm 1 Best Bin Estimation (BBE)

input Validation positive (X_p) and unlabeled (X_u) samples. Blackbox model classifier $\widehat{f} : \mathcal{X} \to [0,1]$. Hyperparameter $0 < \delta, \gamma < 1$.

1: $Z_p, Z_u = f(X_p), f(X_u)$.

2: $\widehat{q}_u(z), \widehat{q}_p(z) = \frac{\sum_{z_i\in Z_p} \mathbb{I}[z_i \geqslant z]}{n_p}, \frac{\sum_{z_i\in Z_u} \mathbb{I}[z_i \geqslant z]}{n_u}$ for all $z \in [0,1]$.

3: $\widehat{c} = \arg\min_{c\in[0,1]} \left(\frac{\widehat{q}_u(c)}{\widehat{q}_p(c)} + \frac{1}{\widehat{q}_p(c)}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right)\right)$

output $\frac{\widehat{q}_u(\widehat{c})}{\widehat{q}_p(\widehat{c})} + \left(\frac{1}{\widehat{q}_p(c)} - \frac{1}{\gamma}\right)_+ \left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right)$

PvU training proceeds, purity of the top bin improves for a fixed fraction of samples in the top bin. Moreover, this behavior gets more pronounced when learning the classifier with the CVuO objective proposed in the next section.

Comparison with existing methods Jain et al. (2016) and Ivanov (2019) discuss Bayes optimality of the PvU classifier (or its one-to-one mapping) as a sufficient condition to preserve α in transformed space. By contrast, we show that the milder pure positive bin property is sufficient to guarantee consistency and achieve $\mathcal{O}(1/\sqrt{n})$ rates. Furthermore, in a simple toy setup (in App. E), we show that even when the hypothesis class is well specified for PvN learning, PvU training will not recover the Bayes-optimal scoring function, even in population. Moreover, we show that any monotonic mapping of the Bayes-optimal PvU scoring function induces a positive pure top bin property.

4. PU-Learning

In supervised learning with separable data (e.g., cleanly labeled image data), overparameterized models generalize well even after achieving near-zero training error. However, with PvU training over-parameterized models can memorize the unlabeled positives, assigning them confidently to the negative class, which can severely hurt generalization on PN data (Zhang et al., 2016). Moreover, while unbiased losses exist that estimate the PvN loss given PU data and the mixture proportion α, this unbiasedness only holds before the loss is optimized, and becomes ineffective with powerful deep learning models capable of memorization.

A variety of heuristics, including ad-hoc early stopping criteria, have been explored (Ivanov, 2019), where training proceeds until the loss on unseen PU data ceases to decrease. However, this approach leads to severe under-fitting (results in App. I.2). On the other hand, by regularizing the loss function, nnPU (Kiryo et al., 2017) mitigates overfitting issues due to memorization. However, we observe that nnPU still leaves a substantial accuracy gap when compared to a model trained just on the positive and negative (from the unlabeled) data (ref. experiment in App. I.1).

Algorithm 2 PU learning with Conditional Value under Optimism (CVuO) objective

input Labeled positive training data (X_p) and unlabeled training samples (X_u). Mixture proportion estimate α.

Initialize a model f_θ and an optimization algorithm $\mathcal{A}$.
$X_n := X_u$.
while $\widehat{\mathcal{E}}^+(f_\theta; X_p) + \widehat{\mathcal{E}}^-(f_\theta; X_n)$ not converged **do**
  Rank samples $x_u \in X_u$ according to their loss values $\ell(f_\theta(x_u), -1)$.
  $X_n := X_{u,1-\alpha}$ where $X_{u,1-\alpha}$ denote the lowest ranked $1-\alpha$ fraction of samples.
  Shuffle (X_p, X_n) into B mini-batches. With (X_p^i, X_n^i) we denote i-th mini-batch.
  for $i = 1$ to B **do**
    Set the gradient $\nabla_\theta \left[\widehat{L}^+(f_\theta; X_p^i) + \widehat{L}^-(f_\theta; X_n^i)\right]$ and update θ with algorithm $\mathcal{A}$.
  end for
end while

output Trained classifier f_θ

This leads us to ask the following question: *can we improve performance over nnPU of a model just trained with PU data and bridge this gap?* In an ideal scenario, if we could identify and remove all the positive points from the unlabeled data during training then we can hope to achieve improved performance over nnPU. Indeed, in practice, we observe that in the initial stages of PvU training, the model assigns much higher scores to positives than to negatives in the unlabeled data (Fig. 2(right)).

Inspired by this observation, we propose CVuO, a simple yet effective objective for PU learning. First, we present our method assuming access to the true MPE. Given a training set of positives X_p and unlabeled X_u and the mixture proportion α, we begin by ranking the unlabeled data according the predicted probability (of being positive) by our classifier. Then, in every epoch of training, we create a set of provisionally negative samples X_n by removing α fraction of the unlabeled samples currently scored as most positive. Next, we update our classifier by minimize the loss on the positives X_p and provisional negatives X_n by treating them as negatives. We repeat this procedure until the training error on X_p and X_n converges. Summary in Algorithm 2. Likewise nnPU, note that this procedure does not need early stopping. In App. F, we justify our loss function in the scenario when the positives and negatives are separable.

(TED)^n Integrating BBE and CVuO We observe the interaction between BBE and CVuO objective. Accurate mixture proportion estimate leads to improved classifier, in particular, we reject accurate number of prospective positive samples from unlabeled. Consequently, updating the classifier to minimize loss on positive versus retained unlabeled improves purity of top bin. This leads to an obvious alternating procedure where at each epoch, we first use BBE to estimate $\widehat{\alpha}$ and then update the classifier with CVuO objective with $\widehat{\alpha}$ as input. We repeat this until training error has not converged. Our method Transfer, Estimate and Discard (TED)^n is summarized in Algorithm 3 (in App. C).

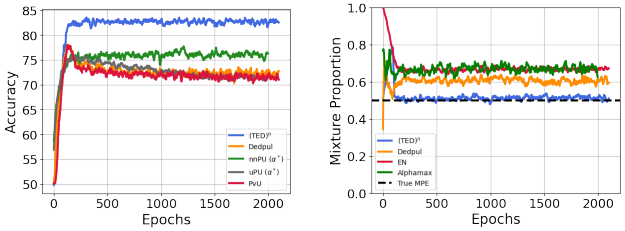


Figure 3. Epoch wise results with ResNet-18 trained on binary-CIFAR. Parallel results on other tasks in App. I.3. Clearly as training proceeds, (TED)^n maintains substantial improvements for both classification and MPE over existing methods.

5. Experiments

Having presented our PU learning and MPE algorithms, we now compare their performance with other methods empirically. We focus on UCI, vision and text datasets in our experiments. We simulate PU tasks on CIFAR-10 (Krizhevsky & Hinton, 2009), MNIST (LeCun et al., 1998), and IMDb sentiment analysis (Maas et al., 2011) datasets. We include results with ResNet-18 on binary CIFAR (i.e., first 5 classes vs rest) in the main paper and include all the other results in App. G and App. I. We include additional details on the datasets and architecture in App. H.

First, we discuss results for MPE (Table 1). We compare our method with KM2, TiCE, Dedpul, AlphaMax and EN (ref Table 1). With the same blackbox classifier, BBE performs similar or better than best alternate(s). Since overparamterized models start memorizing unlabeled samples negatives, the quality of MPE degrades substantially as PvU training proceeds for all methods but (TED)^n as in Fig. 3.

For classification, we compare our method with uPU, nnPU, Dedpul and PvU training (Table 2). Across many tasks, we observe substantial improvements over existing methods. To begin, we note that nnPU and (TED)^n doesn't need early stopping. Moreover, (TED)^n improves over existing baselines even with an oracle early stopping criterion (described in App. G) highlighting the significance of our procedure.

Conclusion In this paper, we proposed two practical algorithms, BBE (for MPE) and CVuO optimization (for PU learning). Our methods out perform others empirically on benchmarked datasets and BBE leverages black box classifiers to produce (nearly) consistent estimates with finite sample convergence guarantees whenever we possess a classifier with a (nearly) pure top bin. An important next direction is to extend our work to multiclass problems, bridging work on label shift and PU learning.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, 2016.

Alexandari, A., Kundaje, A., and Shrikumar, A. Adapting to label shift with bias-corrected calibration. In *arXiv preprint arXiv:1901.06852*, 2019.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations (ICLR)*, 2019.

Bekker, J. and Davis, J. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Bekker, J. and Davis, J. Learning from positive and unlabeled data: a survey. *Mach. Learn.*, 2020.

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

De Comité, F., Denis, F., Gilleron, R., and Letouzey, F. Positive and unlabeled examples help learning. In *International Conference on Algorithmic Learning Theory*, pp. 219–230. Springer, 1999.

Denis, F. Pac learning from positive statistical queries. In *International Conference on Algorithmic Learning Theory*, pp. 112–126. Springer, 1998.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Du Plessis, M., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pp. 1386–1394, 2015.

Du Plessis, M. C. and Sugiyama, M. Class prior estimation from positive and unlabeled data. *IEICE TRANSACTIONS on Information and Systems*, 97(5):1358–1362, 2014a.

Du Plessis, M. C. and Sugiyama, M. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014b.

Du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27:703–711, 2014.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

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

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. C. A unified view of label shift estimation. *arXiv preprint arXiv:2003.07554*, 2020.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Ivanov, D. DEDPUL: Difference-of-estimated-densities-based positive-unlabeled learning. *arXiv preprint arXiv:1902.06965*, 2019.

Jain, S., White, M., Trosset, M. W., and Radivojac, P. Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*, 2016.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv Preprint arXiv:1412.6980*, 2014.

Kiryo, R., Niu, G., Du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems*, pp. 1675–1685, 2017.

Krizhevsky, A. and Hinton, G. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86, 1998.

Lee, W. S. and Liu, B. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*, volume 3, pp. 448–455, 2003.

Letouzey, F., Denis, F., and Gilleron, R. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pp. 71–85. Springer, 2000.

Li, X. and Liu, B. Learning to classify texts using positive and unlabeled data. In *IJCAI*. Citeseer, 2003.

Lipton, Z. C., Wang, Y.-X., and Smola, A. Detecting and Correcting for Label Shift with Black Box Predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Liu, B., Lee, W. S., Yu, P. S., and Li, X. Partially supervised classification of text documents. In *ICML*, 2002.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pp. 179–186. IEEE, 2003.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 2019.

Ramaswamy, H., Scott, C., and Tewari, A. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pp. 2052–2060, 2016.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 2002.

Sanderson, T. and Scott, C. Class proportion estimation with application to multiclass anomaly rejection. In *Artificial Intelligence and Statistics*, pp. 850–858, 2014.

Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pp. 838–846, 2015.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Storkey, A. When Training and Test Sets Are Different: Characterizing Learning Transfer. *Dataset Shift in Machine Learning*, 2009.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, D. and Lee, W. S. A simple probabilistic approach to learning from positive and unlabeled examples. In *Proceedings of the 5th annual UK workshop on computational intelligence (UKCI)*, pp. 83–87, 2005.

Appendix

A. Related Work

Research on MPE and PU learning date to (Denis, 1998; De Comité et al., 1999; Letouzey et al., 2000) (see review by (Bekker & Davis, 2020)). Elkan & Noto (2008) first proposed to leverage a PvU classifier to estimate the mixture proportion. Du Plessis & Sugiyama (2014b) propose a different method for estimating the mixture coefficient based on Pearson divergence minimization. While they do not require a PvU classifier, they suffer the same shortcoming. Both methods require that the positive and negative examples have disjoint support. Our requirements are considerably milder. (Blanchard et al., 2010) observe that without assumptions on the underlying positive and negative distributions, the mixture proportion is not identifiable. Furthermore, (Blanchard et al., 2010) provide an *irreducibility* condition that identifies α and propose an estimator that converges to the true α. While their estimator can converge arbitrarily slowly, Scott (2015) showed faster convergence ($\mathcal{O}(1/\sqrt{n})$) under stronger conditions. Unfortunately, despite its appealing theoretical properties Blanchard et al. (2010)'s estimator is computationally infeasible. Building on Blanchard et al. (2010), Sanderson & Scott (2014) and Scott (2015) proposed estimating the mixture proportion from a ROC curve constructed for the PvU classifier. However, when the PvU classifier is not perfect, these methods are not clearly understood. Ramaswamy et al. (2016) proposed the first computationally feasible algorithm for mixture proportion estimation with convergence guarantees to the true proportion. Their method KM, requires embedding distributions onto an RKHS. However, their estimator underperforms on high dimensional datasets and scales poorly with large datasets. Bekker & Davis (2018) proposed TIcE, hoping to identify a positive subdomain in the input space using decision tree induction. This method also underperforms in high-dimensional settings.

In the most similar works, (Jain et al., 2016) and Ivanov (2019) explore dimensionality reduction using a PvU classifier. Both methods estimate α through a procedure operating on the PvU classifier's output. However, neither methods has provided theoretical backing (Ivanov, 2019) concede that their method often fails and returns a zero estimate, requiring that they fall back to a different estimator. Moreover while both papers state that their method require the Bayes-optimal PvU classifier to identify α in the transformed space, we prove that even when hypothesis class is well specified for PvN learning, PvU training can fail to recover Bayes-optimal scoring function. By contrast, our estimator BBE is theoretically coherent under mild conditions and outperforms both of these methods empirically.

Given α, Elkan & Noto (2008) propose a transformation via Bayes rule to obtain the PvN classifier. They also propose a weighted objective, with weights given by the PvU classifier. Other propose unbiased risk estimators (Du Plessis et al., 2014; 2015) which require the mixture proportion α. Du Plessis et al. (2014) proposed an unbiased estimator with non-convex loss functions satisfying a specific symmetric condition, and subsequently Du Plessis et al. (2015) generalized it to convex loss functions (denoted uPU in our experiments). in our experiments. Noting the problem of overfitting in modern overparameterized models, Kiryo et al. (2017) propose a regularized extension that clips the loss on unlabeled data to zero. This is considered the current state-of-the-art in PU literature (denoted nnPU in our experiments). More recently, Ivanov (2019) proposed Dedpul, which finetunes the PvU classifiers using several heuristics, Bayes rule, and Expectation Maximization (EM). Since their method only applies a post-processing procedure, they rely on a good domain discriminator classifier in the first place and several hyperparameters for their heuristics. Several classical methods attempt to learn weights that identify reliable negative examples (Liu et al., 2002; Li & Liu, 2003; Lee & Liu, 2003; Liu et al., 2003; Zhang & Lee, 2005). While our loss may be similar in spirit, these earlier methods have not been successful with modern deep learning models.

B. Impossibility Absent Assumptions

Indeed, if $\mathrm{P}_u = \alpha\mathrm{P}_p + (1-\alpha)\mathrm{P}_n$, then any alternate decomposition of the form $\mathrm{P}_u = (\alpha-\gamma)\mathrm{P}_p + (1-\alpha+\gamma)\mathrm{P}'_n$, for $\gamma \in [0, \alpha)$ and $\mathrm{P}'_n = (1-\alpha+\gamma)^{-1}(\gamma\mathrm{P}_p + (1-\alpha)\mathrm{P}_n)$, is also valid. Blanchard et al. (2010) formulate an *irreducibility* condition under which α is identifiable. Intuitively, the condition restricts P_n to ensure that it can not be a (non-trivial) mixture of P_p and any other distribution. While this irreducibility condition makes α identifiable, in the worst-case, the parameter α can be difficult to estimate and any estimator must suffer an arbitrarily slow rate of convergence (Blanchard et al., 2010).

C. Integrating BBE and CVuO

Algorithm 3 Transform-Estimate-Discard (TED)^n

input : Positive data (X_p) and unlabeled samples (X_u). Hyperparameter W, δ.
1: Initialize a training model f_θ and an stochastic optimization algorithm $\mathcal{A}$.
2: Randomly split positive and unlabeled data into training X_p^1, X_u^1 and hold-out set (X_p^2, X_u^2).
3: $X_n^1 := X_u^1$.
{// Warm start with domain discrimination training}
4: **for** $i = 1$ to W **do**
5: Shuffle (X_p^1, X_n^1) into B mini-batches. With $({X_p^1}^i, {X_n^1}^i)$ we denote i-th mini-batch.
6: **for** $i = 1$ to B **do**
7: Set the gradient $\nabla_\theta \left[\widehat{L}^+(f_\theta; {X_p^1}^i) + \widehat{L}^-(f_\theta; {X_n^1}^i)\right]$ and update θ with algorithm $\mathcal{A}$.
8: **end for**
9: **end for**
10: **while** training error $\widehat{\mathcal{E}}^+(f_\theta; X_p^1) + \widehat{\mathcal{E}}^-(f_\theta; X_n^1)$ is not converged **do**
11: Estimate $\widehat{\alpha}$ using Algorithm 1 with (X_p^2, X_u^2) and f_θ as input.
12: Rank samples $x_u \in X_u^1$ according to their loss values $l(f_\theta(x_u), -1)$.
13: $X_n^1 := X_{u,1-\widehat{\alpha}}^1$ where $X_{u,1-\widehat{\alpha}}^1$ denote the lowest ranked $1-\widehat{\alpha}$ fraction of samples.
14: Train model f_θ for one epoch on (X_p^1, X_n^1) as in Lines 4-7.
15: **end while**
output : Trained classifier f_θ

Note that we need to warm start with PvU (positive versus negative) training, since in the initial stages mixture proportion estimate is often close to 1 rejecting all the unlabeled examples. We fix $W = 100$ in our experiments. However, in App. I.4, we show that our 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.

D. Proofs from Sec. 3

Lemma 1. *For every $\delta > 0$, with probability at least $1-\delta$, we have for all $c \in [0,1]$*

$$\left|\frac{\widehat{q}_u(c)}{\widehat{q}_p(c)} - \frac{q_u(c)}{q_p(c)}\right| \leqslant \frac{1}{\widehat{q}_p(c)}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \frac{q_u(c)}{q_p(c)}\sqrt{\frac{\log(4/\delta)}{2n_p}}\right).$$

Proof of Lemma 1. The proof primarily involves using DKW inequality (Dvoretzky et al., 1956) on $\widehat{q}_u(c)$ and $\widehat{q}_p(c)$ to show convergence to their respective means $q_u(c)$ and $q_p(c)$. First, we have

$$\begin{aligned}\left|\frac{\widehat{q}_u(c)}{\widehat{q}_p(c)} - \frac{q_u(c)}{q_p(c)}\right| &= \frac{1}{\widehat{q}_u(c)\cdot q_u(c)}\left|\widehat{q}_u(c)\cdot q_p(c) - q_p(c)\cdot q_u(c) + q_p(c)\cdot q_u(c) - \widehat{q}_p(c)\cdot q_u(c)\right| \\ &\leqslant \frac{1}{\widehat{q}_p(c)}\left|\widehat{q}_u(c) - q_u(c)\right| + \frac{q_u(c)}{\widehat{q}_u(c)\cdot q_u(c)}\left|\widehat{q}_p(c) - q_p(c)\right|. \end{aligned} \tag{1}$$

Using DKW inequality, we have with probability $1-\delta$: $|\widehat{q}_p(c) - q_p(c)| \leqslant \sqrt{\frac{\log(2/\delta)}{2n_p}}$ for all $c \in [0,1]$. Similarly, we have with probability $1-\delta$: $|\widehat{q}_u(c) - q_u(c)| \leqslant \sqrt{\frac{\log(2/\delta)}{2n_u}}$ for all $c \in [0,1]$. Plugging this in (1), we have

$$\left|\frac{\widehat{q}_u(c)}{\widehat{q}_p(c)} - \frac{q_u(c)}{q_p(c)}\right| \leqslant \frac{1}{\widehat{q}_p(c)}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \frac{q_u(c)}{q_p(c)}\sqrt{\frac{\log(4/\delta)}{2n_p}}\right).$$

□

Proof of Theorem 1. The main idea of the proof is to use the confidence bound derived in Lemma 1 at $\widehat{c}$ and use the fact that $\widehat{c}$ minimizes the upper confidence bound.

First, we establish lower bound on $\widehat{\alpha}$. From Lemma 1, we have the following inequality at $\widehat{c}$

$$\frac{q_u(\widehat{c})}{q_p(\widehat{c})} \leqslant \frac{\widehat{q}_u(\widehat{c})}{\widehat{q}_p(\widehat{c})} + \frac{1}{\widehat{q}_p(\widehat{c})}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \frac{q_u(\widehat{c})}{q_p(\widehat{c})}\sqrt{\frac{\log(4/\delta)}{2n_p}}\right). \tag{2}$$

Moreover since $q_p(c) \geqslant q_n(c)$ for all $c \in [0,1]$, we have

$$q_u(\widehat{c})/q_p(\widehat{c}) \leqslant 1\,. \tag{3}$$

Additionally, we have

$$\alpha^* = \min_{c\in[0,1]} q_u(c)/q_p(c) \leqslant q_u(\widehat{c})/q_p(\widehat{c})\,. \tag{4}$$

Plugging in (3) and (4) in (2), we have

$$\alpha^* \leqslant \frac{\widehat{q}_u(\widehat{c})}{\widehat{q}_p(\widehat{c})} + \frac{1}{\widehat{q}_p(\widehat{c})}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right). \tag{5}$$

Recall $\widehat{\alpha} = \frac{\widehat{q}_u(\widehat{c})}{\widehat{q}_p(\widehat{c})} + (1/\widehat{q}_p(c) - 1/\gamma)_+\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right)$ for some constant $\gamma \in [0,1]$. Plugging this back in (5), we get

$$\begin{aligned}\alpha^* - \left(\frac{1}{\widehat{q}_p(\widehat{c})} - \left(\frac{1}{\widehat{q}_p(\widehat{c})} - \frac{1}{\gamma}\right)_+\right)\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right)& \\ \leqslant \frac{\widehat{q}_u(\widehat{c})}{\widehat{q}_p(\widehat{c})} + \left(\frac{1}{\widehat{q}_p(\widehat{c})} - \frac{1}{\gamma}\right)_+\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right) = \widehat{\alpha}\,.&\end{aligned} \tag{6}$$

Therefore, we have

$$\alpha^* - \frac{1}{\gamma}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right) \leqslant \widehat{\alpha}\,. \tag{7}$$

Next, we derive upper bound on $\widehat{\alpha}$. Since, we minimize RHS of (5) in Algorithm 1, we simply have

$$\begin{aligned}\frac{\widehat{q}_u(\widehat{c})}{\widehat{q}_p(\widehat{c})} + \frac{1}{\widehat{q}_p(\widehat{c})}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right)& \\ \leqslant \min_{c\in[0,1]}\left[\frac{\widehat{q}_u(c)}{\widehat{q}_p(c)} + \frac{1}{\widehat{q}_p(c)}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right)\right]& \\ \leqslant \frac{\widehat{q}_u(c^*)}{\widehat{q}_p(c^*)} + \frac{1}{\widehat{q}_p(c^*)}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right),&\end{aligned} \tag{8}$$

for $c^* = \arg\min_{c\in[0,1]} q_u(c)/q_p(c)$. Using Lemma 1 at c^*, we get

$$\begin{aligned}\frac{\widehat{q}_u(c^*)}{\widehat{q}_p(c^*)} &\leqslant \frac{q_u(c^*)}{q_p(c^*)} + \frac{1}{\widehat{q}_p(c^*)}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \frac{q_u(c^*)}{q_p(c^*)}\sqrt{\frac{\log(4/\delta)}{2n_p}}\right) \\ &= \alpha^* + \frac{1}{\widehat{q}_p(c^*)}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \alpha^*\sqrt{\frac{\log(4/\delta)}{2n_p}}\right).\end{aligned} \tag{9}$$

Combining (8) and (9), we have

$$\frac{\widehat{q}_u(\widehat{c})}{\widehat{q}_p(\widehat{c})} + \frac{1}{\widehat{q}_p(\widehat{c})}\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right) \leqslant \alpha^* + \frac{1}{\widehat{q}_p(c^*)}\left(2\sqrt{\frac{\log(4/\delta)}{2n_u}} + (1+\alpha^*)\sqrt{\frac{\log(4/\delta)}{2n_p}}\right). \tag{10}$$

Plugging the estimator $\widehat{\alpha}$ in LHS of (10), we get

$$\begin{aligned}\widehat{\alpha} &\leqslant \alpha^* + \frac{1}{\widehat{q}_p(c^*)}\left(2\sqrt{\frac{\log(4/\delta)}{2n_u}} + (1+\alpha^*)\sqrt{\frac{\log(4/\delta)}{2n_p}}\right) \\ &\quad - \left(\frac{1}{\widehat{q}_p(\widehat{c})} - \left(\frac{1}{\widehat{q}_p(\widehat{c})} - \frac{1}{\gamma}\right)_+\right)\left(\sqrt{\frac{\log(4/\delta)}{2n_u}} + \sqrt{\frac{\log(4/\delta)}{2n_p}}\right) \\ &\leqslant \alpha^* + \frac{1}{\widehat{q}_p(c^*)}\left(2\sqrt{\frac{\log(4/\delta)}{2n_u}} + (1+\alpha^*)\sqrt{\frac{\log(4/\delta)}{2n_p}}\right). \end{aligned} \tag{11}$$

Finally, using DKW inequality on $\widehat{q}_p(c^*)$, we have $\widehat{q}_p(c^*) \geqslant q_p(c^*) - \sqrt{\frac{\log(4/\delta)}{2n_p}}$. Assuming $n_p \geqslant \sqrt{\frac{2\log(4/\delta)}{q_p(c^*)}}$, we have $\widehat{q}_p(c^*) \geqslant q_p(c^*)/2$. Hence, we have the following upper bound:

$$\widehat{\alpha} \leqslant \alpha^* + \frac{2}{q_p(c^*)}\left(2\sqrt{\frac{\log(4/\delta)}{2n_u}} + (1+\alpha^*)\sqrt{\frac{\log(4/\delta)}{2n_p}}\right). \tag{12}$$

□

Corollary 1. *Let $q_p(c) \geqslant q_n(c)$ for all $c \in [0,1]$. Define $c^* = \arg\min_{c\in[0,1]} q_n(c)/q_p(c)$ and assume $n_p \geqslant \sqrt{2\log(4/\delta)/q_p(c^*)}$. For every $\delta > 0$, $\widehat{\alpha}$ (in Algo. 1) satisfies with probability $1-\delta$:*

$$\alpha - c_1 \cdot \left(\sqrt{\log(4/\delta)/n_u} + \sqrt{\log(4/\delta)/n_p}\right) \leqslant \widehat{\alpha}\,, \text{and}$$

$$\widehat{\alpha} \leqslant \alpha + (1-\alpha)\frac{q_n(c^*)}{q_p(c^*)} + \frac{c_2}{q_p(c^*)} \cdot \left(2\sqrt{\log(4/\delta)/n_u} + \left(1 + \frac{q_u(c^*)}{q_p(c^*)}\right)\sqrt{\log(4/\delta)/n_p}\right),$$

for some constant $c_1, c_2 \geqslant 0$.

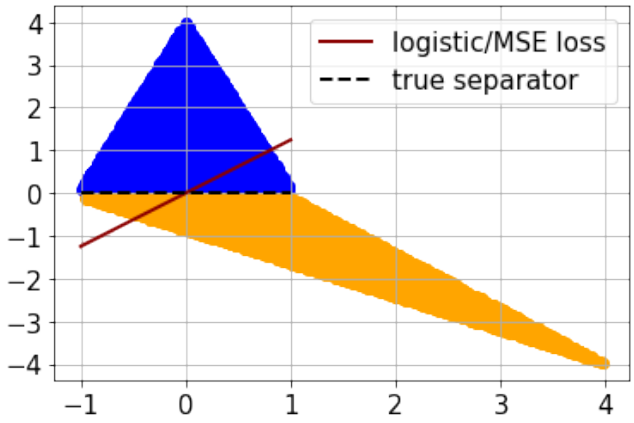
Proof of Corollary 1. Note that since $\alpha \leqslant \alpha^*$, the lower bound remains the same as in Theorem 1. For upper bound, plugging in $q_u(c) = \alpha q_p(c) + (1-\alpha)q_n(c)$, we have $\alpha^* = \alpha + (1-\alpha)q_n(c^*)/q_p(c^*)$ and hence, the required upper bound. □

D.1. Note on γ in Algorithm 1

According to Lemma 1, when $\widehat{q}_p(\widehat{c})$ is arbitrary low, the upper bound in Lemma 1 can be large and hence the concentration to the true mixture proportion can be loose. Thus, we add the confidence bound to our ratio estimate ($\widehat{q}_u(\widehat{c})/\widehat{q}_p(\widehat{c})$) to yield an estimate with tighter guarantees in Theorem 1. However, since we are minimizing the upper confidence bound, in practice, we never observe $\widehat{q}_p(\widehat{c})$ taking arbitrarily small values. In our experiments, we fixed γ at 0.1 and we didn't observe $\widehat{q}_p(\widehat{c})$ taking smaller values than γ.

E. Toy setup

Jain et al. (2016) and Ivanov (2019) discuss Bayes optimality of the PvU classifier (or its one-to-one mapping) as a sufficient condition to preserve α in transformed space. However, in a simple toy setup (in App. E), we show that even when the



(a)

Figure 4. Blue points show samples from the positive distribution and orange points show samples from the negative distribution. Unalabeled data is obtained by mixing positive and negative distribution with equal proportion. BCE (or Brier) loss minimization on P vs U data leads to a classifiers that is not consistent with the ranking of the Bayes optimal score function.

hypothesis class is well specified for PvN learning, it will not in general contain the Bayes optimal scoring function for PvU data and thus PvU training will not recover the Bayes-optimal scoring function, even in population.

Consider a scenario with $\mathcal{X} = \mathbb{R}^2$. Assume points from the positive class are sampled uniformly from the interior of the triangle defined by coordinates $\{(-1, 0.1), (0, 4), (1, 0.1)\}$ and negative points are sampled uniformly from the interior of triangle defined by coordinates $\{(-1, -0.1), (4, -4), (1, -0.1)\}$. Ref. to Fig. 4 for a pictorial representation. Let mixture proportion be 0.5 for the unlabeled data. Given access to distribution of positive data and unlabeled data, we seek to train a linear classifier to minimize logistic or Brier loss for PvU training.

Since we need a monotonic transformation of the Bayes optimal scoring function, we want to recover a predictor parallel to x-axis, the Bayes optimal classifier for PvN training. However, minimizing the logistic loss (or Brier loss) using numerical methods, we obtain a predictor that is inclined at a non-zero acute angle to the x-axis. Thus, the PvU classifier obtained fails to satisfy the sufficient condition from Jain et al. (2016) and Ivanov (2019). On the other hand, note that the linear classifier obtained by PvU training satisfies the pure positive bin property.

Now we show that under the subdomain assumption (Scott, 2015; Ramaswamy et al., 2016), any monotonic transformation of Bayes optimal scoring function induces positive pure bin property. First, we define the subdomain assumption.

Assumption 1 (Subdomain assumption)**.** *A family of subsets $\mathcal{S} \subseteq 2^{\mathcal{X}}$, and distributions p_p, p_n are said to satisfy the anchor set condition with margin $\gamma > 0$, if there exists a compact set $A \in \mathcal{S}$ such that $A \subseteq supp(p_p)/supp(p_n)$ and $p_p(A) \geqslant \gamma$.*

Note that any monotonic mapping of the Bayes optimal scoring function can be represented by $\tau' = g \circ \tau$, where g is a monotonic function and

$$\tau(x) = \begin{cases} p_p(x)/p_u(x) & \text{if } p_p(x) > 0 \\ 0 & \text{o.w}\,. \end{cases} \tag{13}$$

For any point $x \in A$ and $x' \in \mathcal{X}/A$, we have $\tau(x) > \tau(x')$ which implies $\tau'(x) > \tau'(x')$. Thus, any monotonic mapping of Bayes optimal scoring function yields the positive pure bin property with $\epsilon_p \geqslant \gamma$.

F. Analysis of CVuO in separable case

We now analyse our loss function in the scenario when the support of positives and negatives is separable. We assume that the true alpha α is known and we have access to populations of positive and unlabeled data. We also assume that their exists a separator $f^* : \mathcal{X} \mapsto \{0, 1\}$ that can perfectly separate the positive and negative distribution, i.e., $\int dx p_p(x) \mathbb{I}[f^*(x) \neq 1] + \int dx p_n(x) \mathbb{I}[f^*(x) \neq 0] = 0$. Our learning objective can be written as jointly optimizing a

classifier f and a weighting function w on the unlabeled distribution:

$$\min_{f\in\mathcal{F},w} \int dx p_p(x) l(f(x),1) + \frac{1}{1-\alpha}\int dx p_u(x) w(x) l(f(x),0)\,,$$
$$\text{s.t. } w:\mathcal{X}\mapsto[0,1]\,,\int dx p_u(x)w(x) = 1-\alpha\,. \tag{14}$$

The following proposition shows that minimizing the objective (14) on separable positive and negative distributions gives a perfect classifier.

Proposition 1. *For $\alpha\in(0,1)$, if there exists a classifier $f^*\in\mathcal{F}$ that can perfectly separate the positive and negative distributions, optimizing objective (14) with 0-1 loss leads to a classifier f that achieves 0 classification error on the unlabeled distribution.*

Proof. First we observe that having $w(x) = 1 - f^*(x)$ leads to the objective value being minimized to 0 as well as a perfect classifier f. This is because

$$\frac{1}{1-\alpha}\int dx p_u(x)(1-f^*(x))l(f(x),0) = \int dx p_n(x) l(f(x),0)$$

thus the objective becomes classifying positive v.s. negative, which leads to a perfect classifier if $\mathcal{F}$ contains one. Now we show that for any f such that the classification error is non-zero then the objective (14) must be greater than zero no matter what w is. Suppose f satisfies

$$\int dx p_p(x) l(f(x),1) + \int dx p_n(x) l(f(x),0) > 0\,.$$

We know that either $\int dx p_p(x) l(f(x),1) > 0$ or $\int dx p_n(x) l(f(x),0) > 0$ will hold. If $\int dx p_p(x) l(f(x),1) > 0$ we know that (14) must be positive. If $\int dx p_p(x) l(f(x),1) = 0$ and $\int dx p_n(x) l(f(x),0) > 0$ we have $l(f(x),0) = 1$ almost everywhere in $p_p(x)$ thus

$$\begin{aligned}
&\frac{1}{1-\alpha}\int dx p_u(x) w(x) l(f(x),0)\\
&= \frac{\alpha}{1-\alpha}\int dx p_p(x) w(x) l(f(x),0) + \int dx p_n(x) w(x) l(f(x),0)\\
&= \frac{\alpha}{1-\alpha}\int dx p_p(x) w(x) + \int dx p_n(x) w(x) l(f(x),0)\,.
\end{aligned}$$

If $\int dx p_p(x) w(x) > 0$ we know that (14) must be positive. If $\int dx p_p(x) w(x) = 0$, since we know that

$$\int dx p_u(x) w(x) = \alpha\int dx p_p(x) w(x) + (1-\alpha)\int dx p_n(x) w(x) = 1-\alpha$$

we have $\int dx p_n(x) w(x) = 1$ which means $w(x) = 1$ almost everywhere in $p_n(x)$. This leads to the fact that $\int dx p_n(x) l(f(x),0) > 0$ indicates $\int dx p_n(x) w(x) l(f(x),0) > 0$, which concludes the proof.

□

G. Summary of Experiments

Datasets and Evaluation We simulate PU tasks on CIFAR-10 (Krizhevsky & Hinton, 2009), MNIST (LeCun et al., 1998), and IMDb sentiment analysis (Maas et al., 2011) datasets. We consider binarized versions of CIFAR-10 and MNIST. On CIFAR-10 dataset, we consider two classification problems: (i) binarized CIFAR, i.e., first 5 classes vs rest; (ii) Dog vs Cat in CIFAR. Similarly on MNIST, we consider: (i) binarized MNIST, i.e., digits 0-4 vs 5-9; (ii) MNIST17, i.e., digit 1 vs 7. IMDb dataset is binary. For MPE, we use a held out PU validation set. To evaluate PU classifiers, we calculate accuracy on held out positive versus negative dataset. For baselines that suffer from issues due to overfitting on unlabeled data, we report results with an *oracle early stopping* criterion. We report the accuracy averaged over 10 iterations of the best performing model. With nnPU and $(\text{TED})^n$, we report average accuracy over 10 iterations of the final model.

Dataset	Model	(TED)^n	BBE*	Dedpul*	AlphaMax*	EN	KM2	TiCE
Binarized CIFAR	ResNet	**0.018**	0.072	0.075	0.125	0.175		
	All Conv	0.041	**0.038**	0.046	0.09	0.23	0.181	0.251
	FCN	0.184	0.175	**0.151**	0.3	0.355		
CIFAR Dog vs Cat	ResNet	**0.074**	0.120	0.113	0.17	0.205	0.11	0.203
	All Conv	**0.073**	0.093	0.098	0.19	0.274		
Binarized MNIST	FCN	**0.021**	0.028	0.027	0.09	0.067	0.102	0.247
MNIST17	FCN	**0.003**	0.008	0.006	0.075	0.065	0.03	0.117
IMDb	BERT	**0.008**	0.011	0.016	0.07	0.12	-	-

Table 1. Absolute estimation error when α is 0.5. "*" denote oracle early stopping as defined in Sec. 5. Results reported by aggregating absolute error over 10 epochs and 3 seeds.

Dataset	Model	(TED)^n (unknown α)	CVuO (known α)	PvU* (known α)	Dedpul* (unknown α)	nnPU (known α)	uPU* (known α)
Binarized CIFAR	ResNet	**82.7**	82.6	78.3	78.4	76.8	75.8
	All Conv	76.8	**77.1**	74.1	76.9	72.1	71.3
	FCN	63.2	**65.9**	61.4	62.5	63.9	64.8
CIFAR Dog vs Cat	ResNet	**76.1**	74.0	71.6	70.9	72.6	69.5
	All Conv	**72.2**	71.0	70.1	70.5	68.4	65.2
Binarized MNIST	FCN	95.9	**96.4**	94.5	95.2	95.9	95.0
MNIST17	FCN	**98.6**	**98.6**	93.7	98.1	98.2	98.4
IMDb	BERT	**87.6**	87.4	86.1	87.3	86.2	85.9

Table 2. Accuracy for PvN classification with PU learning. "*" denote oracle early stopping as defined in Sec. 5. Results reported by aggregating aggregating over 10 epochs and 3 seeds.

Architectures For CIFAR datasets, we consider (fully connected) multilayer perceptrons (MLPs) with ReLU activations, all convolution nets (Springenberg et al., 2014), and ResNet18 (He et al., 2016). For MNIST, we consider multilayer perceptrons (MLPs) with ReLU activations For the IMDb dataset, we fine-tune an off-the-shelf uncased BERT model (Devlin et al., 2018; Wolf et al., 2020). We did not tune hyperparameters or the optimization algorithm—instead we use the same benchmarked hyperparameters and optimization algorithm for each dataset. For our method, we use cross-entropy loss. For uPU and nnPU, we use Adam (Kingma & Ba, 2014) with sigmoid loss. We provide additional details about the datasets and architectures in App. H.

Mixture Proportion Estimation We compare our method with KM2, TiCE, Dedpul, AlphaMax and EN. For KM2 and TiCE, datasets are reduced to 100 dimensions with PCA. We use existing implementation for other methods[1]. For our method, Dedpul and Alphamax, we use the same PvU classifier as input. On all datasets, classifier based estimators outperform KM and TiCE. With the same blackbox classifier, BBE performs similar or better than best alternate(s). Since overparamterized models start memorizing unlabeled samples negatives, the quality of MPE degrades substantially as PvU training proceeds for all methods but (TED)^n as in Fig. 3 (epoch-wise results for on other tasks in App. I.3).

Classification with known MPE Now, we discuss results for classification with known α. We compare our method with uPU, nnPU[2], Dedpul and PvU training. Although, we solve both MPE and classification, some comparison methods do not. Ergo, we compare our classification algorithm with known MPE (Algorithm 2).

[1]Dedpul: https://github.com/dimonenka/DEDPUL, KM: https://web.eecs.umich.edu/~cscott/code.html#kmpe, TiCE: https://dtai.cs.kuleuven.be/software/tice, and AlphaMax: https://github.com/Dzeiberg/AlphaMax

[2]uPU and nnPU: https://github.com/kiryor/nnPUlearning

To begin, first we note that nnPU and PvU training with CVuO doesn't need early stopping. For all other methods, we report the best performance dictated by the aforementioned oracle stopping criterion. On all datasets, PvU training with CVuO leads to improved classification performance when compared with alternate approaches (Table 2). Moreover, as training proceeds (Fig. 3), the performance of Dedpul, PvU training and uPU substantially degrade. We repeated experiments with the early stopping criterion mentioned in Dedpul (App. I.2), however, their early stopping criterion is too pessimistic resulting in poor results due to under-fitting.

Classification with unknown MPE Finally, we evaluate (TED)^n, our alternating procedure for MPE and PU learning. Across many tasks, we observe substantial improvements over existing methods. Note that these improvements often are over an oracle early stopping baselines highlighting significance of our procedure.

In App. I.4, we show that our procedure is not sensitive to warm start epochs W, and in many tasks with $W = 0$, we observe minor-to-no differences in the performance of (TED)^n. While for the experiments in this section, we used fixed $W = 100$, in the Appendix we show behavior with varying W. We also include ablations with different mixture proportions α.

H. Experimental Details

Below we present dataset details. We present experiments with MNIST Overlap in App. I.6.

Dataset	Simulated PU Dataset	P vs N	#Positives		#Unlabeled	
			Train	Val	Train	Val
CIFAR10	Binarized CIFAR	[0-4] vs [5-9]	12500	12500	2500	2500
	CIFAR Dog vs Cat	3 vs 5	2500	2500	500	500
MNIST	Binarized MNIST	[0-4] vs [5-9]	15000	15000	2500	2500
	MNIST 17	1 vs 7	3000	3000	500	500
	MNIST Overlap	[0-7] vs [3-9]	150000	15000	2500	2500
IMDb	IMDb	pos vs neg	6250	6250	5000	5000

For CIFAR dataset, we also use the standard data augementation of random crop and horizontal flip. PyTorch code is as follows:

```
(transforms.RandomCrop(32, padding=4),
transforms.RandomHorizontalFlip())
```

H.1. Architecture and Implementation Details

All experiments were run on NVIDIA GeForce RTX 2080 Ti GPUs. We used PyTorch (Paszke et al., 2019) and Keras with Tensorflow (Abadi et al., 2016) backend for experiments.

For CIFAR10, we experiment with convolutional nets and FCN. For MNIST, we train FCN. In particular, we use ResNet18 (He et al., 2016) and all convolution net (Springenberg et al., 2014) . Implementation adapted from: `https://github.com/kuangliu/pytorch-cifar.git`. We consider a 4-layered MLP. The PyTorch code for 4-layer MLP is as follows:

```
 nn.Sequential(nn.Flatten(),
nn.Linear(input_dim, 5000, bias=True),
nn.ReLU(),
nn.Linear(5000, 5000, bias=True),
nn.ReLU(),
nn.Linear(5000, 50, bias=True),
nn.ReLU(),
```

```
nn.Linear(50, 2, bias=True)
)
```

For all architectures above, we use Xaviers initialization (Glorot & Bengio, 2010). For all methods except nnPU and uPU, we do cross entropy loss minimization with SGD optimizer with momentum 0.9 and learning rate 0.1. For nnPU and uPU, we minimize sigmoid loss with ADAM optimizer with learning rate 0.0001 as advised in its original paper. For all methods, we fix the weight decay param at 0.0005.

For IMDb dataset, we fine-tune an off-the-shelf uncased BERT model (Devlin et al., 2018). Code adapted from Hugging Face Transformers (Wolf et al., 2020): `https://huggingface.co/transformers/v3.1.0/custom_datasets.html`. For all methods except nnPU and uPU, we do cross entropy loss minimization with Adam optimizer with learning rate 0.00005 (default params). With the same hyperparameters and Sigmoid loss, we could not train BERT with nnPU and uPU due to vanishing gradients. Instead we use learning rate 0.00001.

I. Additional Experiments

I.1. nnPU vs PN classification

In this section, we compare the performance of nnPU and PvN training on the same positive and negative (from the unlabeled) data at $\alpha = 0.5$. We highlight the huge classification performance gap between nnPU and PvN training and show that training with CVuO objective partially recovers the performance gap. Note, to train PvN classifier, we use the same hyperparameters as that with PvU training.

Dataset	Model	nnPU (known α)	PvN	CVuO (known α)	(TED)^n (unknown α)
Binarized CIFAR	ResNet	76.8	86.9	82.6	82.7
	All Conv	72.1	76.7	77.1	76.8
	FCN	63.9	65.1	65.9	63.2
CIFAR Dog vs Cat	ResNet	72.6	80.4	74.0	76.1
	All Conv	68.4	77.9	71.0	72.2
Binarized MNIST	FCN	95.9	96.7	96.4	95.9
MNIST17	FCN	98.2	99.0	98.6	98.6
IMDb	BERT	86.2	89.1	87.4	88.1

Table 3. Accuracy for PvN classification with nnPU, PvN, CVuO objective and (TED)^n training. Results reported by aggregating aggregating over 10 epochs.

I.2. Under-Fitting due to pessimistic early stopping

Ivanov (2019) explored the following heuristics for ad-hoc early stopping criteria: training proceeds until the loss on unseen PU data ceases to decrease. In particular, the authors suggested early stopping criterion based on the loss on unseen PU data doesn't decrease in epochs separated by a pre-defined window of length l. The early stopping is done when this happens consecutively for l epochs. However, this approach leads to severe under-fitting. When we fix $l = 5$, we observe a significant performance drop in CIFAR classification and MPE.

With PvU training, the performance of ResNet model on Binarized CIFAR (in Table 2) drops from 78.3 (orcale stopping) to 60.4 (with early stopping). Similar on CIFAR CAT vs Dog, the performance of the same architecture drops from 71.6 (orcale stopping) to 58.4 (with early stopping). Note that the decrease in accuracy is less or not significant for MNIST. With PvU training, the performance of FCN model on Binarized MNIST (in Table 2) drops from 94.5 (orcale stopping) to 94.1 (with early stopping). This is because we obtain good performance on MNIST early in training.

I.3. Results parallel to Fig. 3

Epoch wise results for best performing model for CIFAR Dog vs Cat (Fig. 5), Binarized MNIST (Fig. 6), MNIST 17 (Fig. 7) and IMDb (Fig. 8)

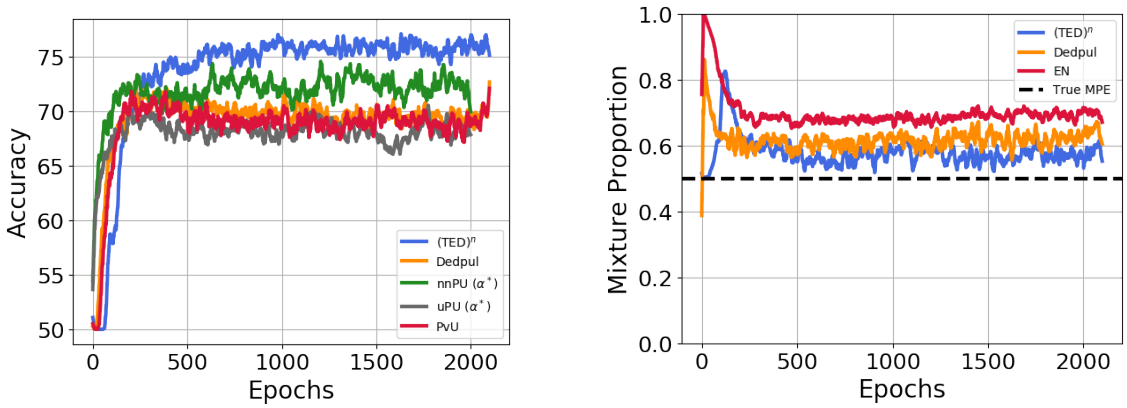


Figure 5. Epoch wise results with ResNet-18 trained on CIFAR Dog vs Cat.

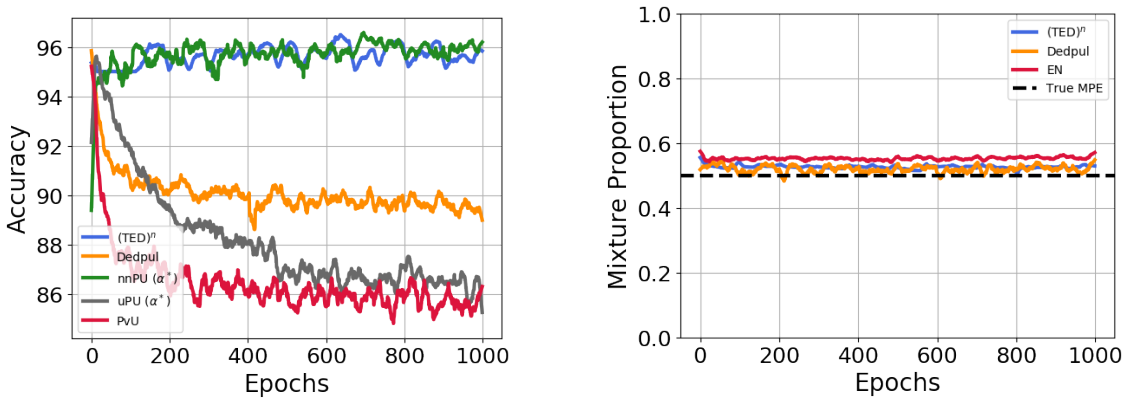


Figure 6. Epoch wise results with FCN trained on Binarized MNIST.

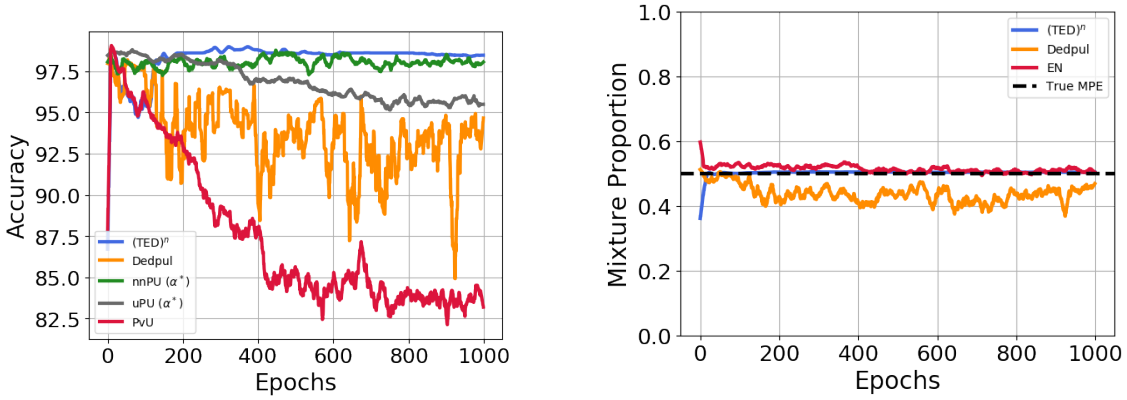


Figure 7. Epoch wise results with FCN trained on MNIST 17.

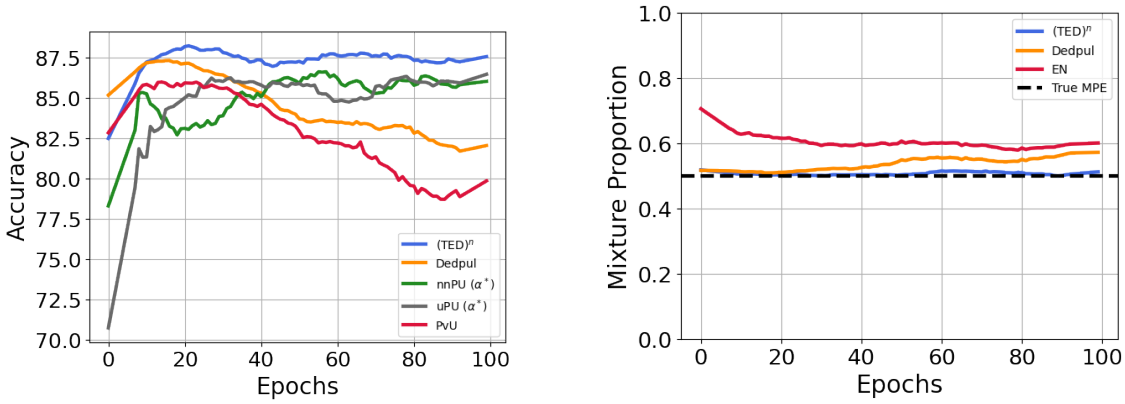


Figure 8. Epoch wise results with BERT trained on IMDb.

I.4. Ablations to (TED)^n

Varying the number of warm start epochs We now vary the number of warm start epochs with (TED)^n. We observe that increasing the number of warm start epochs doesn't hurt (TED)^n even when the classifier at the end of the warm start training memorized PU training data due PvU training. While in many cases (TED)^n training without warm start is able to recover the same performance, it fails to learn anything for CIFAR Dog vs Cat with all convolutional neural network. This highlights the need for warm start training with (TED)^n.

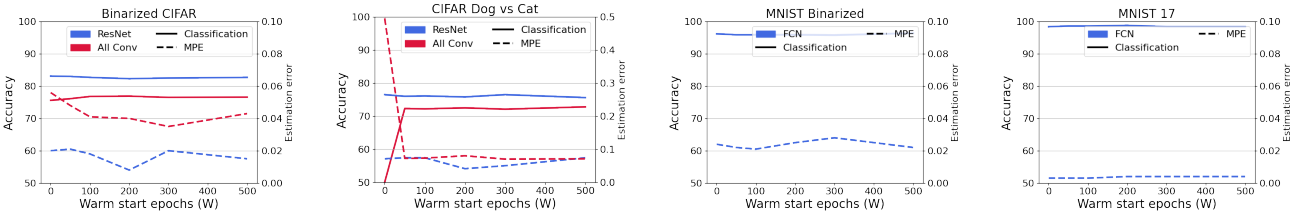


Figure 9. Classification and MPE results with varying warm start epochs W with (TED)^n

Varying the true mixture proportion α Next, we vary α, the true mixture proportion and present results for MPE and classification in Fig. 10. Overall, across all α, our method (TED)^n is able to achieve superior performance as compared to alternate algorithms. We omit high α for CIFAR and IMDb datasets as all the methods result in trivial accuracy and mixture proportion estimate.

I.5. Experiments on UCI dataset

In this section, we will present results on 5 UCI datasets.

Dataset	#Positives		#Unlabeled	
	Train	Val	Train	Val
concrete	162	162	81	81
mushroom	1304	1304	652	652
landsat	946	946	472	472
pageblock	185	185	92	92
spambase	604	604	302	302

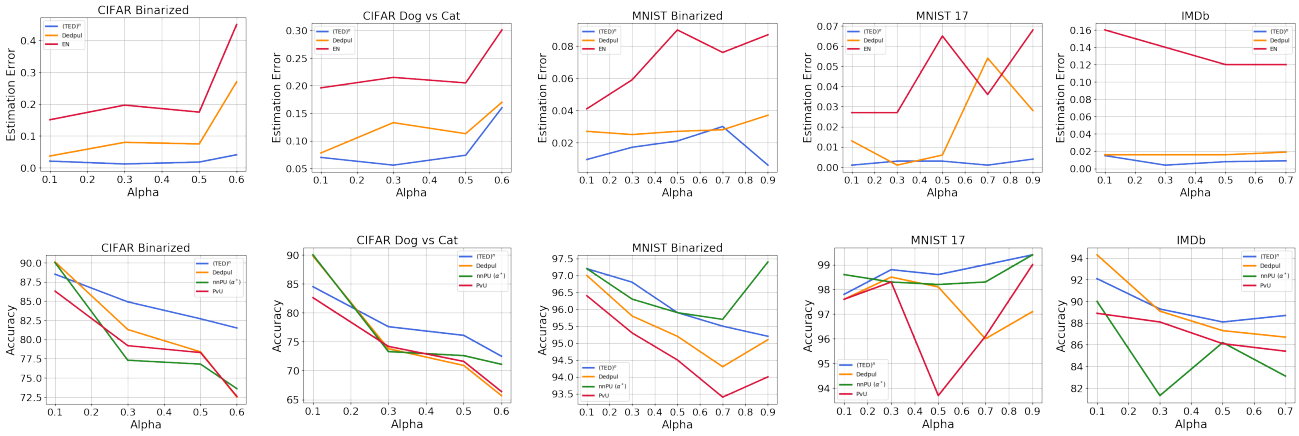


Figure 10. MPE and Classification results with varying mixture proportion. For each method we show results with the best performing architecture.

We train a FCN with 2 hidden layers each with 512 units. The PyTorch code for 4-layer MLP is as follows:

```
 nn.Sequential(nn.Flatten(),
nn.Linear(input_dim, 512, bias=True),
nn.ReLU(),
nn.Linear(512, 512, bias=True),
nn.ReLU(),
nn.Linear(512, 2, bias=True),
)
```

Similar to vision datasets and architectures, we do cross entropy loss minimization with SGD optimizer with momentum 0.9 and learning rate 0.1. For nnPU and uPU, we minimize sigmoid loss with ADAM optimizer with learning rate 0.0001 as advised in its original paper. For all methods, we fix the weight decay param at 0.0005.

Dataset	$(\text{TED})^n$	BBE*	Dedpul*	EN*	KM2	TiCE
concrete	**0.071**	0.152	0.176	0.239	0.099	0.268
mushroom	**0.001**	0.015	0.014	0.013	0.038	0.069
landsat	0.022	0.021	**0.012**	0.080	0.037	0.027
pageblock	**0.007**	0.066	0.041	0.135	0.008	0.298
spambase	**0.006**	0.047	0.077	0.127	0.062	0.276

Table 4. Absolute estimation error when α is 0.5. "*" denote oracle early stopping as defined in Sec. 5. Results reported by aggregating absolute error over 10 epochs.

Dataset	$(\text{TED})^n$ (unknown α)	CVuO (known α)	PvU* (known α)	Dedpul* (unknown α)	nnPU (known α)	uPU* (known α)
concrete	**86.3**	80.1	83.1	83.7	83.2	84.4
mushroom	96.4	96.3	**98.7**	**98.7**	97.5	93.9
landsat	**93.8**	93.1	93.4	92.4	92.9	92.3
pageblock	**95.7**	**95.7**	95.1	94.5	93.9	93.9
spambase	**89.4**	88.1	89.2	86.8	88.5	87.7

Table 5. Accuracy for PvN classification with PU learning. "*" denote oracle early stopping as defined in Sec. 5. Results reported by aggregating aggregating over 10 epochs.

On 4 out of 5 UCI datasets, our proposed methods are better than the best performing alternatives (Table 4 and Table 5).

I.6. Experiments on MNIST Overlap

Similar to binarized MNIST, we create a new dataset called MNIST Overlap, where the positive class contains digits from 0 to 7 and the negative class contains digits from 3 to 9. This creates a dataset with overlap between positive and negative support. Note that while the supports overlap, we sample images from the overlap classes with replacement, and hence, in absence of duplicates in the dataset, exact same images don't appear both in positive and negative subsets.

We train FCN with the same hyperparameters as before. Our findings in Table 6 and Table 7 highlight superior performance of the proposed approaches in the cases of support overlap.

Dataset	$(\text{TED})^n$	BBE*	Dedpul*	EN*	KM2	TiCE
MNIST Overlap	**0.035**	0.100	0.104	0.196	0.099	0.074

Table 6. Absolute estimation error when α is 0.5. "*" denote oracle early stopping as defined in Sec. 5. Results reported by aggregating absolute error over 10 epochs.

Dataset	$(\text{TED})^n$ (unknown α)	CVuO (known α)	PvU* (known α)	Dedpul* (unknown α)	nnPU (known α)	uPU* (known α)
MNIST Overlap	**79.0**	78.4	77.4	77.5	78.6	78.8

Table 7. Accuracy for PvN classification with PU learning. "*" denote oracle early stopping as defined in Sec. 5. Results reported by aggregating aggregating over 10 epochs.