Estimating Expected Calibration Error for Positive-Unlabeled Learning

Anonymous authors
Paper under double-blind review

Abstract

The reliability of probabilistic classifiers hinges on their calibration—the property that their confidence accurately reflect the true class probabilities. The expected calibration error (ECE) is a standard metric for quantifying the calibration of classifiers. However, its estimation presumes access to ground-truth labels. In positive-unlabeled (PU) learning, only positive and unlabeled data are available, which makes the standard ECE estimator inapplicable. Although PU learning has been extensively studied for risk estimation and classifier training, calibration in this setting has received little attention. In this paper, we present PU-ECE, the first ECE estimator for PU data. We provide non-asymptotic bias bounds and prove convergence rates that match those of the fully supervised ECE with an optimal bin size. Furthermore, we develop an information-theoretic generalization error analysis of PU-ECE by formalizing the conditional mutual information (CMI) for a PU setting. Experiments on synthetic and real-world benchmark datasets validate our theoretical analysis and demonstrate that our PU-based ECE estimator achieves performance comparable to that of the fully-labeled ECE estimator.

1 Introduction

Deep learning models achieve state-of-the-art accuracy in various classification tasks (Devlin et al., 2019; Dosovitskiy et al., 2021; Radford et al., 2021). However, their deployment in safety-critical applications requires more than just high prediction accuracy. For instance, in medical diagnosis (Jiang et al., 2012), autonomous systems (Muthali et al., 2023), and financial decision-making (Bahnsen et al., 2014), the confidence of the predictions provided by machine learning models must accurately reflect the true probability of the predicted events. This property, known as *calibration* (Dawid, 1982; Guo et al., 2017), is essential for reliable decision-making. The calibration error, or *true calibration error* (TCE) (Naeini et al., 2015; Guo et al., 2017; Roelofs et al., 2022), is a widely used metric for evaluating the calibration of classifiers. The TCE is typically estimated by partitioning the confidence score space into bins, followed by the calculation of the average difference between the confidence score and the empirical probability of the true label given the confidence score in each bin (Zadrozny & Elkan, 2001; Naeini et al., 2015). The resulting estimator is called the *expected calibration error* (ECE)[1] (Roelofs et al., 2022). Theoretical analysis of the ECE has also been conducted, establishing bias bounds and the consistency of the ECE to the TCE as the number of samples increases, with an appropriate binning strategy Gupta & Ramdas (2021); Futami & Fujisawa (2024). The results show an $\mathcal{O}(n^{-1/3})$ convergence rate of the ECE, where n is the number of test samples. The naive computation of the ECE requires a fully labeled dataset, where each sample has both confidence scores and true labels.

However, in practice, fully labeled data may not be available or may be costly to obtain (Jaskie & Spanias, 2019; Li et al., 2022; Bepler et al., 2019). For instance, when identifying protein particles in microscopy images, annotating negative data is costly due to diverse backgrounds (Bepler et al., 2019). In such cases, positive-unlabeled (PU) learning, which aims to train a classifier from only positive and unlabeled data, is a promising approach (Liu et al., 2002; Elkan & Noto, 2008; Natarajan et al., 2013; du Plessis et al., 2014).

---

[1]As some researchers have pointed out, both TCE and ECE are sometimes called "ECE". For clarity, we follow the convention of Futami & Fujisawa (2024).

Despite extensive research on PU learning methods, the majority of these methods have focused on maximizing classification accuracy (Elkan & Noto, 2008; du Plessis et al., 2015; Kiryo et al., 2017; Chen et al., 2020; Xinrui et al., 2023). On the other hand, the critical issue of calibration in the PU setting has been largely overlooked. As a key motivation for this paper, we first reveal a surprising phenomenon: in PU learning, the standard binary cross-entropy loss, which is strictly proper in supervised learning (Good, 1952; Savage, 1971), can lead to significantly worse calibration than a noise-robust loss that is not strictly proper (Charoenphakdee et al., 2019) but often used in PU learning (Kiryo et al., 2017; Chen et al., 2020; Zhao et al., 2022) (details in Section 5.2). However, evaluating calibration in the PU setting is challenging—the absence of negative labels makes it impossible to apply the standard ECE estimator, which prevents practitioners from evaluating the calibration of PU classifiers.

In this paper, we address the problem of estimating the calibration error in PU learning settings. We introduce *PU-ECE*, a novel method for estimating the ECE using only PU data, without requiring negative labels. The core idea is to reframe the ECE computation by leveraging the statistical properties of PU datasets. By applying Bayes' rule within each bin, we can estimate from PU data the conditional expectation of the true label given the confidence score, which is necessary for the ECE computation. Our theoretical analysis establishes non-asymptotic bias bounds of PU-ECE, which guarantee the precision of our estimator from the TCE in a finite sample. With these bounds, we prove that with an optimal binning strategy, our estimator achieves a convergence rate that matches the convergence rate of the ECE in fully-supervised settings of $\mathcal{O}(n^{-1/3})$, where n is the number of test samples (Futami & Fujisawa, 2024). An information-theoretic generalization error analysis of PU-ECE is also provided by formalizing the functional conditional mutual information (fCMI) (Steinke & Zakynthinou, 2020; Hellström & Durisi, 2022) for a PU setting. In addition, we demonstrate through extensive experiments on synthetic and benchmark datasets that our approach achieves estimation biases comparable to those of the fully-supervised method.

The rest of this paper is organized as follows. In Section 2, we formally describe the problem setting and define notations. In Section 3, we present the PU-ECE method and its theoretical properties, and in Section 4, we present its information-theoretic analysis. Section 5 provides experimental results to validate our approach. Finally, we conclude the paper in Section 6.

2 Problem Formulation

We denote random variables by capital letters and their realizations by lowercase letters. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{0, 1\}$ be the label space. Let $P(X, Y)$ be the probability distribution over $\mathcal{X} \times \mathcal{Y}$, $P(X)$ be the marginal distribution of X, $P(X \mid Y = y)$ be the conditional distribution of X given $Y = y \in \mathcal{Y}$, and $\pi_{\mathrm{P}} = P(Y = 1)$ be the class prior. The hypothesis space is denoted by $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ and a classifier $h \in \mathcal{H}$ outputs a real-valued score for a given input $x \in \mathcal{X}$. The model's confidence score is then obtained by applying an output activation function $\sigma : \mathbb{R} \to [0, 1]$, which we denote by $f(x) := \sigma(h(x))$. We assume σ is measurable and monotone increasing. Examples include the sigmoid function and the probit function. If the sigmoid function $\sigma_{\mathrm{sig}}(z) := (1 + e^{-z})^{-1}$ is used as σ, we refer to $h(x)$ as the logit score. We define $\mathcal{F} = \{\sigma(h(\cdot)) \mid h \in \mathcal{H}\}$ for a fixed σ. We denote the parameterized classifier and its confidence score function by h_w and f_w with $w \in \mathcal{W}$, where $\mathcal{W}$ is the parameter space. If the parameter w is fixed, we simply write h and f. Let $\mathbb{1}_A$ be the indicator function, which returns 1 if the condition A is true and 0 otherwise. We define $[n] = \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$.

2.1 Supervised Learning

In supervised learning, the dataset is a set of independent and identically distributed (i.i.d.) samples drawn from $P(X, Y)$. Let $l : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be the loss function, e.g., the 0-1 loss (Wald, 1945) for the misclassification rate. Let $S_{\mathrm{PN}} := \left\{ \left(X_m^{\mathrm{PN}}, Y_m^{\mathrm{PN}} \right) \right\}_{m=1}^{n_{\mathrm{PN}}}$ be a dataset of n_{PN} samples drawn from $P(X, Y)$. The expected risk of h on $P(X, Y)$ is given by $\mathbb{E}_{X,Y}\left[l(h(X), Y)\right]$. The empirical risk on S_{PN} is given by $\frac{1}{n} \sum_{m=1}^{n} l(h(X_m^{\mathrm{PN}}), Y_m^{\mathrm{PN}})$. Classifiers are trained by the empirical risk minimization (ERM) (Vapnik & Chervonenkis, 1971).

2.2 Calibration Error Estimation

A classifier is said to be perfectly calibrated if its confidence score matches the true class probability (Murphy & Epstein, 1967; Gupta & Ramdas, 2021). Mathematically, this means that for all confidence scores $p \in [0, 1]$, we have $P(Y = 1 \mid f(X) = p) = p$ almost surely (Gupta & Ramdas, 2021). However, this ideal state is rarely achieved in practice (Guo et al., 2017). To quantify the deviation from this perfect calibration, the true calibration error (TCE) (Naeini et al., 2015; Gupta & Ramdas, 2021; Roelofs et al., 2022; Futami & Fujisawa, 2024) is used as a calibration metric.

2.2.1 True Calibration Error (TCE)

The TCE of f on $P(X, Y)$ is defined as

$$\mathrm{TCE}(f) = \mathbb{E}_X \left[\left| \mathbb{E}\left[Y \mid f(X) \right] - f(X) \right| \right].$$

This measures the average absolute difference between the true conditional expectation of Y given $f(X)$ and the confidence score $f(X)$. It quantifies how well the confidence scores align with the true probabilities of the positive class. A lower TCE indicates better calibration, meaning that the confidence scores are closer to the true probabilities. However, calculation of $\mathrm{TCE}(f)$ requires access to the conditional distribution $P(Y \mid f(X))$, which is usually unavailable.

2.2.2 Expected Calibration Error (ECE)

Instead, to estimate $\mathrm{TCE}(f)$, we often use a binning method (Zadrozny & Elkan, 2001; Guo et al., 2017; Futami & Fujisawa, 2024). This method involves a set of $B \in \mathbb{N}$ bins (intervals), denoted by $\mathcal{I} = \{I_b\}_{b=1}^{B}$, that covers the entire range of f. We define the true binned function as

$$f_{\mathcal{I}}(X) = \sum_{b=1}^{B} \mathbb{E}\left[f(X) \mid f(X) \in I_b \right] \mathbb{1}_{f(X) \in I_b}.$$

Then its TCE is calculated as follows (Futami & Fujisawa, 2024):

$$\mathrm{TCE}(f_{\mathcal{I}}) := \sum_{b=1}^{B} P(f(X) \in I_b) \left| \mathbb{E}\left[Y \mid f(X) \in I_b \right] - \mathbb{E}\left[f(X) \mid f(X) \in I_b \right] \right|.$$

$\mathrm{TCE}(f_{\mathcal{I}})$ can be estimated from a supervised evaluation dataset $S_{\mathrm{e}} := \{(X_m^{\mathrm{e}}, Y_m^{\mathrm{e}})\}_{m=1}^{n_{\mathrm{e}}}$ of size $n_{\mathrm{e}} \in \mathbb{N}$, which can be a validation dataset or a test dataset, by the plug-in method. The resulting empirical estimate is called the expected calibration error (ECE):

$$\mathrm{ECE}(f, S_{\mathrm{e}}) := \sum_{b=1}^{B} p_b \left| \bar{y}_{b, S_{\mathrm{e}}} - \bar{f}_{b, S_{\mathrm{e}}} \right|, \tag{1}$$

where $p_b := \frac{\sum_{m=1}^{n_{\mathrm{e}}} \mathbb{1}_{f(X_m^{\mathrm{e}}) \in I_b}}{n_{\mathrm{e}}}$ is the fraction of samples in bin b, $\bar{y}_{b, S_{\mathrm{e}}} := \frac{\sum_{m=1}^{n_{\mathrm{e}}} \mathbb{1}_{f(X_m^{\mathrm{e}}) \in I_b} Y_m^{\mathrm{e}}}{\sum_{m=1}^{n_{\mathrm{e}}} \mathbb{1}_{f(X_m^{\mathrm{e}}) \in I_b}}$ is the average true label in bin b, and $\bar{f}_{b, S_{\mathrm{e}}} := \frac{\sum_{m=1}^{n_{\mathrm{e}}} \mathbb{1}_{f(X_m^{\mathrm{e}}) \in I_b} f(X_m^{\mathrm{e}})}{\sum_{m=1}^{n_{\mathrm{e}}} \mathbb{1}_{f(X_m^{\mathrm{e}}) \in I_b}}$ is the average confidence score in bin b. We assume $n_{\mathrm{e}} \geq 2B$, where n_{e} is the supervised evaluation dataset size and B is the number of bins.

For $\mathcal{I}$, there are two common choices: uniform width binning (UWB) (Guo et al., 2017) and uniform mass binning (UMB) (Zadrozny & Elkan, 2001). In UWB, the bins are defined as $I_b = \left(\frac{b-1}{B}, \frac{b}{B} \right]$ for $b \in [B]$. In UMB, the bins are defined as $I_b = (u_{b-1}, u_b]$ for $b \in [B]$, where $u_0 = 0$, $u_b = f_{(k_b)}$ for $b \in [B-1]$, and $u_B = 1$. Here, $k_b = \lfloor \frac{n_{\mathrm{e}} b}{B} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function that returns the greatest integer less than or equal to the input value, and $f_{(k)}$ denotes the k-th order statistic of the set $\{f(X_m^{\mathrm{e}})\}_{m=1}^{n_{\mathrm{e}}}$.

There are extensions of the TCE and ECE, such as L_p-TCE and L_p-ECE (Kumar et al., 2019; Wenger et al., 2020; Popordanoska et al., 2022), which generalize the TCE and ECE by using the p-th power of the absolute difference. In this paper, we focus on the case of $p = 1$, which is the most commonly used case (Guo et al., 2017; Naeini et al., 2015).

2.3 Positive-Unlabeled Learning

Positive-unlabeled (PU) learning is a classification problem where only positive and unlabeled data are observed, but no negative data. Among various settings of PU learning (Elkan & Noto, 2008; Natarajan et al., 2013; du Plessis et al., 2014; Kato et al., 2019; Bekker et al., 2020), we focus on the selected-completely-at-random (SCAR) and case-control scenario. Concretely, we observe two independent sample sets: a set of unlabeled samples $S_{\text{tr,U}} := \{X_m^{\text{tr,U}}\}_{m=1}^{n_{\text{tr,U}}} \overset{\text{i.i.d.}}{\sim} P(X)$ and a set of positive samples $S_{\text{tr,P}} := \{X_m^{\text{tr,P}}\}_{m=1}^{n_{\text{tr,P}}} \overset{\text{i.i.d.}}{\sim} P(X \mid Y = 1)$, with the full training data denoted by $S_{\text{tr}} = (S_{\text{tr,U}}, S_{\text{tr,P}})$ (du Plessis et al., 2014). This SCAR and case-control scenario is commonly used in the PU learning literature (Natarajan et al., 2013; du Plessis et al., 2014; Kiryo et al., 2017; Yao et al., 2021). Throughout the paper, we assume that the class prior $\pi_{\text{P}} = P(Y = 1)$ is known—in practice, it can be estimated from data (Blanchard et al., 2010; Garg et al., 2021; Zhu et al., 2023).

In PU learning, an unbiased risk estimator of h on S_{tr} is given as follows (Natarajan et al., 2013; du Plessis et al., 2014; 2015):

$$\mathcal{L}_{\text{uPU}}(h) = \frac{\pi_{\text{P}}}{n_{\text{tr,P}}} \sum_{m=1}^{n_{\text{tr,P}}} \left\{ l(h(X_m^{\text{tr,P}}), 1) - l(h(X_m^{\text{tr,P}}), 0) \right\} + \frac{1}{n_{\text{tr,U}}} \sum_{m=1}^{n_{\text{tr,U}}} l(h(X_m^{\text{tr,U}}), 0). \tag{2}$$

The estimator is suitable for estimating the expected risk $\mathbb{E}[l(h(X), Y)]$ or training simple classifiers (du Plessis et al., 2014; 2015), but it does not work well for training more complex classifiers such as deep neural networks due to the *negative risk* issue (Kiryo et al., 2017). To overcome this issue, non-negative risk estimators were proposed (Kiryo et al., 2017; Lu et al., 2020). The non-negative risk estimator is given by

$$\mathcal{L}_{\text{nnPU}}(h) = \frac{\pi_{\text{P}}}{n_{\text{tr,P}}} \sum_{m=1}^{n_{\text{tr,P}}} l(h(X_m^{\text{tr,P}}), 1) + l_{\text{cc}} \left(\frac{1}{n_{\text{tr,U}}} \sum_{m=1}^{n_{\text{tr,U}}} l(h(X_m^{\text{tr,U}}), 0) - \frac{\pi_{\text{P}}}{n_{\text{tr,P}}} \sum_{m=1}^{n_{\text{tr,P}}} l(h(X_m^{\text{tr,P}}), 0) \right), \tag{3}$$

where $l_{\text{cc}} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is called a consistent-correction function. Several functions have been proposed for use as l_{cc}, e.g., $\max\{x, 0\}$ (Kiryo et al., 2017; Lu et al., 2020).

3 PU-ECE: An Estimator for Calibration Error in PU Learning

Conventional ECE estimation, as shown in Eq. (1), requires labeled data (both positive and negative), but in the PU environment, negative data are unavailable, as introduced in Section 2.3. The main challenges are (i) estimating the true positive rate within bins (ii) the data are not from a single distribution $P(X, Y)$ but from $P(X)$ and $P(X|Y = 1)$. In this section, we propose a method to estimate the ECE in the PU environment. The proposed method is based on the idea of using positive and unlabeled data to estimate $\mathbb{E}[Y \mid f(X) \in I_b]$. We denote the PU evaluation data as $S_{\text{U}} := \{X_m^{\text{U}}\}_{m=1}^{n_{\text{U}}}$, $S_{\text{P}} := \{X_m^{\text{P}}\}_{m=1}^{n_{\text{P}}}$, and $S_{\text{PU}} := (S_{\text{U}}, S_{\text{P}})$. The evaluation data are independent of the training data and f, and can be a validation dataset or a test dataset in practice.

3.1 Proposed Estimator

Since $Y \in \{0, 1\}$, the conditional expectation can be expressed as

$$\mathbb{E}[Y \mid f(X) \in I_b] = P(Y = 1 \mid f(X) \in I_b) = \frac{P(Y = 1, f(X) \in I_b)}{P(f(X) \in I_b)} = \frac{\pi_{\text{P}} P(f(X) \in I_b \mid Y = 1)}{P(f(X) \in I_b)}. \tag{4}$$

In the PU learning setting, we have access to positive samples $S_{\text{P}} = \{X_m^{\text{P}}\}_{m=1}^{n_{\text{P}}}$ drawn i.i.d. from $P(X|Y = 1)$ and unlabeled samples $S_{\text{U}} = \{X_m^{\text{U}}\}_{m=1}^{n_{\text{U}}}$ drawn i.i.d. from $P(X)$. The term $P(f(X) \in I_b \mid Y = 1)$ in Eq. (4) can be estimated by the empirical frequency of positive samples falling into bin I_b: $p_{b,\text{P}} := \frac{1}{n_{\text{P}}} \sum_{m=1}^{n_{\text{P}}} \mathbb{1}_{f(X_m^{\text{P}}) \in I_b}$. Similarly, $P(f(X) \in I_b)$ can be estimated by the empirical frequency of unlabeled samples falling into bin I_b: $p_{b,\text{U}} := \frac{1}{n_{\text{U}}} \sum_{m=1}^{n_{\text{U}}} \mathbb{1}_{f(X_m^{\text{U}}) \in I_b}$. These estimators are unbiased, that is,

$\mathbb{E}\left[p_{b,\mathrm{P}}\right] = P(f(X) \in I_b \mid Y = 1)$ and $\mathbb{E}\left[p_{b,\mathrm{U}}\right] = P(f(X) \in I_b)$. Let $|I_{b,\mathrm{P}}| = \sum_{m=1}^{n_\mathrm{P}} \mathbb{1}_{f(X_m^\mathrm{P}) \in I_b}$ and $|I_{b,\mathrm{U}}| = \sum_{m=1}^{n_\mathrm{U}} \mathbb{1}_{f(X_m^\mathrm{U}) \in I_b}$ be the number of positive and unlabeled samples in bin I_b, respectively. Then $p_{b,\mathrm{P}} = |I_{b,\mathrm{P}}|/n_\mathrm{P}$ and $p_{b,\mathrm{U}} = |I_{b,\mathrm{U}}|/n_\mathrm{U}$. Using these estimates, an estimator for the conditional probability $P(Y = 1 \mid f(X) \in I_b)$ is naively obtained as $\hat{y}_{b,S_\mathrm{PU}} := \frac{\pi_\mathrm{P} p_{b,\mathrm{P}}}{p_{b,\mathrm{U}}}$. The ECE for PU learning (PU-ECE) can then be estimated as

$$\sum_{b=1}^{B} p_{b,\mathrm{U}} \left| \hat{y}_{b,S_\mathrm{PU}} - \bar{f}_{b,S_\mathrm{PU}} \right|,$$

where $\bar{f}_{b,S_\mathrm{PU}} := \frac{\sum_{m=1}^{n_\mathrm{U}} \mathbb{1}_{f(X_m^\mathrm{U}) \in I_b} f(X_m^\mathrm{U})}{|I_{b,\mathrm{U}}|}$ is the average confidence score for unlabeled samples in bin I_b. If $|I_{b,\mathrm{U}}| = 0$, then $p_{b,\mathrm{U}} = 0$. In this case, $\hat{y}_{b,S_\mathrm{PU}}$ would involve division by zero. However, this is a removable singularity. For all cases where $p_{b,\mathrm{U}} > 0$, we can algebraically reformulate the estimator as follows:

$$\mathrm{ECE}_\mathrm{PU}(f, S_\mathrm{PU}) = \sum_{b=1}^{B} \left| \frac{\pi_\mathrm{P}}{n_\mathrm{P}} \sum_{m=1}^{n_\mathrm{P}} \mathbb{1}_{f(X_m^\mathrm{P}) \in I_b} - \frac{1}{n_\mathrm{U}} \sum_{m=1}^{n_\mathrm{U}} \mathbb{1}_{f(X_m^\mathrm{U}) \in I_b} f(X_m^\mathrm{U}) \right|.$$

This new form is mathematically equivalent for all $p_{b,\mathrm{U}} > 0$, and remains well-defined at $p_{b,\mathrm{U}} = 0$, and its value at $p_{b,\mathrm{U}} = 0$ coincides with the limit of the estimator as $p_{b,\mathrm{U}} \to 0$.

We define UWB and UMB for the PU setting as follows: in UWB, the bins are defined in the same way as in supervised learning, while in UMB, the bins are defined by *the unlabeled data*; $I_b = (u_{b-1}, u_b]$ for $b \in [B]$, where $u_0 = 0$, $u_b = f_{(k_b)}$ for $b \in [B-1]$, and $u_B = 1$. Here, $k_b = \lfloor \frac{n_\mathrm{U} b}{B} \rfloor$, and $f_{(k)}$ denotes the k-th order statistic of the set $\{f(X_m^\mathrm{U})\}_{m=1}^{n_\mathrm{U}}$, i.e., the k-th smallest value.

3.2 Bias Analysis of PU-ECE

We analyze the bias of the proposed estimator given the evaluation data. In the analysis, we assume the following:

Assumption 1. *The number of positive samples n_P satisfies $n_\mathrm{P} \geq 1$. The number of unlabeled samples n_U satisfies $n_\mathrm{U} \geq 2B$ for UMB and $n_\mathrm{U} \geq 1$ for UWB.*

Assumption 2. *$f(X)$ is absolutely continuous with respect to the Lebesgue measure, i.e., $f(X)$ has a probability density function.*

Assumption 3. *$\mathbb{E}\left[Y \mid f(X)\right]$ satisfies L-Lipschitz continuity with respect to $f(X)$, i.e., $|\mathbb{E}\left[Y \mid f(X) = v\right] - \mathbb{E}\left[Y \mid f(X) = v'\right]| \leq L|v - v'|$ for all $v, v' \in [0, 1]$.*

The (total) bias of the PU-ECE estimator is defined as

$$\mathrm{Bias}_\mathrm{tot}(f, S_\mathrm{PU}) := |\mathrm{ECE}_\mathrm{PU}(f, S_\mathrm{PU}) - \mathrm{TCE}(f)|.$$

Then, we have the following theorem, which provides non-asymptotic bias bounds for the PU-ECE estimator.

Theorem 1. *Under Assumptions 1, 2 and 3, we have*

$$\mathbb{E}_{S_\mathrm{PU}}\left[\mathrm{Bias}_\mathrm{tot}(f, S_\mathrm{PU})\right] \leq \begin{cases} \frac{1+L}{B} + \sqrt{2\left(\frac{\pi_\mathrm{P}^2}{n_\mathrm{P}} + \frac{1}{n_\mathrm{U}}\right) B \log 2} & \text{(for UWB)}, \\ \frac{1+L}{B} + \frac{(3+2L)B}{n_\mathrm{U} - B} + \sqrt{2B \log 2}\left(\frac{1+L}{\sqrt{n_\mathrm{U} - B}} + \sqrt{\frac{\pi_\mathrm{P}^2}{n_\mathrm{P}} + \frac{1}{n_\mathrm{U} - B}}\right) & \text{(for UMB)}. \end{cases}$$

Furthermore, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over S_PU, we have

$$\mathrm{Bias}_\mathrm{tot}(f, S_\mathrm{PU}) \leq \begin{cases} \frac{1+L}{B} + \sqrt{2\left(\frac{\pi_\mathrm{P}^2}{n_\mathrm{P}} + \frac{1}{n_\mathrm{U}}\right)(B \log 2 + \log(1/\delta))} & \text{(for UWB)}, \\ \frac{1+L}{B} + \frac{(3+2L)B}{n_\mathrm{U} - B} + \sqrt{2(B \log 2 + \log(1/\delta))}\left(\frac{1+L}{\sqrt{n_\mathrm{U} - B}} + \sqrt{\frac{\pi_\mathrm{P}^2}{n_\mathrm{P}} + \frac{1}{n_\mathrm{U} - B}}\right) & \text{(for UMB)}. \end{cases}$$

The proof of Theorem 1 is provided in Appendix A.1. Briefly, the proof is based on decomposing the total bias into two parts: the bias due to binning and the bias due to the estimation by the empirical frequency, and then bounding each part separately by concentration inequalities.

Minimizing the upper bound of the total bias with respect to B, we have the following upper bound orders:

Corollary 1. *Under Assumption 1, the scale of the optimal B that minimizes the upper bound in Theorem 1 for both UWB and UMB is given by*

$$B = \Theta\left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{1}{n_{\mathrm{U}}} \right)^{-\frac{1}{3}} \right),$$

and the corresponding expected total bias convergence rate is given by

$$\mathbb{E}_{S_{\mathrm{PU}}}\left[\mathrm{Bias}_{\mathrm{tot}}(f, S_{\mathrm{PU}}) \right] = \mathcal{O}\left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{1}{n_{\mathrm{U}}} \right)^{\frac{1}{3}} \right).$$

The proof can be found in Appendix A.2. In practice, this means that to minimize the bias of the PU-ECE estimator, the number of bins B should be chosen according to the formula above, balancing the sample sizes of the positive and unlabeled data. The result corresponds to the convergence rate of the supervised setting; the optimal B is $\mathcal{O}(n^{\frac{1}{3}})$ and the expected total bias is $\mathcal{O}(n^{-\frac{1}{3}})$, where n is the size of the fully-labeled evaluation data (Futami & Fujisawa, 2024). The term π_{P}^2 in these expressions is a constant factor that does not depend on the sample size, but it indicates that a smaller class prior π_{P} leads to a smaller expected bias.

In practical PU learning scenarios, the true class prior π_{P} is often unknown and must be estimated from data (Bepler et al., 2019; Ito & Sugiyama, 2023), which introduces additional estimation error into the calibration process. It is therefore important to understand how inaccuracies in class prior estimation affect PU-ECE. We also analyze how the class prior estimation error affects the PU-ECE estimation.

Theorem 2 (Error bound due to class prior estimation error). *Let $\hat{\pi}_{\mathrm{P}}$ be an estimate of the class prior π_{P}, and define the PU-ECE estimator using $\hat{\pi}_{\mathrm{P}}$ as*

$$\widehat{\mathrm{ECE}}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) = \sum_{b=1}^{B} \left| \frac{\hat{\pi}_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{m=1}^{n_{\mathrm{P}}} \mathbb{1}_{f(X_m^{\mathrm{P}}) \in I_b} - \frac{1}{n_{\mathrm{U}}} \sum_{m=1}^{n_{\mathrm{U}}} f(X_m^{\mathrm{U}}) \mathbb{1}_{f(X_m^{\mathrm{U}}) \in I_b} \right|.$$

Then, we have the following bound on the difference between the PU-ECE with the true prior and that with the estimated prior:

$$\left| \mathrm{ECE}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) - \widehat{\mathrm{ECE}}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) \right| \leq \left| \pi_{\mathrm{P}} - \hat{\pi}_{\mathrm{P}} \right|.$$

The proof is provided in Appendix A.3. This inequality does not require expectation and holds for any f and S_{PU}. This theorem shows that when the class prior is unknown and must be estimated, the calibration error estimation incurs an additional error of at most the class prior estimation error.

4 Information-Theoretic Generalization Analysis of PU-ECE

The previous analysis focused on the bias of the ECE estimation using evaluation data, which is independent of the training data used for learning f_w. However, in practice, we may recalibrate classifiers using a recalibration dataset and then evaluate the PU-ECE on the same recalibration dataset (Kumar et al., 2019), or we may calibrate classifiers using the training dataset (Futami & Fujisawa, 2024). In these cases, the data used for PU-ECE evaluation and the classifier recalibration/training are no longer independent. Therefore, analyzing the generalization error of the PU-ECE is crucial to ensure reliable model selection and evaluation in practice. In this section, we extend the conditional mutual information (CMI) framework (Steinke & Zakynthinou, 2020; Haghifam et al., 2021) used in supervised learning to the PU setting and analyze the generalization error of PU-ECE using it.

4.1 Information-Theoretic Generalization Error Analysis for Supervised Learning

A rigorous understanding of the generalization error is crucial for evaluating the reliability of machine learning algorithms. Classical paradigms for assessing generalization include the concepts of Vapnik-Chervonenkis (VC) dimension (Vapnik & Chervonenkis, 1968; Blumer et al., 1989) and metric entropy (Kolmogorov & Tikhomirov, 1959; Dudley, 1974). The VC dimension quantifies how many points a hypothesis class can shatter, while metric entropy measures how finely a function class can be approximated using a finite set of representative functions at a given precision. Despite the established significance of these tools, recent advances in information-theoretic perspectives, particularly those centered on CMI (Steinke & Zakynthinou, 2020; Haghifam et al., 2021), have provided new insights into generalization. Intuitively, CMI quantifies the amount of information that a model retains about the training data after observing a test sample. This offers an alternative view for analyzing the generalization of a learned model. By bounding this data-model dependency, CMI-based analysis methods provide new insights into why certain learning algorithms exhibit strong generalization performance. For instance, Futami & Fujisawa (2024) applied this approach to derive non-asymptotic generalization error bounds for the ECE in the supervised setting.

4.2 Conditional Mutual Information (CMI) for PU Learning

To analyze the generalization error, we extend the conditional mutual information (CMI) framework to the PU setting. We define the functional conditional mutual information (fCMI) for PU data, denoted by fCMI, which measures the information between the model's confidence score and the PU training data. We also define a variant, fCMI_U, which considers only the unlabeled training data. The formal definitions of these quantities are technical and are detailed in Appendix B. By the data processing inequality (Cover & Thomas, 2006), we have $\text{fCMI}_\text{U} \leq \text{fCMI}$. The PU CMI metrics can be bounded by data-independent model complexity measures. For instance, if the model has a VC dimension d_VC, $\text{fCMI} \leq \max\{(d_\text{VC}+1)\log 2, d_\text{VC}\log(2e(n_\text{tr,U}+n_\text{tr,P})/d_\text{VC})\}$ (Steinke & Zakynthinou, 2020; Harutyunyan et al., 2021). If $f_{\mathcal{A}(\tilde{X}_M, R)}$ is a d-dimensional parametric function that is L-Lipschitz continuous on $\|\cdot\|_\infty$ and whose domain is $[0,1]^d$ and the function class has an ϵ-cover, then $\text{fCMI} \lesssim d\log(L/\epsilon)$ and $\text{fCMI}_\text{U} \lesssim d\log(L/\epsilon)$ (Wainwright, 2019; Futami & Fujisawa, 2024).

4.3 Generalization Bound for PU-ECE

This subsection analyzes the generalization error of PU-ECE using the fCMIs defined above. We follow Futami & Fujisawa (2024) and use $\text{Bias}_\text{tot}(f_W, S_\text{tr}) := |\text{TCE}(f_W) - \text{ECE}_\text{PU}(f_W, S_\text{tr})|$ as the generalization error of $\text{ECE}_\text{PU}(f_W, S_\text{tr})$, where we denote by $W = \mathcal{A}(S_\text{tr}, R) \in \mathcal{W}$ the parameters of f_W learned by a randomized algorithm $\mathcal{A}$ on the training data S_tr and the source of randomness R.

Theorem 3 (Expected generalization error bound of PU-ECE). *Under Assumptions 2 and 3, we have*

$$\mathbb{E}_{R,S_\text{tr}}\left[\text{Bias}_\text{tot}(f_W, S_\text{tr})\right]$$

$$\leq \begin{cases} \frac{1+L}{B} + \sqrt{8\left(\frac{\pi_\text{P}^2}{n_\text{tr,P}} + \frac{1}{n_\text{tr,U}}\right)\{\text{fCMI} + B\log 2\}} & \text{(for UWB)}, \\ \frac{1+L}{B} + \sqrt{8\left(\frac{\pi_\text{P}^2}{n_\text{tr,P}} + \frac{1}{n_\text{tr,U}}\right)\{\text{fCMI} + B\log 2\}} + (1+L)\sqrt{\frac{8}{n_\text{tr,U}}\{\text{fCMI}_\text{U} + B\log 2\}} & \text{(for UMB)}. \end{cases}$$

The proof of Theorem 3 is provided in Appendix C.1. Briefly, the proof is similar to the proof of Theorem 1, except the dependency on the training data is handled by fCMI.

Similar to Corollary 1, the optimal bin size to minimize the upper bound in Theorem 3 and the order of the upper bound can also be derived if a learning algorithm has a sufficiently small fCMI. For instance, such an assumption holds when $\mathcal{H}$ is a finite VC class since $\mathcal{F}$ is also a finite VC class, and hence fCMI is bounded by $\mathcal{O}(\log(n_\text{tr,P} + n_\text{tr,U}))$ (Steinke & Zakynthinou, 2020). For such a case, we can derive the following result.

Corollary 2 (Total bias upper bound order). *Under Assumption 2 and 3 and assuming* $\text{fCMI} = \mathcal{O}\left(\left(\frac{\pi_\text{P}^2}{n_\text{tr,P}} + \frac{1}{n_\text{tr,U}}\right)^{-1/3}\right)$, *the scale of the optimal B that minimizes the upper bound in Theorem 3 is given*

Dataset	$n_{\mathrm{tr,U}}$	π_{P}	Model	Learning rate
MNIST (LeCun et al., 1998)	60,000	0.49	2-layer MLP (784-100-100-1)	1.0×10^{-3}
CIFAR-10 (Krizhevsky, 2009)	50,000	0.40	ResNet-18	1.0×10^{-5}

Table 1: Specification of benchmark datasets, models, and learning rates. $n_{\mathrm{tr,U}}$ denotes the number of unlabeled samples for training PU classifiers.

by

$$B = \Theta \left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}} \right)^{-1/3} \right),$$

and the corresponding expected generalization error convergence rate is given by

$$\mathbb{E}_{R,S_{\mathrm{tr}}} \left[\mathrm{Bias}_{\mathrm{tot}}(f_W, S_{\mathrm{tr}}) \right] = \mathcal{O} \left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}} \right)^{1/3} \right).$$

The proof is provided in Appendix A.2.

By Corollaries 1 and 2, $\mathbb{E}_{R,S_{\mathrm{tr}}} \left[\mathrm{Bias}_{\mathrm{tot}}(f_W, S_{\mathrm{tr}}) \right] \ll \mathbb{E}_{S_{\mathrm{PU}}} \left[\mathrm{Bias}_{\mathrm{tot}}(f_W, S_{\mathrm{PU}}) \right]$ if $n_{\mathrm{tr,P}} \gg n_{\mathrm{P}}$ and $n_{\mathrm{tr,U}} \gg n_{\mathrm{U}}$ and the classifier generalizes well. This would actually happen in practice since the training dataset is often much larger than the evaluation dataset. Such a property is also observed in the supervised setting (Futami & Fujisawa, 2024). This represents an important future direction for PU learning to develop a method that jointly optimizes the TCE and accuracy, although this is beyond the scope of this paper.

5 Experiments

In this section, we empirically evaluate the theoretical properties of the proposed PU-ECE estimator. We first present a motivational study that highlights the unique challenges of calibration in PU learning (Section 5.2). We then empirically validate the theoretical bias and convergence properties of our proposed PU-ECE estimator on both synthetic and real-world benchmark datasets (Section 5.3).

5.1 Common Setup

For all experiments, we used UMB since UMB has demonstrated smaller biases than UWB (Roelofs et al., 2022). The sigmoid function was used as the output activation function. The code for all experiments was implemented in Python and is available at `https://github.com/<will-be-released>`.

Synthetic data Following previous work (Vaicenavicius et al., 2019; Zhang et al., 2020; Futami & Fujisawa, 2024), we generated data from a logistic model where $P(Y = 1 \mid X = x) = \sigma_{\mathrm{sig}}(2x)$, with the conditional probability densities $p(X \mid Y = 1) = \mathcal{N}(x; 1, 1)$ and $p(X \mid Y = 0) = \mathcal{N}(x; -1, 1)$. We evaluated a classifier $f(x) = \sigma_{\mathrm{sig}}(\beta_0 + \beta_1 x)$ under two settings: a less-calibrated case ($\beta_0 = -0.5, \beta_1 = 1.5$) and a better-calibrated case ($\beta_0 = -0.2, \beta_1 = 1.9$). The ground-truth $\mathrm{TCE}(f)$ was computed by the trapezoidal rule. Further details are provided in Appendix D.

Benchmark data We used the MNIST and CIFAR-10 datasets, with specifications detailed in Table 1. Since these datasets are multiclass, we converted them into binary classification tasks by selecting a positive class and treating all other classes as negative. For MNIST, we used the digits 5-9 as the positive class, while for CIFAR-10, we selected the vehicle class (airplane, automobile, ship, and truck) as the positive class. Classifiers were trained using the standard ERM for supervised learning and the nnPU learning method (Kiryo et al., 2017) for PU learning, and the outputs of the classifiers were used as logit scores. The logistic loss (a.k.a. binary cross-entropy loss) $\log(\sigma_{\mathrm{sig}}(-(2y - 1)f(x)))$ or sigmoid loss $\sigma_{\mathrm{sig}}(-(2y - 1)f(x))$ was used as the loss function, depending on the experiment. The logistic loss was used as a strictly proper loss function in supervised learning (Zadrozny & Elkan, 2001). The sigmoid loss

was used as a noise-robust loss function (Ghosh et al., 2017) and classification calibrated (Charoenphakdee et al., 2019), and hence are widely used in weakly-supervised learning including PU learning (Kiryo et al., 2017; Chen et al., 2020; Zhao et al., 2022; Ye et al., 2023). Although the sigmoid loss is not strictly proper (Charoenphakdee et al., 2019), its confidence scores are conventionally interpreted as class-posterior probabilities (Platt, 1999; Chen et al., 2020; Ye et al., 2023). The other details are provided in Appendix F.

5.2 Motivational Study

We first conducted a controlled study to compare the ECE of classifiers trained from fully-supervised data and PU data to see how calibration evolves with the number of training samples. MNIST was used as the dataset. We evaluated the classifiers trained under four conditions: supervised learning with logistic loss (PN-logistic), supervised learning with sigmoid loss (PN-sigmoid), PU learning trained with the logistic loss (PU-logistic), and PU learning with the sigmoid loss (PU-sigmoid). Results are averaged over 100 trials, each with a different random seed. The number of unlabeled samples is fixed at $60,000$.

As shown in Figure 1, supervised methods improve calibration steadily with more data regardless of the loss function. As expected, the logistic loss leads to better calibration than the sigmoid loss in supervised learning. In contrast, PU learning behaves very differently: PU-sigmoid is better calibrated than PU-logistic.

This collapse is likely due to the non-negative correction in the nnPU risk estimator (Eq. (3)). When the logistic loss is used, the nnPU risk estimator tends to cram the logit scores of positive samples into a small constant value so that the non-negative correction has little effect for positive samples, which
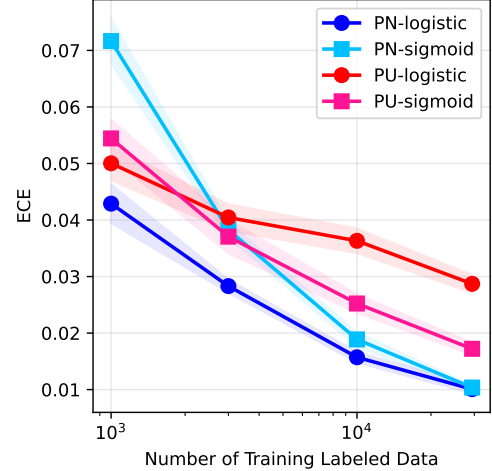

Figure 1: Comparison of ECE for supervised and PU learning on MNIST. The x-axis represents the number of labeled samples used for training. For PU learning, this corresponds to the number of positive samples. The y-axis is the ECE on the test set. The lines represent the ECE values and the shaded areas represent the standard error.

leads to poor calibration. As shown in Figure 2, when trained with the logistic loss, the logit scores for positive examples collapse toward a constant, yielding high ECE. These findings demonstrate that, in practice, PU learning is considerably more prone to miscalibration than standard supervised learning, especially when using the logistic loss, which is strictly proper in supervised learning.

5.3 Validation of Bias and Convergence

Let N denote an integer ranging from 10 to $10,000$. We evaluated the convergence behavior by varying N, comparing PU-ECE in settings with $(n_\mathrm{P} = N, n_\mathrm{U} = \infty)$, $(n_\mathrm{P} = \infty, n_\mathrm{U} = N)$, $(n_\mathrm{P} = N, n_\mathrm{U} = 10N)$, and standard ECE with $n = N$. For settings with either $n_\mathrm{U} = \infty$ or $n_\mathrm{P} = \infty$, the corresponding components (e.g., $P(f(X) \in I_b)$ when $n_\mathrm{U} = \infty$) were computed via numerical integration using the trapezoidal rule with up to 10,000 points.

We also investigated the effect of fixing n_P by comparing PU-ECE with $(n_\mathrm{P} = 100, n_\mathrm{U} = N)$, $(n_\mathrm{P} = 1,000, n_\mathrm{U} = N)$, $(n_\mathrm{P} = 10,000, n_\mathrm{U} = N)$, and $(n_\mathrm{P} = N, n_\mathrm{U} = \infty)$. The number of bins B was set to $\lceil (\pi_\mathrm{P}^2/n_\mathrm{P} + 1/n_\mathrm{U})^{-1/3} \rceil$ for PU-ECE and $n^{1/3}$ or $(n/\pi_\mathrm{P}^2)^{1/3}$ for ECE, following the optimal order derived in our analysis and in Futami & Fujisawa (2024).

We trained the classifiers on the PU training data with 10,000 positive and $n_\mathrm{tr,U}$ unlabeled samples specified in Table 1. Results are averaged over 100 trials, each with a different random seed.

To compute $\mathrm{TCE}(f)$ for these datasets, we first need to assume parametric models for the true conditional probability densities. Fortunately, the TCE can be computed using the distribution of the logit score $h(X)$ or the confidence score $f(X)$, without modeling the distribution of X, which would otherwise suffer from the curse of dimensionality (see Appendix E for details). We observed that the confidence scores were heavily
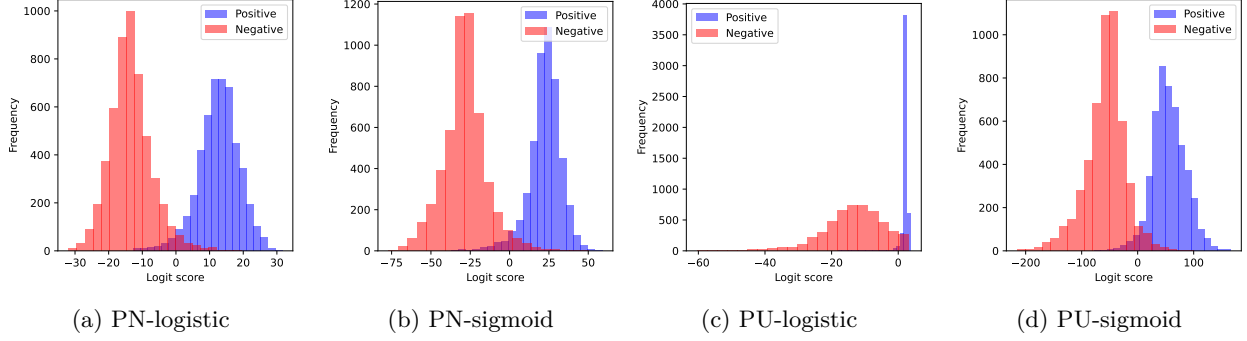
Figure 2: Histogram of logit scores for different training settings on MNIST test data. The x-axis is the logit and the y-axis is the frequency.
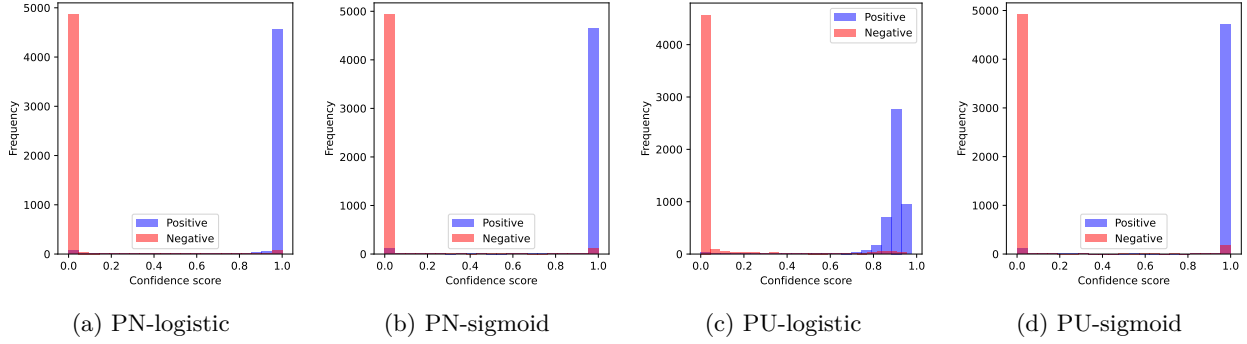

Figure 3: Histogram of the confidence scores for different training settings on MNIST test data. The x-axis is the confidence score and the y-axis is the frequency.

saturated near 0 or 1, while the logit scores were spread out (Figure 2 and 3). Therefore, we modeled the logit score $h(X)$ instead of the confidence score $f(X)$, specifically modeling the distributions of the logit score, $P(h(X) \mid Y = 1)$ and $P(h(X) \mid Y = 0)$. The empirical logit histograms were multi-modal, so a single-mode parametric family, such as the logistic-beta distribution used by Roelofs et al. (2022), is inadequate. For instance, the MNIST logits (Fig. 2) exhibited two clear peaks. Instead, we fitted Gaussian Mixture Models (GMMs) to $P(h(X) \mid Y = 1)$ and $P(h(X) \mid Y = 0)$, selecting the number of components for each GMM from 1 to 10 via the Bayesian Information Criterion (BIC) (Schwarz, 1978). In MNIST, the GMMs with 4 positive class components and 3 negative class components were selected, while in CIFAR-10, the GMMs with 2 positive class components and 2 negative class components were selected. With these estimated densities, we sampled the logit scores from the GMMs. TCE(f) was calculated using the trapezoidal rule at $10,000$ points.

Figure 4 shows the total bias of PU-ECE and standard ECE (absolute bias compared to the TCE) on both synthetic and benchmark datasets. This experiment is designed to validate the theoretical convergence rate of our estimator. According to our analysis, the total bias of PU-ECE is expected to decrease at a rate of $\mathcal{O}((\pi_P^2/n_P + 1/n_U)^{-1/3})$. When both n_P and n_U grow proportionally with N (or are set to ∞), this simplifies to $\mathcal{O}(N^{-1/3})$.

The empirical results clearly exhibit this $\mathcal{O}(N^{-1/3})$ convergence, providing strong evidence in support of our theoretical claims. Notably, this rate matches the known optimal convergence rate for standard ECE (Futami & Fujisawa, 2024), demonstrating that PU-ECE can achieve competitive calibration performance even without access to fully labeled data.

Figure 5 investigates the effect of fixing n_P while increasing n_U. We observe that PU-ECE converges to the limiting case of PU-ECE with $n_P = N, n_U = \infty$ but not to the true TCE. This behavior is consistent with

our theoretical prediction: even as $n_U \to \infty$, the bias remains bounded below by the term depending on n_P. This residual error creates an error floor that prevents convergence to the TCE when n_P is finite.

These results highlight a key practical insight: sufficiently large sets of both positive and unlabeled samples are therefore essential for effectively estimating the TCE.
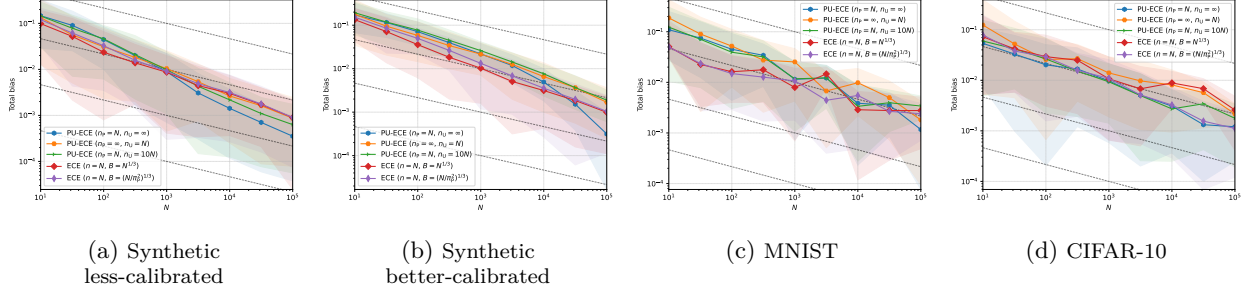

(a) Synthetic less-calibrated (b) Synthetic better-calibrated (c) MNIST (d) CIFAR-10

Figure 4: Total bias of ECE and PU-ECE, as a function of sample size N. Both axes are log-scale. The lines represent the total bias, and the shaded areas represent the 90% percentile interval. The dashed lines represent the theoretical convergence rate of $\mathcal{O}(N^{-1/3})$.

(a) Synthetic less-calibrated (b) Synthetic better-calibrated (c) MNIST (d) CIFAR-10

Figure 5: Convergence of total bias of PU-ECE varying n_P. Both axes are log-scale. The lines represent the total bias, and the shaded areas represent the 90% percentile interval. The dashed lines represent the theoretical convergence rate of $\mathcal{O}(N^{-1/3})$.
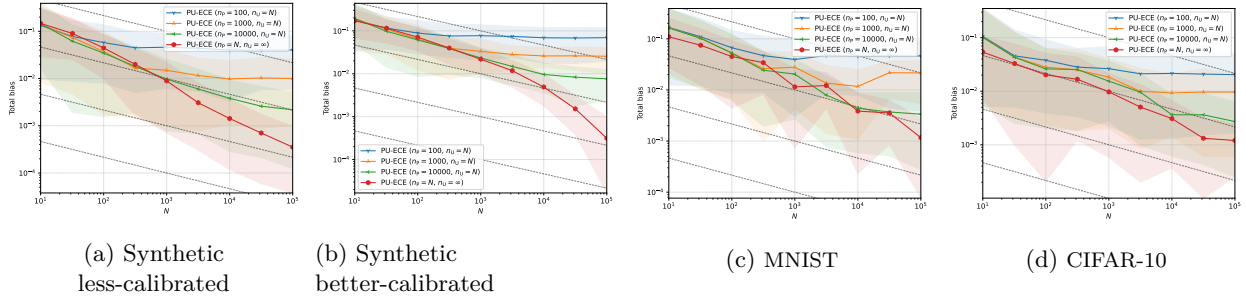
6 Conclusions

We proposed PU-ECE, the first ECE estimator that can be computed from PU data. Our theoretical analysis provides non-asymptotic bias bounds for PU-ECE and shows that, with an optimal binning strategy, the bias of our estimator converges to the TCE at a rate of $\mathcal{O}((\pi_P^2/n_P + 1/n_U)^{1/3})$. This rate matches the convergence rate of the standard ECE in the supervised setting. We further extended the CMI framework to PU learning, deriving generalization error bounds with the same convergence rate as the bias. Experiments on synthetic and benchmark datasets corroborated the theory: PU-ECE performs comparably to the standard ECE, despite lacking negative labels. They also revealed a practical caveat—if the number of positive samples is fixed and the number of unlabeled samples increases, PU-ECE does not converge to the true TCE, which highlights the importance of having sufficient positive data as well as unlabeled data for accurate estimation.

References

Kendall Atkinson. *An introduction to numerical analysis.* John Wiley & Sons, New York, 2 edition, January 1989.

Alejandro Correa Bahnsen, Aleksandar Stojanovic, Djamila Aouada, and Björn Ottersten. Improving credit card fraud detection with calibrated probabilities. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 677–685, Philadelphia, PA, April 2014. Society for Industrial and Applied Mathematics.

Jessa Bekker, Pieter Robberechts, and Jesse Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Machine Learning and Knowledge Discovery in Databases*, pp. 71–85. Springer International Publishing, 2020.

Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 16(11):1153–1160, November 2019.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(99):2973–3009, 2010.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, October 1989.

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 961–970. PMLR, January 2019.

Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, and Zhangyang Wang. Self-PU: Self boosted and calibrated positive-unlabeled training. In Hal Daumé Iii and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1510–1519. PMLR, July 2020.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory: Cover/elements of information theory, second edition.* John Wiley & Sons, Nashville, TN, 2 edition, June 2006.

Alexander P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379): 605–610, 1982.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Monroe D. Donsker and Srinivasa R. S. Varadhan. On a variational formula for the principal eigenvalue for operators with maximum principle. *Proceedings of the National Academy of Sciences of the United States of America*, 72(3):780–783, 1975.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, May 2021.

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

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1386–1394, Lille, France, 2015. PMLR.

Richard M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236, 1974.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, New York, NY, USA, August 2008. ACM.

Futoshi Futami and Masahiro Fujisawa. Information-theoretic generalization analysis for expected calibration error. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 84246–84297. Curran Associates, Inc., 2024.

Saurabh Garg, Yifan Wu, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. Mixture proportion estimation and PU learning:a modern approach. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8532–8544. Curran Associates, Inc., 2021.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, February 2017.

Irving J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14 (1):107–114, 1952.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, August 2017.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3942–3952. PMLR, July 2021.

Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26370–26381. Curran Associates, Inc., 2021.

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy S. Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, pp. 24670–24682. Curran Associates, Inc., 2021.

Fredrik Hellström and Giuseppe Durisi. A new family of generalization bounds using samplewise evaluated cmi. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10108–10121. Curran Associates, Inc., 2022.

Nobutaka Ito and Masashi Sugiyama. Audio signal enhancement with learning from positive and unlabeled data. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

Kristen Jaskie and Andreas Spanias. Positive and unlabeled learning algorithms and applications: A survey. In *2019 10th International Conference on Information, Intelligence, Systems and Applications*, pp. 1–8, 2019.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, March 2012.

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*, May 2019.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Andrey N. Kolmogorov and Vladimir M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14:3–86, 1959.

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

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Ananya Kumar, Percy S. Liang, and Tengyu Ma. Verified uncertainty calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Fuyi Li, Shuangyu Dong, André Leier, Meiya Han, Xudong Guo, Jing Xu, Xiaoyu Wang, Shirui Pan, Cangzhi Jia, Yang Zhang, Geoffrey I Webb, Lachlan J M Coin, Chen Li, and Jiangning Song. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Briefings in Bioinformatics*, 23(1): bbab461, January 2022.

Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. Partially supervised classification of text documents. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 387–394. Morgan Kaufmann Publishers Inc., July 2002.

Nan Lu, Tianyi Zhang, Gang Niu, and Masashi Sugiyama. Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1115–1125. PMLR, August 2020.

Allan H. Murphy and Edward S. Epstein. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology*, 6(5):748–755, October 1967.

Anish Muthali, Haotian Shen, Sampada Deglurkar, Michael H. Lim, Rebecca Roelofs, Aleksandra Faust, and Claire Tomlin. Multi-agent reachability calibration with conformal prediction. In *2023 62nd IEEE Conference on Decision and Control*, pp. 6596–6603, 2023.

Mahdi P. Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, February 2015.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*. MIT Press, 1999.

Teodora Popordanoska, Raphael Sayer, and Matthew B. Blaschko. A consistent and differentiable lp canonical calibration error estimator. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7933–7946. Curran Associates, Inc., 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, July 2021.

Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer texts in statistics. Springer, New York, NY, November 2010.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. In Gustau Camps-Valls, Francisco J R Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4036–4054. PMLR, March 2022.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, March 1978.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pp. 3437–3452. PMLR, July 2020.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3459–3467. PMLR, April 2019.

Vladimir N. Vapnik and Aleksei Y. Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Doklady Akademii Nauk*, 181(4):781–783, 1968.

Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Martin J. Wainwright. *High Dimensional Statistics Non Asymptotic Viewpoint*. Cambridge University Press, 2019.

Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46(2):265–280, 1945.

Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 178–190. PMLR, August 2020.

D S Wilks. *Statistical methods in the atmospheric sciences*. Elsevier, 2019.

Wang Xinrui, Wenhai Wan, Chuanxing Geng, Shao-Yuan Li, and Songcan Chen. Beyond myopia: Learning from positive and unlabeled data through holistic predictive trends. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 67589–67602. Curran Associates, Inc., 2023.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Gang Niu, Masashi Sugiyama, and Dacheng Tao. Rethinking class-prior estimation for positive-unlabeled learning. In *International Conference on Learning Representations*, May 2021.

Xichen Ye, Xiaoqiang Li, songmin dai, Tong Liu, Yan Sun, and Weiqin Tong. Active negative loss functions for learning with noisy labels. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 6917–6940. Curran Associates, Inc., 2023.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 609–616. Morgan Kaufmann Publishers Inc., June 2001.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-match : Ensemble and compositional methods for uncertainty calibration in deep learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11117–11128. PMLR, July 2020.

Yunrui Zhao, Qianqian Xu, Yangbangyan Jiang, Peisong Wen, and Qingming Huang. Dist-PU: Positive-unlabeled learning from a label distribution perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14461–14470, June 2022.

Yilun Zhu, Aaron Fjeldsted, Darren Holland, George Landon, Azaree Lintereur, and Clayton Scott. Mixture proportion estimation beyond irreducibility. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42962–42982. PMLR, 2023.

A Proofs for Section 3.2

A.1 Proof of Theorem 1

First, we define three types of bias, statistical bias, total bias, and binning bias, as follows:

$$\text{Bias}_{\text{tot}}(f, S_{\text{PU}}) := |\text{TCE}(f) - \text{ECE}_{\text{PU}}(f, S_{\text{PU}})|$$
$$\text{Bias}_{\text{bin}}(f, f_{\mathcal{I}}) := |\text{TCE}(f) - \text{TCE}(f_{\mathcal{I}})|$$
$$\text{Bias}_{\text{stat}}(f, S_{\text{PU}}) := |\text{TCE}(f_{\mathcal{I}}) - \text{ECE}_{\text{PU}}(f, S_{\text{PU}})|.$$

The biases satisfy the following relationship:

$$\text{Bias}_{\text{tot}}(f, S_{\text{PU}}) \leq \text{Bias}_{\text{bin}}(f, f_{\mathcal{I}}) + \text{Bias}_{\text{stat}}(f, S_{\text{PU}}). \tag{5}$$

By analyzing the binning bias and the statistical bias, we can derive a bound for the total bias, which measures how PU-ECE is biased from the TCE. Theorem 3 of Futami & Fujisawa (2024) showed the binning bias bound. We can straightforwardly apply it to the PU setting, taking into account that we use unlabeled data for deciding the bin boundaries of UMB.

Lemma 1 (Binning bias analysis (Futami & Fujisawa, 2024)). *Under Assumption 1, 2 and 3, we have*

$$\text{Bias}_{\text{bin}}(f, f_{\mathcal{I}}) \leq \begin{cases} \frac{1+L}{B} & \text{(for UWB),} \\ (1+L)\left(\frac{1}{B} + \sqrt{\frac{2B\log 2}{n_{\text{U}}-B}} + \frac{2B}{n_{\text{U}}-B}\right) & \text{(for UMB).} \end{cases}$$

Note that Lemma 1 holds without considering either expectation or probability, as it is a deterministic bound. Next, we present the following statistical bias analysis.

Lemma 2 (Statistical bias analysis). *Under Assumption 1 and 2, we have*

$$\text{TCE}(f_{\mathcal{I}}) \leq \mathbb{E}_{S_{\text{PU}}}\left[\text{ECE}_{\text{PU}}(f, S_{\text{PU}})\right],$$

$$\mathbb{E}_{S_{\text{PU}}}\left[\text{Bias}_{\text{stat}}(f, S_{\text{PU}})\right] \leq \begin{cases} \sqrt{2\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)B\log 2} & \text{(for UWB),} \\ \sqrt{2\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}-B}\right)B\log 2} + \frac{B}{n_{\text{U}}-B} & \text{(for UMB).} \end{cases}$$

For any $\delta \in (0,1)$, with probability at least $1 - \delta$ over S_{PU}, we have

$$\text{Bias}_{\text{stat}}(f, S_{\text{PU}}) \leq \begin{cases} \sqrt{2\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)(B\log 2 + \log(1/\delta))} & \text{(for UWB),} \\ \sqrt{2\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}-B}\right)(B\log 2 + \log(1/\delta))} + \frac{B}{n_{\text{U}}-B} & \text{(for UMB).} \end{cases}$$

It is worth noting that for the L_1-ECE, the proof can handle all B bins simultaneously, which leads to a tighter high-probability bound than a naive union bound over the bins. An approach based on a union bound would typically introduce a $B \log B$ term, whereas our method results in a dependency of B, which is an improvement for a large B.

Proof. The proof goes similarly to the proof of Theorem 2 in Futami & Fujisawa (2024). We first prove the inequality $\mathrm{TCE}(f_{\mathcal{I}}) \leq \mathbb{E}_{S_{\mathrm{PU}}}\left[\mathrm{ECE}_{\mathrm{PU}}(f, S_{\mathrm{PU}})\right]$.

$$
\begin{aligned}
\mathrm{TCE}(f_{\mathcal{I}}) &= \sum_{b=1}^{B} P(f(X) \in I_b) \left| \mathbb{E}\left[Y \mid f(X) \in I_b\right] - \mathbb{E}\left[f(X) \mid f(X) \in I_b\right] \right| \\
&= \sum_{b=1}^{B} P(f(X) \in I_b) \left| \frac{\pi_{\mathrm{P}} P(f(X) \in I_b \mid Y = 1)}{P(f(X) \in I_b)} - \frac{\mathbb{E}\left[f(X) \cdot \mathbb{1}_{f(X) \in I_b}\right]}{P(f(X) \in I_b)} \right| \\
&= \sum_{b=1}^{B} \left| \pi_{\mathrm{P}} P(f(X) \in I_b \mid Y = 1) - \mathbb{E}\left[f(X) \mathbb{1}_{f(X) \in I_b}\right] \right| \\
&= \sum_{b=1}^{B} \left| \mathbb{E}_{S_{\mathrm{P}}}\left[\frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{m=1}^{n_{\mathrm{P}}} \mathbb{1}_{f(X_m^{\mathrm{P}}) \in I_b} \right] - \mathbb{E}_{S_{\mathrm{U}}}\left[\frac{1}{n_{\mathrm{U}}} \sum_{m=1}^{n_{\mathrm{P}}} f(X_m^{\mathrm{U}}) \mathbb{1}_{f(X_m^{\mathrm{U}}) \in I_b} \right] \right| \qquad (6) \\
&\leq \mathbb{E}_{S_{\mathrm{PU}}}\left[\mathrm{ECE}_{\mathrm{PU}}(f, S_{\mathrm{PU}})\right].
\end{aligned}
$$

The last inequality is obtained by Jensen's inequality. Next, we prove the bound for $\mathbb{E}_{S_{\mathrm{PU}}}\left[\mathrm{Bias}_{\mathrm{stat}}(f, S_{\mathrm{PU}})\right]$.

UWB: For convenience, denote by $l_{\mathrm{P}}(t, b) := \mathbb{1}_{t \in I_b}$ and $l_{\mathrm{U}}(t, b) := t \mathbb{1}_{t \in I_b}$. We also define $\mathcal{L}(S_{\mathrm{PU}}, b) := \frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{m=1}^{n_{\mathrm{P}}} l_{\mathrm{P}}(f(X_m^{\mathrm{P}}), b) - \frac{1}{n_{\mathrm{U}}} \sum_{m=1}^{n_{\mathrm{U}}} l_{\mathrm{U}}(f(X_m^{\mathrm{U}}), b)$. Note that $0 \leq l_{\mathrm{P}}(t, b) \leq 1$, $0 \leq l_{\mathrm{U}}(t, b) \leq t$, and $\mathrm{TCE}(f_{\mathcal{I}}) = \mathbb{E}_{S_{\mathrm{PU}}}\left[\mathcal{L}(S_{\mathrm{PU}}, b)\right]$. By the triangle inequality, we have

$$
\begin{aligned}
\left| \mathrm{TCE}(f_{\mathcal{I}}) - \mathrm{ECE}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) \right| &= \left| \sum_{b=1}^{B} \left| \mathbb{E}_{S_{\mathrm{PU}}}\left[\mathcal{L}(S_{\mathrm{PU}}, b)\right] \right| - \sum_{b=1}^{B} \left| \mathcal{L}(S_{\mathrm{PU}}, b) \right| \right| \\
&\leq \sum_{b=1}^{B} \left| \mathbb{E}_{S_{\mathrm{PU}}}\left[\mathcal{L}(S_{\mathrm{PU}}, b)\right] - \mathcal{L}(S_{\mathrm{PU}}, b) \right|. \qquad (7)
\end{aligned}
$$

We use the following inequality derived from the Jensen's inequality:

$$
\mathbb{E}\left[|X|\right] \leq \frac{1}{t} \log \mathbb{E}\left[e^{t|X|}\right] \quad (t > 0),
$$

and let $g(S_{\mathrm{PU}}, b) = \mathcal{L}(S_{\mathrm{PU}}, b) - \mathbb{E}_{S_{\mathrm{PU}}}\left[\mathcal{L}(S_{\mathrm{PU}}, b)\right]$. Then we have

$$
\begin{aligned}
\mathbb{E}_{S_{\mathrm{PU}}}\left[\left| \mathrm{TCE}(f_{\mathcal{I}}) - \mathrm{ECE}(f, S_{\mathrm{PU}}) \right|\right] &\leq \frac{1}{t} \log \mathbb{E}_{S_{\mathrm{PU}}}\left[e^{t \sum_{b=1}^{B} |g(S_{\mathrm{PU}}, b)|} \right] \\
&= \frac{1}{t} \log \mathbb{E}_{S_{\mathrm{PU}}}\left[\prod_{b=1}^{B} e^{t|g(S_{\mathrm{PU}}, b)|} \right] \\
&\leq \frac{1}{t} \log \mathbb{E}_{S_{\mathrm{PU}}}\left[\prod_{b=1}^{B} \left\{ e^{tg(S_{\mathrm{PU}}, b)} + e^{-tg(S_{\mathrm{PU}}, b)} \right\} \right] \\
&\leq \frac{1}{t} \log \mathbb{E}_{S_{\mathrm{PU}}}\left[\sum_{v_1, \ldots, v_B = 0, 1} e^{t \sum_{b=1}^{B} (-1)^{v_b} g(S_{\mathrm{PU}}, b)} \right], \qquad (8)
\end{aligned}
$$

where the summation $\sum_{v_1, \ldots, v_B = 0, 1}$ ranges over all 2^B possible binary vectors $v = (v_1, \ldots, v_B) \in \{0, 1\}^B$. Let $Z^{v_b} = (Z_1^v, \ldots, Z_{n_{\mathrm{P}} + n_{\mathrm{U}}}^v)$ be a random variable vector defined as follows:

$$
\begin{aligned}
Z_m^v &:= \sum_{b=1}^{B} (-1)^{v_b} \frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}} l_{\mathrm{P}}(f(X_m^{\mathrm{P}}), b) && (\text{for } m = 1, \ldots, n_{\mathrm{P}}) \\
Z_m^v &:= - \sum_{b=1}^{B} (-1)^{v_b} \frac{1}{n_{\mathrm{U}}} l_{\mathrm{U}}(f(X_m^{\mathrm{U}}), b) && (\text{for } m = n_{\mathrm{P}} + 1, \ldots, n_{\mathrm{P}} + n_{\mathrm{U}}).
\end{aligned}
$$

18

Then we can rewrite the expectation in Eq. (8) as:

$$\mathbb{E}_{S_{\text{PU}}}\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\sum_{b=1}^{B}(-1)^{v_b}g(S_{\text{PU}},b)}\right] = \mathbb{E}_{S_{\text{PU}}}\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\left(\sum_{m=1}^{n_{\text{P}}+n_{\text{U}}} Z_m^v - \mathbb{E}_{Z^v}\left[\sum_{m=1}^{n_{\text{P}}+n_{\text{U}}} Z_m^v\right]\right)}\right] \qquad (9)$$

We calculate the bound of Z_m^v so that we can apply the Hoeffding's lemma to the above inequality.

Lemma 3 (Hoeffding's lemma). *Let $Z_1,\ldots,Z_n$ be independent random variables such that $Z_i \in [a_i,b_i]$ for all i. Then*

$$\mathbb{E}_{Z_1,\ldots,Z_n}\left[\exp\left(t\left\{\sum_{i=1}^{n}Z_i - \mathbb{E}_{Z_1,\ldots,Z_n}\left[\sum_{i=1}^{n}Z_i\right]\right\}\right)\right] \le \prod_{i=1}^{n}\exp\left(\frac{t^2(b_i-a_i)^2}{8}\right)$$

for any $t > 0$.

The interval length of Z_m^v $(m = 1,\ldots,n_{\text{P}})$ is $\frac{2\pi_{\text{P}}}{n_{\text{P}}}$ since the range of Z_m^v (for $m = 1,\ldots,n_{\text{P}}$) is

$$-\frac{\pi_{\text{P}}}{n_{\text{P}}} \le Z_m^v \le \frac{\pi_{\text{P}}}{n_{\text{P}}}.$$

This follows from the fact that $f(X_m^{\text{P}})$ falls into exactly one bin, which means $l_{\text{P}}(f(X_m^{\text{P}}),b) = 1$ for a single b and is zero for all others. Similarly, the interval length of $Z_{m+n_{\text{P}}}^v$ $(m = 1,\ldots,n_{\text{U}})$ is $\frac{2}{n_{\text{U}}}$. Applying Hoeffding's lemma yields:

$$\mathbb{E}_{S_{\text{PU}}}\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\sum_{b=1}^{B}(-1)^{v_b}g(S_{\text{PU}},b)}\right] \le \sum_{v_1,\ldots,v_B=0,1} e^{\frac{t^2}{8}\left(n_{\text{P}}\frac{4\pi_{\text{P}}^2}{n_{\text{P}}^2}+n_{\text{U}}\frac{4}{n_{\text{U}}^2}\right)}$$

$$= 2^B e^{\frac{t^2}{2}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}}+\frac{1}{n_{\text{U}}}\right)}. \qquad (10)$$

Therefore, from Eq. (9),

$$\mathbb{E}_{S_{\text{PU}}}\left[|\text{TCE}(f_{\mathcal{I}}) - \text{ECE}(f,S_{\text{PU}})|\right] \le \frac{1}{t}\log\left(2^B e^{\frac{t^2}{2}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}}+\frac{1}{n_{\text{U}}}\right)}\right)$$

$$= \frac{B\log 2}{t} + \frac{t}{2}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}}+\frac{1}{n_{\text{U}}}\right)$$

$$\le \sqrt{2\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}}+\frac{1}{n_{\text{U}}}\right)B\log 2}.$$

UMB: The proof is similar to the UWB case, but we need to consider the dependence of the bin boundaries on the data. Let $S_{(B)} \subset S_{\text{U}}$ be the set of $B-1$ unlabeled samples used to determine the bin boundaries, and let $S_{\text{U-B}} = S_{\text{U}} \setminus S_{(B)}$ be the remaining unlabeled samples and let $S_{\text{PU-B}} := (S_{\text{U-B}}, S_{\text{P}})$. We have

$$|\text{TCE}(f_{\mathcal{I}}) - \text{ECE}_{\text{PU}}(f,S_{\text{PU}})|$$

$$\le \sum_{b=1}^{B}|\mathbb{E}_{S_{\text{PU}}}\left[\mathcal{L}(S_{\text{PU}},b)\right] - \mathcal{L}(S_{\text{PU}},b)|$$

$$\le \sum_{b=1}^{B}|\mathbb{E}_{S_{\text{PU-B}}}\left[\mathcal{L}(b,S_{\text{PU-B}})\right] - \mathcal{L}(b,S_{\text{PU-B}})|$$

$$+ \sum_{b=1}^{B}\left|\mathbb{E}_{S_{\text{U}}}\left[\frac{1}{|S_{\text{U-B}}|}\sum_{X_m^{\text{U}}\in S_{\text{U-B}}} l_{\text{U}}(f(X_m^{\text{U}}),b) - \frac{1}{n_{\text{U}}}\sum_{m=1}^{n_{\text{U}}} l_{\text{U}}(f(X_m^{\text{U}}),b)\right]\right|. \qquad (11)$$

The second term of Eq. (11) includes the bin boundary samples that are not i.i.d.; therefore, we have

$$
\mathbb{E}_{S_{\mathrm{U}}} \left[\frac{1}{|S_{\mathrm{U\text{-}B}}|} \sum_{X_m^{\mathrm{U}} \in S_{\mathrm{U\text{-}B}}} l_{\mathrm{U}}(f(X_m^{\mathrm{U}}), b) - \frac{1}{n_{\mathrm{U}}} \sum_{m=1}^{n_{\mathrm{U}}} l_{\mathrm{U}}(f(X_m^{\mathrm{U}}), b) \right]
$$

$$
= \left| \left(\frac{1}{n_{\mathrm{U}} - B + 1} - \frac{1}{n_{\mathrm{U}}} \right) \sum_{X_m^{\mathrm{U}} \in S_{\mathrm{U\text{-}B}}} l_{\mathrm{U}}(f(X_m^{\mathrm{U}}), b) - \frac{1}{n_{\mathrm{U}}} \sum_{X_m^{\mathrm{U}} \in S_{(B)}} l_{\mathrm{U}}(f(X_m^{\mathrm{U}}), b) \right|.
$$

Here, we use the following inequality:

$$
\frac{1}{n_{\mathrm{U}} - B + 1} - \frac{1}{n_{\mathrm{U}}} = \frac{n_{\mathrm{U}} - n_{\mathrm{U}} + B - 1}{(n_{\mathrm{U}} - B + 1)n_{\mathrm{U}}}
$$

$$
= \frac{B - 1}{(n_{\mathrm{U}} - B + 1)n_{\mathrm{U}}}
$$

$$
\leq \frac{B - 1}{n_{\mathrm{U}} - B + 1},
$$

which leads to

$$
\mathbb{E}_{S_{\mathrm{U}}} \left[\frac{1}{|S_{\mathrm{U\text{-}B}}|} \sum_{X_m^{\mathrm{U}} \in S_{\mathrm{U\text{-}B}}} l_{\mathrm{U}}(f(X_m^{\mathrm{U}}), b) - \frac{1}{n_{\mathrm{U}}} \sum_{m=1}^{n_{\mathrm{U}}} l_{\mathrm{U}}(f(X_m^{\mathrm{U}}), b) \right]
$$

$$
\leq \max \left\{ \frac{B - 1}{n_{\mathrm{U}} - B + 1}, \frac{B - 1}{n_{\mathrm{U}}} \right\}
$$

$$
\leq \frac{B}{n_{\mathrm{U}} - B}, \tag{12}
$$

where the first inequality is from Eq. (7) and the second is obtained by the triangle inequality. For the first term of Eq. (11), we need to consider the dependence of the bin boundaries on the data, but we can still apply Hoeffding's lemma to it in the same manner as the UWB case thanks to Lemma 2 in Gupta & Ramdas (2021); data points inside a bin are i.i.d. Therefore, we have

$$
\mathbb{E}_{S_{\mathrm{PU}}} \left[|\mathrm{TCE}(f_{\mathcal{I}}) - \mathrm{ECE}(f, S_{\mathrm{PU}})| \right] \leq \frac{1}{t} \log \mathbb{E}_{S_{\mathrm{PU}}} \left[e^{t \sum_{b=1}^{B} |g(b, S_{\mathrm{PU\text{-}B}})| + \frac{tB}{n_{\mathrm{U}} - B}} \right]
$$

$$
\leq \frac{1}{t} \log \left\{ 2^B e^{\frac{t^2}{2} \left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{1}{n_{\mathrm{U}} - B} \right) + \frac{tB}{n_{\mathrm{U}} - B}} \right\}
$$

$$
= \frac{1}{t} \left\{ B \log 2 + \frac{t^2}{2} \left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{1}{n_{\mathrm{U}} - B} \right) + \frac{tB}{n_{\mathrm{U}} - B} \right\}
$$

$$
= \frac{B \log 2}{t} + \frac{t}{2} \left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{1}{n_{\mathrm{U}} - B} \right) + \frac{B}{n_{\mathrm{U}} - B}. \tag{13}
$$

By minimizing the above expression with respect to t, we have

$$
\mathbb{E}_{R, \tilde{X}, M} \left[\sum_{b=1}^{B} \Delta_{\hat{P}}(W, \tilde{X}, M, b) \right] \leq \inf_{t > 0} \left\{ \frac{I(f_W(\tilde{X}); M | \tilde{X}) + \mathbb{E}_{\tilde{X}} \left[\log \mathbb{E}_{R, M, M'} \left[e^{t \Delta_{\hat{P}}(W, \tilde{X}, M', b)} \right] \right]}{t} \right\}
$$

$$
\leq \inf_{t > 0} \left\{ \frac{\mathrm{fCMI}_{\mathrm{U}} + B \log 2 + \frac{2t^2}{2 n_{\mathrm{tr,U}}}}{t} \right\}
$$

$$
\leq \sqrt{\frac{8}{n_{\mathrm{tr,U}}} \{ \mathrm{fCMI}_{\mathrm{U}} + B \log 2 \}},
$$

Finally, we prove the high probability bound. By the Chernoff bound, we have

$$
P(\mathrm{Bias}_{\mathrm{stat}}(f, S_{\mathrm{PU}}) > \epsilon) \leq \inf_{t > 0} \mathbb{E}_{S_{\mathrm{PU}}} \left[e^{t \, \mathrm{Bias}_{\mathrm{stat}}(f, S_{\mathrm{PU}})} \right] e^{-t\epsilon}.
$$

UWB: By using the Eq (7), (8) and (10), we have

$$\inf_{t>0} \mathbb{E}_{S_{\mathrm{PU}}}\left[e^{t\,\mathrm{Bias}_{\mathrm{stat}}(f,S_{\mathrm{PU}})}\right]e^{-t\epsilon} \leq \inf_{t>0} \mathbb{E}_{S_{\mathrm{PU}}}\left[e^{t\sum_{b=1}^{B}|g(S_{\mathrm{PU}},b)|}\right]e^{-t\epsilon}$$

$$\leq \inf_{t>0} \mathbb{E}_{S_{\mathrm{PU}}}\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\sum_{b=1}^{B}(-1)^{v_b}g(S_{\mathrm{PU}},b)}\right]e^{-t\epsilon}$$

$$\leq \inf_{t>0} 2^B e^{\frac{t^2}{2}\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}}\right)}e^{-t\epsilon}$$

$$= e^{-\frac{\epsilon^2}{2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}}\right)}+B\log 2}.$$

Let $\delta := e^{-\frac{\epsilon^2}{2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}}\right)}+B\log 2}$, then we have

$$\epsilon = \sqrt{2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}}\right)(B\log 2 + \log(1/\delta))}.$$

Substituting this into the above inequality, we have

$$P\left(\mathrm{Bias}_{\mathrm{stat}}(f,S_{\mathrm{PU}}) > \sqrt{2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}}\right)(B\log 2 + \log(1/\delta))}\right) \leq \delta.$$

UMB: The proof is similar to the UWB case, but we need to consider the dependence of the bin boundaries on the data similarly to the proof of the expected statistical bias. By using the Eq (11), (12) and (13), we have

$$\inf_{t>0} \mathbb{E}_{S_{\mathrm{PU}}}\left[e^{t\,\mathrm{Bias}_{\mathrm{stat}}(f,S_{\mathrm{PU}})}\right]e^{-t\epsilon} \leq \inf_{t>0} 2^B e^{\frac{t^2}{2}\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}-B}\right)+\frac{tB}{n_{\mathrm{U}}-B}}e^{-t\epsilon}$$

$$= e^{-\frac{1}{2}\left(\epsilon-\frac{B}{n_{\mathrm{U}}-B}\right)^2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}-B}\right)^{-1}+B\log 2}.$$

Let $\delta := e^{-\frac{1}{2}\left(\epsilon-\frac{B}{n_{\mathrm{U}}-B}\right)^2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}-B}\right)^{-1}+B\log 2}$, then we have

$$\epsilon = \frac{B}{n_{\mathrm{U}}-B} + \sqrt{2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}-B}\right)(B\log 2 + \log(1/\delta))}.$$

Substituting this into the above inequality, we have

$$P\left(\mathrm{Bias}_{\mathrm{stat}}(f,S_{\mathrm{PU}}) > \frac{B}{n_{\mathrm{U}}-B} + \sqrt{2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}}+\frac{1}{n_{\mathrm{U}}-B}\right)(B\log 2 + \log(1/\delta))}\right) \leq \delta.$$

$\square$

Applying Lemma 1 and 2 to Eq. (5), we have the total bias bound of Theorem 1.

A.2 Proof of Corollary 1

Proof. We first consider the UWB case and then the UMB case.

UWB: We regard the total bias as a function of B, $h(B) = \mathbb{E}_{S_{\text{PU}}}[\text{Bias}_{\text{tot}}(f, S_{\text{PU}})]$. The first derivative of $h(B)$ is given by

$$
h'(B) = -\frac{1+L}{B^2} + \frac{1}{2}\sqrt{\frac{2}{B}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)\log 2}
$$

$$
= \frac{-(1+L) + B^{\frac{3}{2}}\sqrt{\frac{\log 2}{2}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)}}{B^2}.
$$

The optimal B is given by the solution of $h'(B) = 0$, i.e.

$$
B = \left(\frac{(1+L)}{\sqrt{\frac{\log 2}{2}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)}}\right)^{\frac{2}{3}}
$$

$$
= \left(\frac{2(1+L)^2}{\log 2}\right)^{\frac{1}{3}}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)^{-\frac{1}{3}}.
$$

The corresponding total bias bound is given by

$$
h(B) = (1+L)\left(\frac{\log 2}{2(1+L)^2}\right)^{\frac{1}{3}}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)^{\frac{1}{3}} + \sqrt{2\left(\frac{2(1+L)^2}{\log 2}\right)^{\frac{1}{3}}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)^{-\frac{1}{3}}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)\log 2}
$$

$$
= 3\left\{\frac{(1+L)\log 2}{2}\right\}^{\frac{1}{3}}\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)^{\frac{1}{3}}.
$$

UMB: We further upper-bound the total bias for simplicity

$$
\mathbb{E}_{S_{\text{PU}}}[\text{Bias}_{\text{tot}}(f, S_{\text{PU}})] \leq \frac{1+L}{B} + (3+2L)\frac{B}{n_{\text{U}} - B} + \sqrt{2B\log 2}\left(\frac{1+L}{\sqrt{n_{\text{U}} - B}} + \sqrt{\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}} - B}}\right).
$$

By the assumption $n_{\text{U}} \geq 2B$, we have $\frac{1}{n_{\text{U}} - B} \leq \frac{2}{n_{\text{U}}}$. Therefore, we have

$$
(1+L)\sqrt{\frac{1}{n_{\text{U}} - B}} + \sqrt{\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}} - B}} \leq (2+L)\sqrt{2}\sqrt{\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}}.
$$

We also have

$$
(3+2L)\frac{B}{n_{\text{U}} - B} \leq 4(2+L)\sqrt{\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)}B.
$$

Therefore, we have

$$
\mathbb{E}_{S_{\text{PU}}}[\text{Bias}_{\text{tot}}(f, S_{\text{PU}})] \leq \frac{2+L}{B} + 4(2+L)\sqrt{\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)B} + 2(2+L)\sqrt{\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)B\log 2}
$$

$$
= (2+L)\left\{\frac{1}{B} + 2(2+\sqrt{\log 2})\sqrt{\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)B}\right\}
$$

Denote the RHS by $h(B)$. The first derivative of $h(B)$ is given by

$$
h'(B) = (2+L)\left\{-\frac{1}{B^2} + (2+\sqrt{\log 2})\sqrt{\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)\frac{1}{B}}\right\}
$$

$$
= \frac{2+L}{B^2}\left\{-1 + (2+\sqrt{\log 2})B^{\frac{3}{2}}\sqrt{\left(\frac{\pi_{\text{P}}^2}{n_{\text{P}}} + \frac{1}{n_{\text{U}}}\right)}\right\}.
$$

The optimal B is given by the solution of $h'(B) = 0$, i.e.

$$B = (2 + \sqrt{\log 2})^{-\frac{2}{3}} \left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{1}{n_{\mathrm{U}}} \right)^{-\frac{1}{3}}.$$

The corresponding total bias bound is given by

$$h(B) = 3(2 + L)(2 + \sqrt{\log 2})^{\frac{2}{3}} \left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{1}{n_{\mathrm{U}}} \right)^{\frac{1}{3}}.$$

$\square$

A.3   Proof of Theorem 2

Proof. The difference is:

$$\left| \mathrm{ECE}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) - \widehat{\mathrm{ECE}}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) \right| = \left| \sum_{b=1}^{B} \left| \frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{m=1}^{n_{\mathrm{P}}} \mathbb{1}_{f(X_m^{\mathrm{P}}) \in I_b} - \frac{1}{n_{\mathrm{U}}} \sum_{m=1}^{n_{\mathrm{U}}} f(X_m^{\mathrm{U}}) \mathbb{1}_{f(X_m^{\mathrm{U}}) \in I_b} \right| \right.$$
$$\left. - \sum_{b=1}^{B} \left| \frac{\hat{\pi}_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{m=1}^{n_{\mathrm{P}}} \mathbb{1}_{f(X_m^{\mathrm{P}}) \in I_b} - \frac{1}{n_{\mathrm{U}}} \sum_{m=1}^{n_{\mathrm{U}}} f(X_m^{\mathrm{U}}) \mathbb{1}_{f(X_m^{\mathrm{U}}) \in I_b} \right| \right|.$$

By the reverse triangle inequality, we have

$$\left| \mathrm{ECE}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) - \widehat{\mathrm{ECE}}_{\mathrm{PU}}(f, S_{\mathrm{PU}}) \right| \leq \sum_{b=1}^{B} \left| \frac{\pi_{\mathrm{P}} - \hat{\pi}_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{m=1}^{n_{\mathrm{P}}} \mathbb{1}_{f(X_m^{\mathrm{P}}) \in I_b} \right|$$
$$= \frac{|\pi_{\mathrm{P}} - \hat{\pi}_{\mathrm{P}}|}{n_{\mathrm{P}}} \sum_{b=1}^{B} \left| \sum_{m=1}^{n_{\mathrm{P}}} \mathbb{1}_{f(X_m^{\mathrm{P}}) \in I_b} \right|$$
$$= \frac{|\pi_{\mathrm{P}} - \hat{\pi}_{\mathrm{P}}|}{n_{\mathrm{P}}} \sum_{b=1}^{B} \sum_{m=1}^{n_{\mathrm{P}}} \mathbb{1}_{f(X_m^{\mathrm{P}}) \in I_b}$$
$$= |\pi_{\mathrm{P}} - \hat{\pi}_{\mathrm{P}}|,$$

where the third equality holds because the term inside the absolute value is non-negative. $\square$

B   Details of CMI for PU Learning in Section 4.2

Denote by $D_{\mathrm{KL}}(P \| Q)$ the Kullback-Leibler divergence of distribution P from distribution Q. We assume that $D_{\mathrm{KL}}(P \| Q) < \infty$ in this paper; i.e., P is absolutely continuous with respect to Q. The CMI is defined as $I(X; Y | Z) := \mathbb{E}_Z \left[D_{\mathrm{KL}}(P(X, Y | Z) \| P(X | Z) P(Y | Z)) \right]$.

In the CMI setting for PU learning, we consider a supersample $\tilde{X} = [\tilde{X}^{\mathrm{U}}; \tilde{X}^{\mathrm{P}}] \in \mathcal{X}^{(n_{\mathrm{tr},\mathrm{U}} + n_{\mathrm{tr},\mathrm{P}}) \times 2}$, which is a $(n_{\mathrm{tr},\mathrm{U}} + n_{\mathrm{tr},\mathrm{P}}) \times 2$ array of samples. [2] This array is composed of two subarrays. The first subarray is the unlabeled supersample $\tilde{X}^{\mathrm{U}} \in \mathcal{X}^{n_{\mathrm{tr},\mathrm{U}} \times 2}$, where each entry is drawn i.i.d. from $P(X)$. The second subarray is the positive supersample $\tilde{X}^{\mathrm{P}} \in \mathcal{X}^{n_{\mathrm{tr},\mathrm{P}} \times 2}$, where each entry is drawn i.i.d. from $P(X \mid Y = 1)$. Note that the standard CMI setting of supervised learning uses input-output data pairs as the supersample (Steinke & Zakynthinou, 2020; Hellström & Durisi, 2022), while in our setting, we use only the input data samples. For convenience, we use the 0-start-index for the columns of the arrays and the 1-start-index for the rows of the arrays. We denote the i-th row of $\tilde{X}^{\mathrm{U}}$ by $\tilde{X}_i^{\mathrm{U}}$ and the i-th row of $\tilde{X}^{\mathrm{P}}$ by $\tilde{X}_i^{\mathrm{P}}$. Analogous to $\tilde{X}$, we collect positive and unlabeled membership indicators into a single supersample membership vector

---

[2]We use the term "array" for notational convenience, although each entry is an element of $\mathcal{X}$, which can be a high-dimensional vector rather than a scalar.

$M := [M^{\mathrm{U}}; M^{\mathrm{P}}] \in \{0,1\}^{n_{\mathrm{tr,U}}+n_{\mathrm{tr,P}}}$, where $M^{\mathrm{U}} \in \{0,1\}^{n_{\mathrm{tr,U}}}$ and $M^{\mathrm{P}} \in \{0,1\}^{n_{\mathrm{tr,P}}}$ are the membership vectors for the unlabeled and positive supersamples, respectively. Using the same index convention as above (i.e., rows are 1-start and columns are 0-start), the entry M_i specifies which column of the i-th row $\tilde{X}_i$ is retained for the i-th data point: $M_i = 0$ selects the left column (index 0) and $M_i = 1$ selects the right column (index 1). Similarly, M_i^{U} and M_i^{P} specify the column selection for the unlabeled and positive samples, respectively. All entries are drawn independently as $M_i \sim \mathrm{Bern}(1/2)$. We denote the training vector by $\tilde{X}_M := [\tilde{X}_{M^{\mathrm{U}}}^{\mathrm{U}}; \tilde{X}_{M^{\mathrm{P}}}^{\mathrm{P}}]$ and the unused (test) vector by $\tilde{X}_{\bar{M}} := [\tilde{X}_{\bar{M}^{\mathrm{U}}}^{\mathrm{U}}; \tilde{X}_{\bar{M}^{\mathrm{P}}}^{\mathrm{P}}]$. Let $\mathcal{A} : \mathcal{X}^{n_{\mathrm{tr,U}}+n_{\mathrm{tr,P}}} \times \mathcal{R} \to \Theta$ be a randomized algorithm that maps the training samples to the parameters of a predictor $f_{\mathcal{A}(\tilde{X}_M,R)} \in \mathcal{F}$, where $R \in \mathcal{R}$ is a randomness of an algorithm, independent of any other random variables.

We define the functional conditional mutual information (fCMI) for PU data as

$$\mathrm{fCMI} := I(f_{\mathcal{A}(\tilde{X}_M,R)}(\tilde{X}); M|\tilde{X}),$$

where $f_{\mathcal{A}(\tilde{X}_M,R)}(\tilde{X})$ is the matrix calculated by the elementwise application of $f_{\mathcal{A}(\tilde{X}_M,R)}$ to $\tilde{X}$. Similarly, we also define fCMI on $\tilde{X}^{\mathrm{U}}$ as

$$\mathrm{fCMI}_{\mathrm{U}} := I(f_{\mathcal{A}(\tilde{X}_M,R)}(\tilde{X}^{\mathrm{U}}); M|\tilde{X}).$$

By the data processing inequality, we have $\mathrm{fCMI}_{\mathrm{U}} \leq \mathrm{fCMI}$.

These PU CMI metrics differ slightly from the original definitions (Steinke & Zakynthinou, 2020; Hellström & Durisi, 2022) because they consist of two data distributions, while the ordinary CMI metrics consist of one data distribution. Due to this dependency on two distributions, existing CMI bounds that depend on both the data and the model cannot be directly applied to the PU CMI metrics. Nevertheless, the PU CMI metrics can be bounded by data-independent model complexity measures. For instance, if the model has a VC dimension d_{VC}, $\mathrm{fCMI} \leq \max\{(d_{\mathrm{VC}}+1)\log 2, d_{\mathrm{VC}}\log(2e(n_{\mathrm{tr,U}}+n_{\mathrm{tr,P}})/d_{\mathrm{VC}})\}$ (Steinke & Zakynthinou, 2020; Harutyunyan et al., 2021). If $f_{\mathcal{A}(\tilde{X}_M,R)}$ is a d-dimensional parametric function that is L-Lipschitz continuous on $\|\cdot\|_\infty$ and whose domain is $[0,1]^d$ and the function class has an ϵ-cover, then $\mathrm{fCMI} \lesssim d\log(L/\epsilon)$ and $\mathrm{fCMI}_{\mathrm{U}} \lesssim d\log(L/\epsilon)$ (Wainwright, 2019; Futami & Fujisawa, 2024).

C Proofs in Section 4

C.1 Proof of Theorem 3

We assume the PU CMI setting; we set $n_{\mathrm{U}} = n_{\mathrm{tr,U}}$ and $n_{\mathrm{tr,P}} = n_{\mathrm{tr,P}}$, and consider $S_{\mathrm{tr}} = \tilde{X}_M$ and $S_{\mathrm{PU}} = \tilde{X}_{\bar{M}}$, which does not lose generality in the analysis.

The generalization error of the PU-ECE can be decomposed similarly to the Eq. (5)

$$\mathrm{Bias}_{\mathrm{tot}}(f_W, S_{\mathrm{tr}}) \leq \mathrm{Bias}_{\mathrm{bin}}(f_W, f_{\mathcal{I}}) + \mathrm{Bias}_{\mathrm{stat}}(f_W, S_{\mathrm{tr}})$$

where

$$\mathrm{Bias}_{\mathrm{bin}}(f_W, f_{\mathcal{I}}) = |\mathrm{TCE}(f_W) - \mathrm{TCE}(f_{\mathcal{I}})|$$
$$\mathrm{Bias}_{\mathrm{stat}}(f_W, S_{\mathrm{tr}}) = |\mathrm{TCE}(f_{\mathcal{I}}) - \mathrm{ECE}_{\mathrm{PU}}(f_W, S_{\mathrm{tr}})|.$$

First, we derive the bounds for $\mathrm{Bias}_{\mathrm{stat}}(f_W, S_{\mathrm{tr}})$. Instead of directly analyzing it, we analyze an upper bound of it, which is easier to analyze.

Lemma 4. *The statistical bias is upper-bounded as follows:*

$$\mathrm{Bias}_{\mathrm{stat}}(f_W, S_{\mathrm{tr}}) \leq \mathbb{E}_{S_{\mathrm{PU}}}\left[\Delta_{\mathcal{L}}(W, \tilde{X}, M)\right]$$

where

$$\Delta_{\mathcal{L}}(W, \tilde{X}, M) = \sum_{b=1}^{B} \left|\mathcal{L}(W, \tilde{X}_M, b) - \mathcal{L}(W, \tilde{X}_{\bar{M}}, b)\right|$$

$$\mathcal{L}(W, \tilde{X}_m, b) = \frac{\pi_{\mathrm{P}}}{n_{\mathrm{tr,P}}} \sum_{i=1}^{n_{\mathrm{tr,P}}} \mathbb{1}_{f_W(\tilde{X}_{i,m_i^{\mathrm{P}}}^{\mathrm{P}}) \in I_b} - \frac{1}{n_{\mathrm{tr,U}}} \sum_{m=1}^{n_{\mathrm{tr,U}}} f_W(\tilde{X}_{i,m_i^{\mathrm{U}}}^{\mathrm{U}}) \mathbb{1}_{f_W(\tilde{X}_{i,m_i^{\mathrm{U}}}^{\mathrm{U}}) \in I_b}.$$

The expectation of the statistical bias is upper-bounded as well:

$$\mathbb{E}_{R,S_{\mathrm{tr}}}\left[\mathrm{Bias}_{\mathrm{stat}}(f_W, S_{\mathrm{tr}})\right] \leq \mathbb{E}_{R,\tilde{X},M}\left[\Delta_{\mathcal{L}}(\mathcal{A}(\tilde{X}_M, R), \tilde{X}, M)\right].$$

Proof. We have

$$\mathrm{Bias}_{\mathrm{stat}}(f_W, S_{\mathrm{tr}}) = |\mathrm{TCE}(f_{\mathcal{I}}) - \mathrm{ECE}_{\mathrm{PU}}(f_W, S_{\mathrm{tr}})|$$

$$= \left|\sum_{b=1}^{B} |\mathbb{E}_{S_{\mathrm{PU}}}\left[\mathcal{L}(W, S_{\mathrm{PU}}, b)\right]| - \sum_{b=1}^{B} |\mathcal{L}(W, S_{\mathrm{tr}}, b)|\right|$$

$$\leq \sum_{b=1}^{B} |\mathbb{E}_{S_{\mathrm{PU}}}\left[\mathcal{L}(W, S_{\mathrm{PU}}, b)\right] - \mathcal{L}(W, S_{\mathrm{tr}}, b)|$$

$$\leq \mathbb{E}_{S_{\mathrm{PU}}}\left[\sum_{b=1}^{B} |\mathcal{L}(W, S_{\mathrm{PU}}, b) - \mathcal{L}(W, S_{\mathrm{tr}}, b)|\right]$$

$$= \mathbb{E}_{S_{\mathrm{PU}}}\left[\Delta_{\mathcal{L}}(W, \tilde{X}, M)\right]$$

where the second line is obtained by the Eq((6)), the third line is obtained by the triangle inequality and the fifth line is obtained by Jensen's inequality. By taking the expectation over the training data S_{tr}, we have

$$\mathbb{E}_{R,S_{\mathrm{tr}}}\left[\mathrm{Bias}_{\mathrm{stat}}(f_W, S_{\mathrm{tr}})\right] \leq \mathbb{E}_{R,\tilde{X}_M,\tilde{X}_{\bar{M}}}\left[\sum_{b=1}^{B} |\mathcal{L}(W, \tilde{X}_{\bar{M}}, b) - \mathcal{L}(W, \tilde{X}_M, b)|\right]$$

$$= \mathbb{E}_{R,\tilde{X},M}\left[\Delta_{\mathcal{L}}(W, \tilde{X}, M)\right].$$

$\square$

Using the above lemma, we can derive the statistical bias of the PU-ECE on the training data.

Lemma 5 (Statistical bias analysis on training data). *Under Assumption 2, we have*

$$\mathbb{E}_{S_{\mathrm{tr}}}\left[\mathrm{Bias}_{\mathrm{stat}}(f_W, S_{\mathrm{tr}})\right] \leq \sqrt{8\left\{\mathrm{fCMI} + B\log 2\right\}\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)}.$$

Proof. By Lemma 4 and using the Donsker-Varadhan lemma (Donsker & Varadhan, 1975; Steinke & Zakynthinou, 2020), we have

$$\mathbb{E}_{R,\tilde{X},M}\left[\Delta_{\mathcal{L}}(W, \tilde{X}, M)\right] \leq \inf_{t>0}\left\{\frac{I(f_W(\tilde{X}); M|\tilde{X}) + \mathbb{E}_{\tilde{X}}\left[\log \mathbb{E}_{R,M,M'}\left[e^{t\Delta_{\mathcal{L}}(W, \tilde{X}, M')}\right]\right]}{t}\right\}$$

$$\leq \inf_{t>0}\left\{\frac{I(f_W(\tilde{X}); M|\tilde{X}) + \mathbb{E}_{\tilde{X}}\left[\log \mathbb{E}_{R,M,M'}\left[e^{t\sum_{b=1}^{B}|g(\tilde{X},M,R,M',b)|}\right]\right]}{t}\right\},$$

where

$$g(\tilde{X}, M, R, M', b) = \mathcal{L}(\mathcal{A}(\tilde{X}_M, R), \tilde{X}_{\bar{M}'}, b) - \mathcal{L}(\mathcal{A}(\tilde{X}_M, R), \tilde{X}_{M'}, b),$$

and $M' = \left(\left(M^{\mathrm{U}}\right)', \left(M^{\mathrm{P}}\right)'\right)$ is an independent copy of M, i.e., $\left(M^{\mathrm{U}}\right)'_i \sim \mathrm{Bern}(1/2)$ and $\left(M^{\mathrm{P}}\right)'_i \sim \mathrm{Bern}(1/2)$ independently. Similarly to the proof of Lemma 2, we upper-bound the second term.

$$\mathbb{E}_{M'}\left[e^{t\sum_{b=1}^{B}|g(\tilde{x},m,r,M',b)|}\right] \leq \mathbb{E}_{M'}\left[\prod_{b=1}^{B}\left(e^{tg(\tilde{x},m,r,M',b)} + e^{-tg(\tilde{x},m,r,M',b)}\right)\right]$$

$$= \mathbb{E}_{M'}\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\sum_{b=1}^{B}(-1)^{v_b}g(\tilde{x},m,r,M',b)}\right] \qquad (14)$$

We calculate the range of the exponent of the above expression w.r.t. $\left(M^{\mathrm{U}}\right)'$ and $\left(M^{\mathrm{P}}\right)'$ in order to apply Hoeffding's lemma. We define $l_{\mathrm{P}}(w,x,b) := \mathbb{1}_{f_w(x)\in I_b}$ and $l_{\mathrm{U}}(w,x,b) := f_w(x)\mathbb{1}_{f_w(x)\in I_b}$. Let $Z^v = (Z_1^v,\ldots,Z_{n_{\mathrm{tr,P}}+n_{\mathrm{tr,U}}}^v)$ be a random variable vector defined as follows:

$$Z_i^v = \sum_{b=1}^{B}(-1)^{v_b}\frac{\pi_{\mathrm{tr,P}}}{n_{\mathrm{tr,P}}}\left\{l_{\mathrm{P}}(w,\tilde{x}_{i,(\bar{M}^{\mathrm{P}})_i'}^{\mathrm{P}},b) - l_{\mathrm{P}}(w,\tilde{x}_{i,(M^{\mathrm{P}})_i'}^{\mathrm{P}},b)\right\} \qquad (\text{for } i=1,\ldots,n_{\mathrm{tr,P}})$$

$$Z_i^v = \sum_{b=1}^{B}(-1)^{v_b}\frac{1}{n_{\mathrm{tr,U}}}\left\{l_{\mathrm{U}}(w,\tilde{x}_{i,(\bar{M}^{\mathrm{U}})_i'}^{\mathrm{U}},b) - l_{\mathrm{U}}(w,\tilde{x}_{i,(M^{\mathrm{U}})_i'}^{\mathrm{U}},b)\right\} \qquad (\text{for } i=n_{\mathrm{tr,P}}+1,\ldots,n_{\mathrm{tr,P}}+n_{\mathrm{tr,U}}).$$

Then we can write the Eq. (14) as:

$$\mathbb{E}_{M'}\left[\sum_{v_1,\ldots,v_B=0,1}e^{t\sum_{b=1}^{B}(-1)^{v_b}g(\tilde{x},m,r,M',b)}\right] = \mathbb{E}_{Z^v}\left[\sum_{v_1,\ldots,v_B=0,1}e^{t\sum_{i=1}^{n_{\mathrm{tr,P}}+n_{\mathrm{tr,U}}}Z_i^v}\right].$$

Note that $\mathbb{E}\left[Z_i^v\right]=0$ for all i. The interval length of Z_i^v for $i=1,\ldots,n_{\mathrm{tr,P}}$ is $\frac{4\pi_{\mathrm{tr,P}}}{n_{\mathrm{tr,P}}}$ since the range of each is

$$-\frac{2\pi_{\mathrm{tr,P}}}{n_{\mathrm{tr,P}}} \le Z_i^v \le \frac{2\pi_{\mathrm{tr,P}}}{n_{\mathrm{tr,P}}} \qquad (i=1,\ldots,n_{\mathrm{tr,P}}).$$

Similarly, the interval length of Z_i^v for $i=n_{\mathrm{tr,P}}+1,\ldots,n_{\mathrm{tr,P}}+n_{\mathrm{tr,U}}$ is $\frac{2}{n_{\mathrm{tr,U}}}$. Unlike the proof of Lemma 2, since we consider independent membership vectors instead of the training samples themselves, we do not need to account for the dependence of UMB on training samples. Therefore, by Hoeffding's lemma, we simply have

$$\mathbb{E}_Z\left[\sum_{v_1,\ldots,v_B=0,1}e^{t\sum_{i=1}^{n_{\mathrm{tr,P}}+n_{\mathrm{tr,U}}}Z_i^v}\right] \le \sum_{v_1,\ldots,v_B=0,1}e^{\frac{t^2}{8}\left(n_{\mathrm{tr,P}}\frac{16\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}^2}+n_{\mathrm{tr,U}}\frac{16}{n_{\mathrm{tr,U}}^2}\right)}$$

$$= 2^B e^{2t^2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}}+\frac{1}{n_{\mathrm{tr,U}}}\right)}$$

By substituting the above result into the previous equation, we have

$$\mathbb{E}_{R,\tilde{X},M}\left[\Delta_{\mathcal{L}}(W,\tilde{X},M)\right] \le \inf_{t>0}\frac{I(f_W(\tilde{X});M|\tilde{X})+\log 2^B e^{2t^2\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}}+\frac{1}{n_{\mathrm{tr,U}}}\right)}}{t}$$

$$\le \sqrt{8\{\mathrm{fCMI}+B\log 2\}\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}}+\frac{1}{n_{\mathrm{tr,U}}}\right)}$$

$\square$

Note about Theorem 4 in Futami & Fujisawa (2024) Theorem 4 in Futami & Fujisawa (2024) states that the statistical bias of the ECE is upper-bounded by the expected absolute difference of the ECE on evaluation data and the ECE on training data. However, this is not true in general. The second last line and the last line of the Eq.(98) should be not equal, but the last line should be $\le$. This leads to the wrong conclusion that the statistical bias of the ECE is bounded by eCMI of the absolute difference of the two ECEs. Our proof shows that the statistical bias of the ECE is bounded by the expected absolute difference of accuracy.

Furthermore, their "eCMI($\tilde{l}$)" is actually not the eCMI of the original definition, since their $\tilde{l}$ depends on the membership vector U directly.

We can also analyze the binning bias of PU-ECE on the training data. We can directly apply the binning bias analysis in Futami & Fujisawa (2024). According to Theorem 8 in Futami & Fujisawa (2024), we have

Lemma 6 (Binning bias analysis on training data). *Under Assumption 2 and 3, we have*

$$
\mathbb{E}_{R,S_{\mathrm{tr}}}\left[\mathrm{Bias}_{\mathrm{bin}}(f_W, f_{\mathcal{I}})\right] \leq
\begin{cases}
\frac{1+L}{B} & \text{(for UWB)} \\
(1+L)\left(\frac{1}{B} + \sqrt{\frac{8}{n_{\mathrm{tr,U}}}(\mathrm{fCMI_U} + B\log 2)}\right) & \text{(for UMB)}
\end{cases}
$$

Proof. The UMB case immediately follows from Lemma 1, thus we only need to prove the UWB case.

UMB: By Lemma 4 in Futami & Fujisawa (2024) and Assumption 3, we have

$$
\mathbb{E}_{R,S_{\mathrm{tr}}}\left[|\mathrm{TCE}(f_W) - \mathrm{TCE}(f_{\mathcal{I}})|\right]
$$
$$
\leq \mathbb{E}_{R,S_{\mathrm{tr}}}\left[\mathbb{E}_X\left[|\mathbb{E}\left[Y \mid f_W(X)\right] - \mathbb{E}\left[Y \mid f_{\mathcal{I}}(X)\right]|\right] + \mathbb{E}_X\left[|f_W(X) - f_{\mathcal{I}}(X)|\right]\right]
$$
$$
\leq (1+L)\mathbb{E}_{R,S_{\mathrm{tr}}}\left[\mathbb{E}_X\left[|f_W(X) - f_{\mathcal{I}}(X)|\right]\right]
$$
$$
= (1+L)\mathbb{E}_{R,S_{\mathrm{tr}}}\left[\sum_{b=1}^{B} P(f_W(X) \in I_b)\mathbb{E}\left[|\mathbb{E}\left[f_W(X) \mid f_W(X) \in I_b\right] - f_W(X)| \mid f_W(X) \in I_b\right]\right]
$$
$$
\leq (1+L)\sum_{b=1}^{B}\mathbb{E}_{R,S_{\mathrm{tr}}}\left[|P(f_W(X) \in I_b)|\right]\operatorname*{ess\,sup}_{R,S_{\mathrm{tr}}}|\mathbb{E}\left[|\mathbb{E}\left[f_W(X) \mid f_W(X) \in I_b\right] - f_W(X)| \mid f_W(X) \in I_b\right]|\tag{15}
$$

where we use Hölder inequality in the last line. We first bound $P(f_W(X) \in I_b)$. The proof is similar to the derivation of the Eq. (13). We define the empirical estimate of the average confidence score in bin I_b using $\tilde{X}_M$ and $\tilde{X}_{\bar{M}}$ as

$$
\hat{P}(W, \tilde{X}_M, b) := \frac{1}{n_{\mathrm{tr,U}}}\sum_{i=1}^{n_{\mathrm{tr,U}}}\mathbb{1}_{f_W(\tilde{X}^{\mathrm{U}}_{i,M^{\mathrm{U}}_i}) \in I_b}
$$
$$
\hat{P}(W, \tilde{X}_{\bar{M}}, b) := \frac{1}{n_{\mathrm{tr,U}}}\sum_{i=1}^{n_{\mathrm{tr,U}}}\mathbb{1}_{f_W(\tilde{X}^{\mathrm{U}}_{i,\bar{M}^{\mathrm{U}}_i}) \in I_b}.
$$

Then, we have

$$
\mathbb{E}_{R,\tilde{X},M}\left[P(f_W(X) \in I_b)\right] = \mathbb{E}_{R,\tilde{X},M}\left[\hat{P}(W, \tilde{X}_{\bar{M}}, b)\right]
$$
$$
\leq \mathbb{E}_{R,\tilde{X},M}\left[\hat{P}(W, \tilde{X}_M, b) + \left|\hat{P}(W, \tilde{X}_{\bar{M}}, b) - \hat{P}(W, \tilde{X}_M, b)\right|\right]
$$

The first term can be bounded by $1/B$ from the definition of UMB. For the second term, we consider the sum over bins as an upper bound. We define

$$
\Delta_{\hat{P}}(W, \tilde{X}, M, b) := \left|\hat{P}(W, \tilde{X}_{\bar{M}}, b) - \hat{P}(W, \tilde{X}_M, b)\right|.
$$

Then, by the Donsker-Varadhan lemma, we have

$$
\mathbb{E}_{R,\tilde{X},M}\left[\sum_{b=1}^{B}\Delta_{\hat{P}}(W, \tilde{X}, M, b)\right] \leq \inf_{t>0}\left\{\frac{I(f_W(\tilde{X}^{\mathrm{U}}); M|\tilde{X}) + \mathbb{E}_{\tilde{X}}\left[\log\mathbb{E}_{R,M,M'}\left[e^{t\Delta_{\hat{P}}(W, \tilde{X}, M', b)}\right]\right]}{t}\right\}
$$
$$
\leq \inf_{t>0}\left\{\frac{\mathrm{fCMI_U} + \mathbb{E}_{\tilde{X}}\left[\log\mathbb{E}_{R,M,M'}\left[e^{t\sum_{b=1}^{B}|g(\tilde{X},M,R,M',b)|}\right]\right]}{t}\right\},
$$

where

$$
g(\tilde{X}, M, R, M', b) = \hat{P}(\mathcal{A}(\tilde{X}_M, R), \tilde{X}_{\bar{M}'}, b) - \hat{P}(\mathcal{A}(\tilde{X}_M, R), \tilde{X}_{M'}, b),
$$

and $M' = \left((M^{\mathrm{U}})', (M^{\mathrm{P}})'\right)$ is an independent copy of M. Then, we can upper-bound the second term similarly to the proof of Lemma 5.

$$\mathbb{E}_{M'}\left[e^{t\sum_{b=1}^{B}|g(\tilde{x},m,r,M',b)|}\right] \leq \mathbb{E}_{M'}\left[\prod_{b=1}^{B}\left(e^{tg(\tilde{x},m,r,M',b)} + e^{-tg(\tilde{x},m,r,M',b)}\right)\right]$$

$$= \mathbb{E}_{M'}\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\sum_{b=1}^{B}(-1)^{v_b}g(\tilde{x},m,r,M',b)}\right] \tag{16}$$

We calculate the range of the exponent of the above expression w.r.t. $(M^{\mathrm{U}})'$ and $(M^{\mathrm{P}})'$ in order to apply Hoeffding's lemma. Note that we only need to take into account $(M^{\mathrm{U}})'$ because g does not depend on $\tilde{X}^{\mathrm{P}}$. Let $Z^v = (Z_1^v, \ldots, Z_{n_{\mathrm{tr,U}}}^v)$ be a random variable vector defined as follows:

$$Z_i^v = \frac{1}{n_{\mathrm{tr,U}}}\sum_{b=1}^{B}(-1)^{v_b}\left\{\mathbb{1}_{f_w(\tilde{x}_{i,(\bar{M}^{\mathrm{U}})'_i})\in I_b} - \mathbb{1}_{f_w(\tilde{x}_{i,(M^{\mathrm{U}})'_i})\in I_b}\right\}.$$

Then we can rewrite the Eq. (16) as:

$$\mathbb{E}_{M'}\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\sum_{b=1}^{B}(-1)^{v_b}g(\tilde{x},m,r,M',b)}\right] = \mathbb{E}_{Z^v}\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\sum_{i=1}^{n_{\mathrm{tr,U}}}Z_i^v}\right].$$

We have $\mathbb{E}\left[Z_i^v\right] = 0$ because of symmetry. For the interval length of Z_i^v for all i, because M_i^{U} is 0 or 1, we have

$$\sup_{m,m'\in\{0,1\}}\left|\frac{1}{n_{\mathrm{tr,U}}}\sum_{b=1}^{B}(-1)^{v_b}\left\{\mathbb{1}_{f_w(\tilde{x}_{i,m})\in I_b} - \mathbb{1}_{f_w(\tilde{x}_{i,1-m})\in I_b}\right\}\right.$$

$$\left. - \frac{1}{n_{\mathrm{tr,U}}}\sum_{b=1}^{B}(-1)^{v_b}\left\{\mathbb{1}_{f_w(\tilde{x}_{i,m'})\in I_b} - \mathbb{1}_{f_w(\tilde{x}_{i,1-m'})\in I_b}\right\}\right| \tag{17}$$

$$= \sup_{m\in\{0,1\}}\frac{2}{n_{\mathrm{tr,U}}}\left|\sum_{b=1}^{B}(-1)^{v_b}\left\{\mathbb{1}_{f_w(\tilde{x}_{i,m})\in I_b} - \mathbb{1}_{f_w(\tilde{x}_{i,1-m})\in I_b}\right\}\right|$$

$$\leq \frac{4}{n_{\mathrm{tr,U}}}.$$

Therefore, by Hoeffding's lemma, we have

$$\mathbb{E}_Z\left[\sum_{v_1,\ldots,v_B=0,1} e^{t\sum_{i=1}^{n_{\mathrm{tr,U}}}Z_i^v}\right] \leq \sum_{v_1,\ldots,v_B=0,1} e^{\frac{t^2}{8}\left(n_{\mathrm{tr,U}}\frac{16}{n_{\mathrm{tr,U}}^2}\right)}$$

$$= 2^B e^{\frac{2t^2}{n_{\mathrm{tr,U}}}}.$$

By substituting the above result into the previous equation, we have

$$\mathbb{E}_{R,\tilde{X},M}\left[\sum_{b=1}^{B}\Delta_{\hat{P}}(W,\tilde{X},M,b)\right] \leq \inf_{t>0}\frac{\mathrm{fCMI_U} + \mathbb{E}_{\tilde{X}}\left[\log\mathbb{E}_{R,M,M'}\left[e^{t\Delta_{\hat{P}}(W,\tilde{X},M')}\right]\right]}{t}$$

$$\leq \inf_{t>0}\left\{\frac{\mathrm{fCMI_U} + B\log 2 + \frac{2t^2}{2n_{\mathrm{tr,U}}}}{t}\right\}$$

$$\leq \sqrt{\frac{8}{n_{\mathrm{tr,U}}}\{\mathrm{fCMI_U} + B\log 2\}},$$

and for any $b \in [B]$, we have

$$\mathbb{E}_{R,\tilde{X},M}\left[P(f_W(X) \in I_b)\right] \leq \hat{P}(W, \tilde{X}_M, b) + \Delta_{\hat{P}}(W, \tilde{X}, M, b)$$

$$\leq \hat{P}(W, \tilde{X}_M, b) + \sum_{b=1}^{B} \Delta_{\hat{P}}(W, \tilde{X}, M, b)$$

$$\leq \frac{1}{B} + \sqrt{\frac{8}{n_{\text{tr,U}}}\{\text{fCMI}_{\text{U}} + B\log 2\}}, \tag{18}$$

where the first term at the last inequality is obtained from the definition of UMB. Substituting Eq. (18) into Eq. (15), the term without the coefficient $(1 + L)$ can be bounded as follows:

$$\sum_{b=1}^{B} \mathbb{E}_{R,S_{\text{tr}}}\left[P(f_W(X) \in I_b)\right] \operatorname*{ess\,sup}_{R,S_{\text{tr}}} |\mathbb{E}\left[|\mathbb{E}\left[f_W(X) \mid f_W(X) \in I_b\right] - f_W(X)| \mid f_W(X) \in I_b\right]|$$

$$\leq \sum_{b=1}^{B} \left(\frac{1}{B} + \sqrt{\frac{8}{n_{\text{tr,U}}}\{\text{fCMI}_{\text{U}} + B\log 2\}}\right) \operatorname*{ess\,sup}_{R,S_{\text{tr}}} |\mathbb{E}\left[|\mathbb{E}\left[f_W(X) \mid f_W(X) \in I_b\right] - f_W(X)| \mid f_W(X) \in I_b\right]|$$

$$= \left(\frac{1}{B} + \sqrt{\frac{8}{n_{\text{tr,U}}}\{\text{fCMI}_{\text{U}} + B\log 2\}}\right) \sum_{b=1}^{B} \operatorname*{ess\,sup}_{R,S_{\text{tr}}} |\mathbb{E}\left[|\mathbb{E}\left[f_W(X) \mid f_W(X) \in I_b\right] - f_W(X)| \mid f_W(X) \in I_b\right]|.$$

Finally, we use the fact that $\mathbb{E}\left[|\mathbb{E}\left[f_W(X) \mid f_W(X) \in I_b\right] - f_W(X)| \mid f_W(X) \in I_b\right] \leq u_b - u_{b-1}$ holds by the definition of UMB, where $I_b = (u_{b-1}, u_b]$. Therefore, we have

$$\sum_{b=1}^{B} \operatorname*{ess\,sup}_{R,S_{\text{tr}}} |\mathbb{E}\left[|\mathbb{E}\left[f_W(X) \mid f_W(X) \in I_b\right] - f_W(X)| \mid f_W(X) \in I_b\right]|$$

$$\leq \sum_{b=1}^{B} (u_b - u_{b-1}) = u_B - u_0 = 1.$$

Thus, we can conclude that

$$\mathbb{E}_{R,S_{\text{tr}}}\left[\text{Bias}_{\text{bin}}(f_{\mathcal{A}(\tilde{X}_M,R)}, f_{\mathcal{I}})\right] \leq (1 + L)\left(\frac{1}{B} + \sqrt{\frac{8}{n_{\text{tr,U}}}\{\text{fCMI}_{\text{U}} + B\log 2\}}\right)$$

for UMB [3] $\hfill\square$

Combining Lemma 5 and 6, we can derive the total bias of the PU-ECE on the training data.

C.2 Proof of Corollary 2

UWB: We regard the generalization error as a function of B, $h(B) = \mathbb{E}_{R,S_{\text{tr}}}\left[\text{Bias}_{\text{tot}}(f_W, S_{\text{tr}})\right]$. The first derivative of $h(B)$ is given by:

$$h'(B) = -\frac{1+L}{B^2} + \frac{1}{2}\sqrt{8\left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}}\right)\frac{\log 2}{\text{fCMI} + B\log 2}}$$

Setting $h'(B) = 0$ to find the minimum, we have:

$$B^4 = \frac{1+L}{2\log 2}\left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}}\right)^{-1}(\text{fCMI} + B\log 2)$$

---

[3]Different from our result, Theorem 8 in Futami & Fujisawa (2024) states that the coefficient of the term $\sqrt{(\text{fCMI} + B\log 2)/n}$ (corresponding to $\sqrt{(\text{fCMI}_{\text{U}} + B\log 2)/n_{\text{tr,U}}}$) is $\sqrt{2}$, but it should be $\sqrt{8}$. Although the proof in Futami & Fujisawa (2024) was abbreviated, the technique should be the same as the proof of Lemma 5. More specifically, the equivalent form to Eq. (17) should be used for calculating the bound of the bounded differences.

Solving the above equation for an exact solution is complex. However, we can find the order of B for large $n_{\text{tr,P}}$ and $n_{\text{tr,U}}$. We analyze the order of B by considering three cases for the relationship between fCMI and $B \log 2$:

1. If $\text{fCMI} \gg B \log 2$, we can approximate the equation as:

$$B^4 \approx \frac{1+L}{2 \log 2} \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1} \text{fCMI}.$$

This gives:

$$
\begin{aligned}
B &\approx \left(\frac{1+L}{2 \log 2} \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1} \text{fCMI} \right)^{1/4} \\
&\lesssim \left(\frac{1+L}{2 \log 2} \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-4/3} \right)^{1/4} \\
&= \mathcal{O}\left(\left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1/3} \right).
\end{aligned}
$$

2. If $\text{fCMI} \ll B \log 2$, the approximation becomes:

$$B^4 \approx \frac{1+L}{2 \log 2} \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1} B \log 2.$$

Solving for B yields:

$$
\begin{aligned}
B &\approx \left(\frac{1+L}{2 \log 2} \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1} \log 2 \right)^{1/3} \\
&= \Theta\left(\left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1/3} \right).
\end{aligned}
$$

3. If $\text{fCMI} \approx B \log 2$, the analysis is similar to the second case, leading to:

$$B = \Theta\left(\left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1/3} \right).$$

In all cases, the order of the optimal B is $B^* = \Theta\left(\left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1/3} \right)$. This implies there exists a constant $c > 0$ such that:

$$B^* \le c \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1/3}.$$

Substituting the above inequality into the upper bound of the generalization error, we have:

$$
\begin{aligned}
h(B^*) &\le \frac{1+L}{c \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1/3}} + \sqrt{ 8 \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right) \left\{ \text{fCMI} + c \left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{-1/3} \log 2 \right\} } \\
&= \mathcal{O}\left(\left(\frac{\pi_{\text{P}}^2}{n_{\text{tr,P}}} + \frac{1}{n_{\text{tr,U}}} \right)^{1/3} \right).
\end{aligned}
$$

UMB: For UMB, $h(B)$ can be upper-bounded by:

$$h(B) = \frac{1+L}{B} + \sqrt{8\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)\{\mathrm{fCMI} + B\log 2\}} + (1+L)\sqrt{\frac{8}{n_{\mathrm{tr,U}}}\{\mathrm{fCMI_U} + B\log 2\}}$$

$$\leq \frac{1+L}{B} + \sqrt{8\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)\{\mathrm{fCMI} + B\log 2\}} + (1+L)\sqrt{8\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)\{\mathrm{fCMI} + B\log 2\}}$$

$$= \frac{1+L}{B} + (2+L)\sqrt{8\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)\{\mathrm{fCMI} + B\log 2\}},$$

where we use the fact that $\mathrm{fCMI_U} \leq \mathrm{fCMI}$ in the second inequality. This upper-bound is the same as the UWB case up to the coefficient of the second term. Therefore, we can follow the same analysis as the UWB case. The order of the optimal B is $B^* = \Theta\left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)^{-1/3}\right)$ and there is a constant c such that

$$B^* \leq c\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)^{-1/3}.$$

Substituting the above inequality into the upper bound of the generalization error, we have:

$$h(B^*) \leq \frac{1+L}{c\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)^{-1/3}} + (2+L)\sqrt{8\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)\left\{\mathrm{fCMI} + c\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)^{-1/3}\log 2\right\}}$$

$$= \mathcal{O}\left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)^{1/3}\right).$$

Therefore, we can conclude that the optimal number of bins B is $\Theta\left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)^{-1/3}\right)$ and the generalization error converges at a rate of $\mathcal{O}\left(\left(\frac{\pi_{\mathrm{P}}^2}{n_{\mathrm{tr,P}}} + \frac{1}{n_{\mathrm{tr,U}}}\right)^{1/3}\right)$ for both UWB and UMB.

D   Synthetic Dataset Setup

Following Vaicenavicius et al. (2019); Zhang et al. (2020); Futami & Fujisawa (2024), we use the same logistic model to generate synthetic data. The dataset is generated as follows:

- $P(Y = 1) = 1/2$

- The conditional probability density of X given $Y = 1$ is $p(x \mid Y = 1) = \mathcal{N}(x; 1, 1)$

- The conditional probability density of X given $Y = 0$ is $p(x \mid Y = 0) = \mathcal{N}(x; -1, 1)$

- $P(Y = 1 \mid X = x) = 1/(1 + \exp(-2x))$

- $f(x) = 1/(1 + \exp(-\beta_0 - \beta_1 x))$.

Given the above setup, TCE can be reformulated as:

$$\mathrm{TCE}(f) = \int_{x \in \mathcal{X}} p(x)\,|P(Y = 1 \mid f(X) = f(x)) - f(x)|\ dx$$

$$= \int_{x \in \mathcal{X}} p(x)\,|P(Y = 1 \mid X = x) - f(x)|\ dx$$

$$= \int_{x \in \mathcal{X}} p(x)\left|\frac{1}{1 + \exp(-2x)} - \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}\right|\ dx,$$

where we use the fact that $P(Y = 1 \mid f(X) = f(x)) = P(Y = 1 \mid X = x) = 1/(1 + \exp(-2x))$ since $f(x)$ is injective at the second line. Since the integrand can be analytically computed, we can compute TCE with the trapezoidal rule. Compared to previous works (Zhang et al., 2020; Futami & Fujisawa, 2024) that used Monte-Carlo integration, which converges at a rate of $\mathcal{O}(n^{-1/2})$ (Robert & Casella, 2010) where n is the number of samples, the trapezoidal rule achieves a convergence rate of $\mathcal{O}(n^{-2})$ (Atkinson, 1989), where n is the number of points within the interval of integration, making our experiment more efficient and accurate. In addition, the trapezoidal rule is easier to calculate the case when one of n_P or n_U is ∞, where MC integration has to use the time-consuming rejection method to calculate the binned probability, e.g. $P(f(X) \in I_b)$ We use two cases for the parameters β_0 and β_1:

- Case 1: $\beta_0 = -0.5$ and $\beta_1 = 1.5$ (less calibrated)

- Case 2: $\beta_0 = -0.2$ and $\beta_1 = 1.9$ (better calibrated)

E   TCE Computation

For high-dimensional data X, directly modeling its probability density $p(x)$ is difficult due to the curse of dimensionality. However, TCE can be computed by modeling the one-dimensional probability density of the logit score $h(X)$ or the confidence score $f(X)$. We denote the probability density functions of $h(X)$ and $f(X)$ by $p_{h(X)}(v)$ and $p_{f(X)}(w)$, respectively. For simplicity, we use the notation $p(h(X) = v)$ and $p(f(X) = w)$.

Lemma 7. *Under the Assumption 2, the Total Calibration Error (TCE) can be expressed as:*

$$
\begin{aligned}
\mathrm{TCE}(f) &= \mathbb{E}_X \left[|P(Y = 1 \mid f(X) = f(x)) - f(x)|| \right] \\
&= \int_{v \in \mathbb{R}} p_{h(X)}(v) \left| P(Y = 1 \mid f(X) = \sigma_{\mathrm{sig}}(v)) - \sigma_{\mathrm{sig}}(v) \right| dv \\
&= \int_{w \in [0,1]} p_{f(X)}(w) \left| P(Y = 1 \mid f(X) = w) - w \right| dw.
\end{aligned}
$$

Furthermore, TCE can also be expressed as:

$$
\mathrm{TCE}(f) = \int_{v \in \mathbb{R}} \left| \pi_\mathrm{P} P(h(X) = v \mid Y = 1) - \sigma_{\mathrm{sig}}(v) p_{h(X)}(v) \right| dv
$$

Proof. The first equality follows from the definition of TCE. The second equality follows from the law of unconscious statistician, where we change the variable from x to $v = h(x)$. The third equality follows from the fact that $\sigma_{\mathrm{sig}}(v) = f(x)$. The last equality immediately follows from the Bayes rule. $\qquad \square$

F   Benchmark Dataset Setup

Other than the explanation in the main text (Section 5.1), we provide more details about the benchmark datasets used in our experiments. Training was performed for 100 epochs using AdamW with weight decay of 2.5×10^{-4}, a cosine learning rate scheduler with warmup, and a batch size of 256. To select the best model, 10% of the training data was held out as validation data. At each epoch, we evaluated the misclassification rate for supervised learning and the uPU (du Plessis et al., 2014) 0-1 risk for PU learning (Eq (2)) on the validation data and selected the model with the lowest risk.

G   Extension to Reliability Diagrams

Building on our core idea of estimating $P(Y = 1 \mid f(X) \in I_b)$ from PU data by Bayes' rule (Section 3, Eq. (4)), we propose a method to construct reliability diagrams (Guo et al., 2017; Wilks, 2019) for PU learning. This extends our method from a scalar summary (PU-ECE) to a diagnostic visualization, and both share the same statistical foundation and convergence guarantees.

Reliability diagrams are often discussed alongside ECE as its visual counterpart. They show, for each bin, how the mean confidence score aligns with the empirical frequency of the positive class, i.e., the fraction of positives, helping diagnose over- and under-confidence. Uniform Width Binning (UWB) is commonly used for readability of the plots, whereas Uniform Mass Binning (UMB) is often preferable for lower empirical bias in ECE estimation.

For each bin $I_b = (u_{b-1}, u_b]$, we define

$$|I_{b,\mathrm{U}}| := \sum_{m=1}^{n_\mathrm{U}} \mathbb{1}_{f(X_m^\mathrm{U}) \in I_b}, \quad |I_{b,\mathrm{P}}| := \sum_{m=1}^{n_\mathrm{P}} \mathbb{1}_{f(X_m^\mathrm{P}) \in I_b}, \quad p_{b,\mathrm{U}} := |I_{b,\mathrm{U}}|/n_\mathrm{U}, \quad p_{b,\mathrm{P}} := |I_{b,\mathrm{P}}|/n_\mathrm{P}.$$

For bins with $|I_{b,\mathrm{U}}| > 0$, we compute

$$m_b := \frac{u_{b-1} + u_b}{2}$$

$$\bar{f}_{b,S_\mathrm{PU}} := \frac{\sum_{m=1}^{n_\mathrm{U}} \mathbb{1}_{f(X_m^\mathrm{U}) \in I_b} f(X_m^\mathrm{U})}{|I_{b,\mathrm{U}}|},$$

$$\hat{y}_{b,S_\mathrm{PU}} := \frac{\pi_\mathrm{P}\, p_{b,\mathrm{P}}}{p_{b,\mathrm{U}}} = \frac{\frac{\pi_\mathrm{P}}{n_\mathrm{P}} \sum_{m=1}^{n_\mathrm{P}} \mathbb{1}_{f(X_m^\mathrm{P}) \in I_b}}{\frac{1}{n_\mathrm{U}} \sum_{m=1}^{n_\mathrm{U}} \mathbb{1}_{f(X_m^\mathrm{U}) \in I_b}}.$$

The reliability diagram is then constructed by plotting $(\bar{f}_{b,S_\mathrm{PU}}, \hat{y}_{b,S_\mathrm{PU}})$ (or $(m_b, \hat{y}_{b,S_\mathrm{PU}})$) together with the $y = x$ line as a reference for perfect calibration. If $|I_{b,\mathrm{U}}| = 0$, the estimate $\hat{y}_{b,S_\mathrm{PU}}$ is undefined because the empirical mass in that bin is zero. In this case, we simply omit the point from the diagram, following standard practice in the supervised setting. As $n_\mathrm{P}, n_\mathrm{U} \to \infty$, the points on the diagram $(\bar{f}_{b,S_\mathrm{PU}}, \hat{y}_{b,S_\mathrm{PU}})$ converge to the true counterparts, $(\mathbb{E}\left[f(X) \mid f(X) \in I_b\right], P(Y = 1 \mid f(X) \in I_b))$, respectively. Therefore, our proposed diagram is consistent with the true reliability diagram. In finite samples, clipping $\hat{y}_{b,S_\mathrm{PU}}$ to $[0, 1]$ and adding confidence intervals (e.g., bootstrap) improves interpretability.

Figure 6 illustrates the proposed PU reliability diagram on the less-calibrated case of the synthetic dataset defined in Section D, using $n_\mathrm{P} = 1000$ positive and $n_\mathrm{U} = 10000$ unlabeled samples. As shown in the figure, the reliability diagram constructed from PU data accurately approximates the true reliability diagram, without access to full labels.
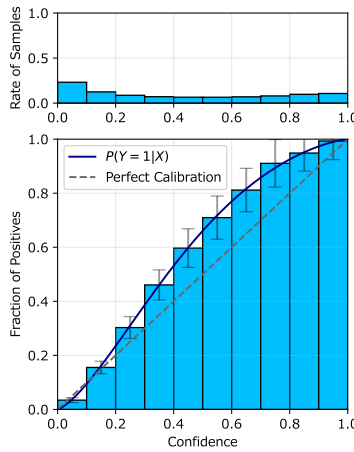

Figure 6: PU reliability diagram on the synthetic dataset (less-calibrated case). The error bars represent the standard deviation over 100 trials with different random seeds.

H   Extension to Multiclass Calibration Metrics

Our proposed PU-ECE can be naturally extended to some multiclass calibration metrics. For example, it can be easily adapted to classwise-ECE (Kull et al., 2019), which evaluates calibration per class. While other multiclass metrics such as top-1 ECE (Guo et al., 2017), may not be directly compatible, they often reduce the multiclass calibration problem to a one-dimensional evaluation based on the maximum predicted score. In this sense, our research provides an essential first step toward calibration in multiclass PU settings.