Mixture Proportion Estimation Beyond Irreducibility

Yilun Zhu [1] **Aaron Fjeldsted** [2] **Darren Holland** [3] **George Landon** [4] **Azaree Lintereur** [2] **Clayton Scott** [1 5]

Abstract

The task of mixture proportion estimation (MPE) is to estimate the weight of a component distribution in a mixture, given observations from both the component and mixture. Previous work on MPE adopts the *irreducibility* assumption, which ensures identifiablity of the mixture proportion. In this paper, we propose a more general sufficient condition that accommodates several settings of interest where irreducibility does not hold. We further present a resampling-based meta-algorithm that takes any existing MPE algorithm designed to work under irreducibility and adapts it to work under our more general condition. Our approach empirically exhibits improved estimation performance relative to baseline methods and to a recently proposed regrouping-based algorithm.

1. Introduction

Mixture proportion estimation (MPE) is the problem of estimating the weight of a component distribution in a mixture. Specifically, let $\kappa^* \in [0, 1]$ and let F, G, and H be probability distributions such that $F = (1 - \kappa^*)G + \kappa^* H$. Given i.i.d. observations

$$\begin{aligned} X_H &:= \{x_1, x_2, \cdots, x_m\} \overset{iid}{\sim} H, \\ X_F &:= \{x_{m+1}, x_{m+2}, \cdots, x_{m+n}\} \overset{iid}{\sim} F, \end{aligned} \tag{1}$$

MPE is the problem of estimating κ^*. A typical application is given some labeled positive reviews X_H, estimate the proportion of positive comments about a product among all comments X_F (González et al., 2017). MPE is also an important component in solving several domain adaptation and weakly supervised learning problems, such as learning from positive and unlabeled examples (LPUE) (Elkan & Noto, 2008; Du Plessis et al., 2014; Kiryo et al., 2017), learning with noisy labels (Lawrence & Schölkopf, 2001; Natarajan et al., 2013; Blanchard et al., 2016), multi-instance learning (Zhang & Goldman, 2001), and anomaly detection (Sanderson & Scott, 2014).

If no assumptions are made on the unobserved component G, then κ^* is not identifiable. Blanchard et al. (2010) proposed the *irreducibility* assumption on G so that κ^* becomes identifiable. Up to now, almost all MPE algorithms build upon the irreducibility assumption (Blanchard et al., 2010; Scott, 2015; Blanchard et al., 2016; Jain et al., 2016; Ramaswamy et al., 2016; Ivanov, 2020; Bekker & Davis, 2020; Garg et al., 2021), or some stricter conditions like non-overlapping support of component distributions (Elkan & Noto, 2008; Du Plessis & Sugiyama, 2014). However, as we discuss below, irreducibility can be violated in several applications, in which case the above methods produce statistically inconsistent estimates. As far as we know, Yao et al. (2022), discussed in Sec. 5, is the first attempt to move beyond irreducibility.

This work proposes a more general sufficient condition than irreducibility, and offers a practical algorithm for estimating κ^* under this condition. We introduce a meta-algorithm that takes as input any MPE method that consistently estimates κ^* under irreducibility, and removes the bias of that method whenever irreducibility does not hold but our more general sufficient condition does. Furthermore, even if our new sufficient condition is not satisfied, our meta-algorithm will not increase the bias of the underlying MPE method. We describe several applications and settings where our framework is relevant, and demonstrate the practical relevance of this framework through extensive experiments. Proofs and additional details can be found in the appendices.

2. Problem Setup and Background

Let G and H be probability measures on a measurable space $(\mathcal{X}, \mathfrak{S})$, and let F be a mixture of G and H

$$F = (1 - \kappa^*)G + \kappa^* H, \tag{2}$$

[1]Department of Electrical Engineering and Computer Science, University of Michigan. [2]Ken and Mary Alice Lindquist Department of Nuclear Engineering, Penn State University. [3]Department of Engineering Physics, Air Force Institute of Technology. [4]School of Engineering and Computer Science, Cedarville University. [5]Department of Statistics, University of Michigan. Correspondence to: Yilun Zhu <allanzhu@umich.edu>, Clayton Scott <clayscot@umich.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

where $0 \leq \kappa^* \leq 1$. With no assumptions on G, κ^* is not uniquely determined by F and H. For example, suppose $F = (1-\kappa^*)G + \kappa^* H$ for some G, and take any $\delta \in [0, \kappa^*]$. Then $F = (1-\kappa^*+\delta)G' + (\kappa^*-\delta)H$, where $G' = (1-\kappa^*+\delta)^{-1}\left[(1-\kappa^*)G + \delta H\right]$, has a different proportion on H (Blanchard et al., 2010).

2.1. Ground-Truth and Maximal Proportion

To address the lack of identifiability, Blanchard et al. (2010) introduced the so-called irreducibility assumption. We now recall this definition and related concepts. Throughout this work we assume that F, G and H have densities f, g and h, defined w.r.t. a common dominating measure μ.

Definition 2.1 (Blanchard et al. (2010))**.** For any two probability distributions F and H, define

$$\kappa(F|H) := \sup\{\kappa \in [0,1] | F = (1-\kappa)G' + \kappa H, \text{ for some distribution } G'\},$$

the maximal proportion of H in F.

This quantity equals the infimum of the likelihood ratio:

Proposition 2.2 (Blanchard et al. (2010))**.** *It holds that*

$$\kappa(F|H) = \inf_{S \in \mathfrak{S}: H(S)>0} \frac{F(S)}{H(S)} = \operatorname*{ess\,inf}_{x: h(x)>0} \frac{f(x)}{h(x)}. \tag{3}$$

By substituting $F = (1-\kappa^*)G + \kappa^* H$ into Eqn. (3), we get that

$$\begin{aligned} \kappa(F|H) &= \inf_{S \in \mathfrak{S}: H(S)>0} \frac{F(S)}{H(S)} \\ &= \kappa^* + (1-\kappa^*) \inf_{S \in \mathfrak{S}: H(S)>0} \frac{G(S)}{H(S)} \\ &= \kappa^* + (1-\kappa^*)\kappa(G|H). \end{aligned} \tag{4}$$

Since $\kappa(F|H)$ is identifiable from F and H, the following assumption on G ensures identifiability of κ^*.

Definition 2.3 (Blanchard et al. (2010))**.** We say that G is *irreducible* with respect to H if $\kappa(G|H) = 0$.

Thus, irreducibility means that there exists *no* decomposition of the form: $G = \gamma H + (1-\gamma)J'$, where J' is some probability distribution and $0 < \gamma \leq 1$. Under irreducibility, κ^* is identifiable, and in particular, equals $\kappa(F|H)$.

2.2. Latent Label Model

We now consider another way of understanding irreducibility in terms of a latent label model. In particular, let X and $Y \in \{0,1\}$ be the random variables characterized by

(a) (X, Y) are jointly distributed

(b) $P(Y=1) = \kappa^*$

(c) $P_{X|Y=0} = G$ and $P_{X|Y=1} = H$.

It follows from these assumptions that the marginal distribution of X is F:

$$P_X = (1-\kappa^*)\, P_{X|Y=0} + \kappa^* P_{X|Y=1} = F.$$

We also take the conditional probability of Y given X to be defined via

$$P(Y=1|X=x) = \begin{cases} \frac{\kappa^* h(x)}{f(x)}, & f(x) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The latent label model is commonly used in the positive unlabeled (PU) learning literature (Bekker & Davis, 2020). MPE is also called class prior/proportion estimation (CPE) in PU learning because $P(Y=1) = \kappa^*$. Y may be viewed as a label indicating which component an observation from F was drawn from. Going forward, we use this latent label model in addition to the original MPE notation.

Proposition 2.4. *Under the latent label model,*

$$\operatorname*{ess\,sup}_x P(Y=1|X=x) = \frac{\kappa^*}{\kappa(F|H)},$$

where $0/0 := 0$.

By definition, we know that G is not irreducible with respect to H iff $\kappa(G|H) > 0$. Combining Proposition 2.4 and Eqn. (4), we conclude that $\operatorname{ess\,sup} P(Y=1|X=x) < 1$ is equivalent to $\kappa(G|H) > 0$.

2.3. Violation of Irreducibility

Up to now, almost all MPE algorithms assume G to be irreducible w.r.t. H (Blanchard et al., 2010; 2016; Jain et al., 2016; Ivanov, 2020), or stricter conditions like the anchor set assumption (Scott, 2015; Ramaswamy et al., 2016), or that G and H have disjoint supports (Elkan & Noto, 2008; Du Plessis & Sugiyama, 2014). These methods return an estimate of $\kappa(F|H)$ as the estimate of κ^*. If irreducibility does not hold and $\kappa^* < 1$, then $\kappa(F|H) > \kappa^*$. Even if these methods are consistent estimators of $\kappa(F|H)$, they are asymptotically biased and thus inconsistent estimators of κ^*.

A sufficient condition for irreducibility to hold is that the support of H is not totally contained in the support of G. In a classification setting where G and H are the class-conditional distributions, this may be quite reasonable. It essentially assumes that each class has at least some subset of instances (with positive measure) that cannot possibly be

confused with the other class. While irreducibility is reasonable in many classification-related tasks, there are also a number of important applications where it does not hold. In this subsection we give three examples of applications where irreducibility is not satisfied.

Ubiquitous Background. In gamma spectroscopy, we may view H as the distribution of the energy of a gamma particle emitted by some source of interest (e.g., Cesium-137), and G as the energy distribution of background radiation. Typically the background emits gamma particles with a wider range of energies than the source of interest does, and therefore its distribution has a wider support: $\text{supp}(H) \subset \text{supp}(G)$, thus violating irreducibility. What's more, G is usually unknown, because it varies according to the surrounding environment (Alamaniotis et al., 2013). The MPE problem is: given observations of the source spectrum H, which may be collected in a laboratory setting, and observations of F in the wild, estimate κ^*. This quantity is important for nuclear threat detection and nuclear safeguard applications.

Global Uncertainty. In marketing, let $Y \in \{0, 1\}$ denote whether a customer does ($Y = 1$) or does not purchase a product (Fei et al., 2013). Let H be the distribution of a feature vector extracted from a customer who buys the product, and G the distribution for those who do not. The MPE problem is: given data from past purchasers of a product (H), and from a target population (F), estimate the proportion κ^* of H in F. This quantity is called the *transaction rate*, and is important for estimating the number of products to be sold. Irreducibility is likely to be violated here because, given a finite number of features, uncertainty about customers should remain bounded away from 1: $\forall x, P(Y = 1|X = x) < 1$. In other words, there is an $\epsilon > 0$ such that, for any feature vector of demographic information, the probability of buying the product is always $< 1 - \epsilon$.

Underreported Outcomes. In public health, let (X, Y, Z) be a jointly distributed triple, where X is a feature vector, $Y \in \{0, 1\}$ denotes whether a person reports a health condition or not, and $Z \in \{0, 1\}$ indicates whether the person truly has the health condition. Here, $H = P_{X|Y=1}, G = P_{X|Y=0}$ and $F = P_X$. The MPE problem is: given data from previous people who reported the condition (H), and from a target group (F), determine the *prevalence* κ^* of the condition for the target group. This helps estimate the amount of resources needed to address the health condition. Assume there are no false reports from those who do not have the medical condition: $\forall x, P(Y = 1|Z = 0, X = x) = 0$. Some medical conditions are underreported, such as smoking and intimate partner violence (Gorber et al., 2009; Shanmugam & Pierson, 2021). If the underreporting happens globally, meaning $e(x) := P(Y = 1|Z = 1, X = x)$ is bounded away from 1, then $\text{ess sup } P(Y = 1|X = x) < 1$. This is because $P(Y = 1|X = x) = P(Y = 1|Z = 0, X = x)P(Z = 0|X = x) + P(Y = 1|Z = 1, X = x)P(Z = 1|X = x) \leq e(x)$. In this situation, irreducibility is again violated.

In the above situations, irreducibility is violated and $\kappa(F|H) > \kappa^*$. Estimating $\kappa(F|H)$ alone leads to bias. In the following, we will re-examine mixture proportion estimation. In particular, we propose a more general sufficient condition than irreducibility, and introduce an estimation strategy that calls an existing MPE method and reduces or eliminates its asymptotic bias.

3. A General Identifiability Condition

Previous MPE works assume irreducibility. We propose a more general sufficient condition for recovering κ^*.

Theorem 3.1 (Identifiability Under Local Supremal Posterior (LSP))**.** *Let A be any non-empty measurable subset of $E_H = \{x : h(x) > 0\}$ and $s = \operatorname{ess\,sup}_{x \in A} P(Y = 1|X = x)$. Then*

$$\kappa^* = s \cdot \inf_{S \subseteq A} \frac{F(S)}{H(S)} = s \cdot \operatorname*{ess\,inf}_{x \in A} \frac{f(x)}{h(x)}.$$

This implies that under LSP, κ^* is identifiable. Two special cases are worthy of comment. First, irreducibility holds when $A = E_H$ and $s = 1$, in which case the above theorem recovers the known result that $\kappa^* = \kappa(F|H)$. Second, when $A = \{x_0\}$ is a singleton, then $s = P(Y = 1|X = x_0)$ and $\kappa^* = P(Y = 1|X = x_0) \cdot \frac{f(x_0)}{h(x_0)}$, which can also be derived directly from the definition of conditional probability.

The above theorem gives a general sufficient condition for recovering κ^*, but estimating $\inf_{S \subseteq A} \frac{F(S)}{H(S)}$ is non-trivial: when $A = E_H$, it can be estimated using existing MPE methods (Blanchard et al., 2016). When A is a proper subset, however, a new approach is needed. We now present a variation of Theorem 3.1 that lends itself to a practical estimation strategy without having to devise a completely new method of estimating $\inf_{S \subseteq A} \frac{F(S)}{H(S)}$.

Theorem 3.2 (Identifiability Under Tight Posterior Upper Bound)**.** *Consider any non-empty measurable set $A \subseteq E_H = \{x : h(x) > 0\}$, and let $s = \operatorname{ess\,sup}_{x \in A} P(Y = 1|X = x)$. Let $\alpha(x)$ be any measurable function satisfying*

$$\alpha(x) \in \begin{cases} [P(Y = 1|X = x), s], & x \in A, \\ [P(Y = 1|X = x), 1], & \textit{o.w.} \end{cases} \tag{6}$$

Define a new distribution $\widetilde{F}$ in terms of its density

$$\widetilde{f}(x) = \frac{1}{c} \cdot \alpha(x) \cdot f(x), \quad \textit{where} \quad c = \int \alpha(x) f(x) dx = \mathbb{E}_{X \sim F}\left[\alpha(X)\right]. \tag{7}$$

Then

$$\kappa^* = c \cdot \kappa(\widetilde{F}|H).$$

The theorem can be re-stated as: κ^* is identifiable given an upper bound of the posterior probability $\alpha(x) \geq P(Y = 1|X = x)$ that is tight for some $x \in A$. One possible choice for $\alpha(x)$ is simply

$$\alpha(x) = \begin{cases} s, & x \in A, \\ 1, & \text{o.w.} \end{cases}$$

If the conditional probability $P(Y = 1|X = x)$ is known for all $x \in A$, then

$$\alpha(x) = \begin{cases} P(Y = 1|X = x), & x \in A, \\ 1, & \text{o.w.}, \end{cases}$$

may be chosen.

Having $\alpha(x)$ satisfying Eqn. (6) ensures identifiablility of κ^*. Relaxing this requirement slightly still guarantees that the bias will not increase.

Corollary 3.3. *Let $\alpha(x)$ be any measurable function with*

$$\alpha(x) \in [P(Y = 1|X = x), 1] \quad \forall x. \tag{8}$$

Define a new distribution $\widetilde{F}$ in terms of its density $\widetilde{f}$ according to Eqn. (7)*. Then*

$$\kappa^* \leq c \cdot \kappa(\widetilde{F}|H) \leq \kappa(F|H).$$

This shows that even if we have a non-tight upper bound on $P(Y = 1|X = x)$, the quantity $c \cdot \kappa(\widetilde{F}|H)$ is still bounded by $\kappa(F|H)$. Therefore, a smaller asymptotic bias may be achieved by estimating $c \cdot \kappa(\widetilde{F}|H)$ instead of $\kappa(F|H)$.

The intuition underlying the above results is that the new distribution $\widetilde{F}$ is generated by throwing away some probability mass from G, and therefore can be viewed as a mixture of H and a new $\widetilde{G}$, but now $\widetilde{G}$ tends to be irreducible w.r.t. H. The proportion $\kappa(\widetilde{F}|H)$ relates to the original proportion κ^* by a scaling constant c. This interpretation is supported mathematically in Appendix B.1.

4. Subsampling MPE (SuMPE)

Theorem 3.2 directly motivates a practical algorithm. We obtain a new distribution $\widetilde{F}$ from F by rejection sampling (MacKay, 2003), which is a Monte Carlo method that generates a sample following a new distribution $\widetilde{Q}$ based on a sample from distribution Q, in terms of their densities $\widetilde{q}$ and q. An instance x drawn from $q(x)$ is kept with acceptance probability $\beta(x) \in [0, 1]$, and rejected otherwise. Appendix B.2 shows the detailed procedure. In our scenario, $\widetilde{Q} = \widetilde{F}$, $Q = F$ and $\beta(x) = \alpha(x)$.

Algorithm 1 Subsampling MPE (SuMPE)

1: **Input:**
   X_F: sample drawn i.i.d. from F
   X_H: sample drawn i.i.d. from H
   $\alpha(x)$: acceptance function
2: **Output:**
   Estimate of κ^*
3: Generate $X_{\widetilde{F}}$ from X_F by rejection sampling (Algorithm 4), with acceptance function $\alpha(x)$.
4: Compute $\widehat{c} = \left|X_{\widetilde{F}}\right| / \left|X_F\right|$, where $|\cdot|$ denotes the cardinality of a set.
5: Apply an off-the-shelf MPE algorithm to produce an estimate $\widehat{\kappa}(\widetilde{F}|H)$ from $X_{\widetilde{F}}$ and X_H.
6: **return** $\widehat{c} \cdot \widehat{\kappa}(\widetilde{F}|H)$

4.1. Method

Our Subsampling MPE algorithm, SuMPE (Algorithm 1), follows directly from Theorem 3.2. It first obtains in line 3 a data sample $X_{\widetilde{F}}$ following distribution $\widetilde{F}$ using rejection sampling and in line 4 estimates the normalizing constant c. Then in line 5, it computes an estimate $\widehat{\kappa}(\widetilde{F}|H)$ using any existing MPE method that consistently estimates the mixture proportion under irreducibility. The final estimate is returned as the product of $\widehat{c}$ and $\widehat{\kappa}(\widetilde{F}|H)$.

Rejection sampling in high dimensional settings may be inefficient due to a potentially low acceptance rate (MacKay, 2003). However, this concern is mitigated in our setting because the acceptance rate can be taken to be 1 except on the set A, which is potentially a small set.

One advantage of building our method around existing MPE methods is that we may adapt known theoretical results to our setting. To illustrate this, we give a rate of convergence result for SuMPE.

Theorem 4.1. *Assume $\alpha(x)$ satisfies the condition in Eqn.* (6)*. After subsampling, assume the resulting $\widetilde{F}$ and H are such that $\operatorname{arg\,min}_{S \in \mathcal{A}: H(S) > 0} \frac{\widetilde{F}(S)}{H(S)}$ exists. Then there exists a constant $C > 0$ and an existing MPE estimator $\widehat{\kappa}$ such that, for m and n sufficiently large, the estimator $\widehat{\kappa}^*$ obtained*

from SuMPE (Algorithm 1) satisfies

$$\Pr\left(|\widehat{\kappa}^* - \kappa^*| \leq C\left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \geq 1 - \mathcal{O}\left(\frac{1}{m} + \frac{1}{n}\right).$$

4.2. Practical Scenarios

Our new sufficient condition assumes knowledge of some set $A \subseteq E_H$ and $s = \operatorname{ess\,sup}_{x \in A} P(Y = 1|X = x)$. However, practically speaking, our algorithm only requires an $\alpha(x)$ satisfying Eqn. (6) for some $A \subseteq E_H$ and the associated value of s, and does not require the explicit knowledge of A and s. Additionally, even if $\alpha(x)$ does not satisfy Eqn. (6) , as long as it satisfies Eqn. (8) (which is easier to achieve), it shall perform no worse than directly applying off-the-shelf MPE methods.

There are settings where a generic construction of $\alpha(x)$ is possible. For example, suppose the user has access to fully labeled data (where it is known which of G or H each instance came from) but only on a subset A of the domain. This may come from an annotator who is only an expert on a subset of instances. This data should be sufficient to get a non-trivial upper bound on the posterior class probability $P(Y|X)$, which in turns leads to an $\alpha(x)$.

More typically, however, it may be necessary to determine $\alpha(x)$ on a case by case basis. This section continues the discussion of the three applications introduced in Sec. 2.3. Each of these three settings leverages different domain-specific knowledge in different ways, and we believe this leads to the best $\alpha(x)$ compared to a one-size-fits-all construction.

4.2.1. UNFOLDING

Unfolding refers to the process of recovering one or more true distributions from contaminated ones (Cowan, 1998). In gamma spectrum unfolding (Li et al., 2019), a gamma ray detector measures the energies of incoming gamma rays. The gamma rays were emitted either by a source of interest or from the background. The measurement is represented as a histogram $f(x)$ where the bins correspond to a quantization of energy. In many settings, the histogram $h(x)$ of measurements from the source of interest is also known. In this case, unfolding amounts to the estimation of the unknown background histogram $g(x)$. Toward this goal, it suffices to estimate the proportion of recorded particles κ^* emanating from the source of interest, since $g(x) = (f(x) - \kappa^* h(x))/(1 - \kappa^*)$. This application corresponds to the "ubiquitious background" setting described in Sec. 2.3, where irreducibility may not hold since the source of interest energies can be a subset of the background energies.

Using existing techniques from the nuclear detection literature (Knoll, 2010; Alamaniotis et al., 2013), we can obtain a lower bound $\rho(x)$ of the quantity $(1 - \kappa^*)g(x)$ on a certain set $A \subset \operatorname{supp}(H)$ (see Appendix B.3 for details). This leads to the acceptance function

$$\alpha(x) = \begin{cases} 1 - \frac{\rho(x)}{f(x)}, & x \in A, \\ 1, & \text{o.w.}, \end{cases} \tag{9}$$

which is an upper bound of $P(Y = 1|X = x)$, satisfying the condition in Corollary 3.3.

4.2.2. CPE UNDER DOMAIN ADAPTATION

In the problem of domain adaptation, the learner is given labeled examples from a source distribution, and the task is to do inference on a potentially different target distribution. Previous work on domain adaptation mainly focuses on classification and typically makes assumptions about which of the four distributions $P_X, P_{Y|X}, P_Y$, and $P_{X|Y}$ vary between the source and target. This leads to situations such as covariate shift (where P_X changes) (Heckman, 1979), posterior drift (where $P_{Y|X}$ changes) (Scott, 2019), prior/target shift (where P_Y changes) (Storkey et al., 2009), and conditional shift (where $P_{X|Y}$ changes) (Zhang et al., 2013). It is also quite commonly assumed that the support of source distribution contains the support of target (Heckman, 1979; Bickel et al., 2009; Gretton et al., 2009; Storkey et al., 2009; Zhang et al., 2013; Scott, 2019).

We study class proportion estimation (CPE) under domain adaptation. Prior work on this topic has considered distributional assumptions like those described above (Saerens et al., 2002; Sanderson & Scott, 2014; González et al., 2017). In this work, we consider the setting where, in addition to labeled examples from the source, the learner has access to labeled positive and unlabeled data from the target. We propose a model that includes covariate shift and posterior drift as special cases. We use P^{sr}_{XY} and P^{tg}_{XY} to denote source and target distributions. In MPE notation, $F = P^{tg}_X$, $G = P^{tg}_{X|Y=0}$ and $H = P^{tg}_{X|Y=1}$.

Definition 4.2 (CSPL)**.** We say that *covariate shift with posterior lift* occurs whenever

$$\forall x \in \operatorname{supp}(P^{sr}_X) \bigcap \operatorname{supp}(P^{tg}_X),$$
$$P^{sr}(Y = 1|X = x) \geq P^{tg}(Y = 1|X = x),$$

and "=" holds for some $x \in \operatorname{supp}(P^{sr}_X) \bigcap \operatorname{supp}(P^{tg}_{X|Y=1})$.

Covariate shift is a special case of CSPL when equality always holds. One motivation for posterior lift is to model

labels produced by an annotator who is biased toward one class. It is a type of posterior drift model wherein the posterior changes from source to target (Scott, 2019; Cai & Wei, 2021; Maity et al., 2023). Also notice that CSPL does not require the support of the source distribution to contain the target, nor irreducibility.

CSPL is motivated by a marketing application mentioned in Sec. 2.3. In marketing, companies often have access to labeled data from a source distribution, such as survey results where customers express their interest in a product. Additionally, they also have access to labeled positive and unlabeled data from the target distribution, which corresponds to actual purchasing behavior. In this scenario, the CSPL assumption is often met as it is more likely for customers to express interest than to actually make a purchase: $P^{sr}(Y=1|X=x) \geq P^{tg}(Y=1|X=x)$.

Although irreducibility is violated in the marketing application due to the "global uncertainty" about a target customer buying the product (see Sec. 2.3), CSPL ensures the identifiability of $\kappa^* = P^{tg}(Y=1)$ because we can choose the set $A = \text{supp}(P_X^{sr}) \bigcap \text{supp}(P_{X|Y=1}^{tg})$ and the acceptance function as

$$\alpha(x) = \begin{cases} P^{sr}(Y=1|X=x), & x \in A, \\ 1, & \text{o.w.}, \end{cases} \tag{10}$$

which satisfies the identifiability criteria in Theorem 3.2. [1] By using the labeled source data, an estimate of $P^{sr}(Y=1|X=x)$ can be obtained and used as the acceptance function $\alpha(x)$ in Algorithm 1 to do CPE.

4.2.3. Selected/Reported at Random

In public health, (X, Y, Z) are jointly distributed, where X is the feature vector, $Y \in \{0,1\}$ denotes whether a person reports a medical condition or not and $Z \in \{0,1\}$ indicates whether a person truly has the medical condition. The goal is to estimate the proportion of people in X_F that report the medical condition. This setting was described in Sec. 2.3 as "underreported outcomes" where it was argued that irreducibility may not hold, in which case estimating $\kappa(F|H)$ overestimates the true value of κ^*. Our SuMPE framework provides a way to eliminate the bias.

The behavior of underreporting can be captured using the *selection bias* model (Kato et al., 2018; Bekker et al., 2019; Gong et al., 2021). Denote the probability of reporting as $e(x) := P(Y=1|X=x, Z=1)$. Assume there is no false report: $\forall x, P(Y=1|X=x, Z=0) = 0$. We use the notation $p(x|\Omega)$ to indicate the conditional density of X given the event Ω. Then $p(x|Y=1) = \frac{e(x)}{\nu} p(x|Z=1)$, where $\nu = P(Y=1|Z=1)$. Under this model, the density of marginal distribution P_X can be decomposed as

$$\begin{aligned} p(x) &= (1-\alpha) \cdot p(x|Z=0) + \alpha \cdot p(x|Z=1) \\ &= (1-\alpha\nu) \cdot p(x|Y=0) + \alpha\nu \cdot p(x|Y=1), \end{aligned}$$

where $\alpha = P(Z=1)$ is the proportion of people having the medical condition. The mixture proportion to be estimated is $\kappa^* = P(Y=1) = P(Y=1, Z=1) = P(Z=1)P(Y=1|Z=1) = \alpha\nu$.

We assume access to i.i.d. sample from $H = P_{X|Y=1}$, representing the public survey data where people report the presence of the medical condition, and from $F = P_X$, representing the target group. Further assume that $A \subseteq \{x : P(Z=1|X=x) = 1\}$. This is a subset of patients who are guaranteed to have the condition, which could be obtained based on historical patient data from hospital. Then the mixture proportion $\kappa^* = \alpha\nu$ can be recovered from Algorithm 1, where the acceptance function is

$$\alpha(x) = \begin{cases} e(x), & x \in A, \\ 1, & \text{o.w.} \end{cases} \tag{11}$$

$\alpha(x)$ satisfies the condition in Theorem 3.2 . This is because under no-false-report assumption, $\forall x \in A, P(Y=1|X=x) = P(Y=1, Z=1|X=x) = P(Y=1|X=x, Z=1) \cdot P(Z=1|X=x) = e(x) \cdot 1 = e(x)$. [2] In practice, $e(x)$ can be estimated from labeled examples $(X, Y, Z=1)$.

5. Limitation of Previous Work

Previous research by Yao et al. (2022) introduced the Regrouping MPE (ReMPE) method [3], which is built on top of any existing MPE estimator (just like our meta-algorithm). They claimed that ReMPE works as well as the base MPE method when irreducibility holds, while improving the performance when it does not. In this section we offer some comments on ReMPE.

Regrouping MPE in theory. Consider any F, H, G, κ^* such that Eqn. (2) holds. Write G as an arbitrary mixture of two distributions $G = \gamma G_1 + (1-\gamma)G_2, \gamma \geq 0$. Then F

[1] To see this, take $A' = \{x \in A : P^{sr}(Y=1|X=x) = P^{tg}(Y=1|X=x)\}$ and $s' = \operatorname{ess\,sup}_{x \in A'} P^{sr}(Y=1|X=x) = \operatorname{ess\,sup}_{x \in A'} P^{tg}(Y=1|X=x)$. Then A' and s' are the A and s in Theorem 3.2.

[2] Take $A' = \{x \in A : e(x) > 0\}$ and $s' = \operatorname{ess\,sup}_{x \in A'} e(x) = \operatorname{ess\,sup}_{x \in A'} P(Y=1|X=x)$. Then A' and s' are the A and s in Theorem 3.2.

[3] Yao et al. (2022) called the method Regrouping CPE (ReCPE).

Algorithm 2 ReMPE-1 (Yao et al., 2022)

Input: Distributions F and H
Obtain set $B = \arg\min_{S \in \mathfrak{S}} \frac{G(S)}{H(S)}$
Generate new distribution H' by Eqn. (12), where $G_1 = G_B$
return $\kappa(F|H')$

can be re-written as

$$\begin{aligned} F &= (1-\kappa^*)G + \kappa^* H \\ &= (1-\kappa^*)\left[\gamma G_1 + (1-\gamma)G_2\right] + \kappa^* H \\ &= (1-\kappa^*)(1-\gamma)G_2 + \underbrace{\left[(1-\kappa^*)\gamma G_1 + \kappa^* H\right]}_{\text{Regrouped}} \quad (12) \\ &= (1-\kappa')G_2 + \kappa' H', \end{aligned}$$

where $\kappa' = \kappa^* + (1-\kappa^*)\gamma$. Yao et al. (2022) assumes there exists a set such that the infimum in Eqn. (3) and (4) can be achieved. They proposed to specify G_1 as the truncated distribution of G in set B, denote as G_B, where $B = \arg\min_{S \in \mathfrak{S}} \frac{G(S)}{H(S)}$. This specific choice causes the resulting distribution G_2 to be irreducible w.r.t. H' and the bias introduced by regrouping $(1-\kappa^*)G(A)$ to be minimal. Denote the above procedure as *ReMPE-1* (Algorithm 2). Theorem 2 in Yao et al. (2022) provides a theoretical justification for ReMPE-1, which we will restate here.

Theorem 5.1 (Yao et al. (2022))**.** *Let $\kappa(F|H')$ be the mixture proportion obtained from ReMPE-1. 1) If G is irreducible w.r.t. H, then $\kappa(F|H') = \kappa^*$. 2) if G is not irreducible w.r.t. H, then $\kappa^* < \kappa(F|H') < \kappa(F|H)$.*

While this theorem is valid, we note that in the case where G is irreducible w.r.t. H, the set B is outside the support of G, and therefore it is not appropriate to describe the procedure as "regrouping G." In fact, performing regrouping ($\gamma > 0$) always introduces a positive bias, because $\kappa(F|H') \geq \kappa' > \kappa^*$. This indicates that any kind of regrouping will have a positive bias under irreducibility.

Regrouping MPE in practice. Yao et al. (2022)'s practical implementation of regrouping deviates from the theoretical proposal. Here, we state and analyze the idealized version of their practical algorithm, referred to as *ReMPE-2* (Algorithm 3). ReMPE-2 does not rely on the knowledge of κ^* and $G(B)$ as outlined in Eqn. (12). Instead, the set B is chosen based solely on F and H, and the distribution H' is obtained through regrouping some probability mass from F rather than G. [4]

[4] Yao et al. (2022)'s real implementation differs a bit from ReMPE-2 in that instead of choosing $\arg\min_{S \in \mathfrak{S}} \frac{F(S)}{H(S)}$, they select $p = 10\%$ of examples drawn from F with smallest estimated score of $f(x)/h(x)$.

Algorithm 3 ReMPE-2 (Yao et al., 2022)

Input: Distributions F and H
Obtain set $B = \arg\min_{S \in \mathfrak{S}} \frac{F(S)}{H(S)}$
Generate new distribution $H' = \frac{1}{1+F(B)}(F_B + H)$
return $\kappa(F|H')$

ReMPE-2 is fundamentally different from ReMPE-1 in that it uses a different way to construct H'. To be specific, when the irreducibility assumption holds, ReMPE-1 suggests regrouping nothing (because $G(B) = 0$), but ReMPE-2 still regroups a proportion $F(B)$ from F to H. Therefore, Theorem 5.1 does not apply to it. Yao et al. (2022) did not analyze ReMPE-2, but the next result shows that it has a negative bias under irreducibility.

Proposition 5.2. *For $\kappa(F|H')$ obtained from ReMPE-2:*

$$\kappa(F|H') < \kappa(F|H).$$

Thus, if irreducibility holds, then ReMPE-2 returns $\kappa(F|H') < \kappa(F|H) = \kappa^*$, which is undesirable. However, when irreducibility does not hold, ReMPE-2 may lead to a smaller asymptotic bias than estimating $\kappa(F|H)$, which could explain why the authors observe empirical improvements in their results. Our theoretical analysis of ReMPE-2 is supported experimentally in Sec. 6 and Appendix D.

To summarize, Yao et al. (2022) proposed a regrouping approach that was the first attempt to tackle the problem of MPE beyond irreducibility and motivated our work. ReMPE-1 recovers κ^* when irreducibility holds (although in this case it is not doing regrouping), and decreases bias when irreducibility does not hold. The more practical algorithm ReMPE-2 might decrease the bias when irreducibility does not hold, but it has a negative bias when irreducibility does hold. Like ReMPE-1, SuMPE draws on some additional information beyond F and H. Both meta-algorithms do not increase the bias, and recover κ^* when irreducibility holds. Unlike ReMPE-1, however, SuMPE is able to recover κ^* under a more general condition. Furthermore, our practical implementations of subsampling are based directly on Theorem 3.2, unlike ReMPE-2 which does not have the desireable theoretical properties of ReMPE-1. Finally, as we argue in the next section, SuMPE offers significant empirical performance gains.

One limitation of our SuMPE framework is that some knowledge of $P(Y|X)$ is needed and that $\alpha(x)$ may need to be developed specifically for different applications.

6. Experiments

We ran our algorithm on nuclear, synthetic and some benchmark datasets taken from the UCI machine learning repository and MNIST, corresponding to all three scenarios

described in Sec. 4.2. We take four MPE algorithms: DPL (Ivanov, 2020), EN (Elkan & Noto, 2008), KM (Ramaswamy et al., 2016) and TIcE (Bekker & Davis, 2018). We compare the original version of these methods together with their regrouping (Re-$(\cdot)$) and subsampling (Su-$(\cdot)$) version. All experiments in this section consider settings where irreducibility is violated. The summarized results are shown in Table 1 and the detailed results are in Appendix C. Overall, the subsampling version of each MPE method outperforms the original and regrouping version. Additional experiments where irreducibility holds are offered in Appendix D, where we find that ReMPE harms the estimation performance while SuMPE does not. The implementation is available at `https://github.com/allan-z/SuMPE`.

6.1. Unfolding: Gamma Ray Spectra Data

The gamma ray spectra data are simulated from the Monte-Carlo N-Particle (MCNP) radiation transport code (Werner et al., 2018). H refers to the distribution of Cesium-137. G is the background distribution, consisting of terrestrial background and Cobalt-60. The goal is to estimate the proportion of Cesium-137 in the measured spectrum F.

$$
\begin{aligned}
&H \sim p(x|\text{Cesium}) \\
&G \sim 0.8 \cdot p(x|\text{Cobalt}) + 0.2 \cdot p(x|\text{terrestrial background}) \\
&F = (1-\kappa^*)G + \kappa^* H.
\end{aligned}
$$

Sample sizes of $m = n = 50,000$ were chosen, which is a reasonable number of counts for many nuclear detection applications. The true mixture proportion κ^* is varied in $\{0.1, 0.25, 0.5, 0.75\}$. The random variable x represents quantized energy, which is one-dimensional and is discrete-valued. Therefore, we directly use the histogram as the estimate of the distribution. We choose the acceptance function $\alpha(x)$ according to the methodology developed in Sec. 4.2.1 and Appendix B.3.

Three of the four baseline MPE methods (DPL, EN, KM) did not work well out-of-the-box in this setting. We therefore re-implemented these methods to explicitly leverage the histogram representation of the probability distributions. This also greatly sped up the KM approach. The results are summarized in Table 2.

6.2. Domain Adaptation: Synthetic Data

Following the setup in Sec. 4.2.2, we specify the target conditional and marginal distributions H, G and F as:

$$
\begin{aligned}
&H \sim \mathcal{N}(\mu_1 = 0, \sigma_1 = 1) \\
&G \sim 0.8 \cdot \mathcal{N}(\mu_2 = 3, \sigma_2 = 2) + 0.2 \cdot \mathcal{N}(\mu_3 = 4, \sigma_3 = 1) \\
&F = (1-\kappa^*)G + \kappa^* H.
\end{aligned}
$$

where G is not irreducible w.r.t. H because it contains a Gaussian distribution with a bigger variance than H. We draw $m = n = 1000$ instances from both H and F. The true mixture proportion κ^* is varied in $\{0.1, 0.25, 0.5, 0.75\}$.

In addition, we draw 4000 labeled instances from the source distribution, where μ_3 is changed to 5. We then truncate the source distribution (and therefore the source sample) to $(-\infty, 2]$. The resulting source and target distribution satisfy the CSPL assumption. A 1 hidden layer neural network with 16 neurons was trained to predict $P^{sr}(Y = 1|X = x)$ for $x \in (-\infty, 2]$, thus $A = (-\infty, 2]$ and $\alpha(x)$ was chosen according to Eqn. (10). This procedure was repeated 10 times with different random seeds. Detailed results are shown in Table 3.

6.3. Domain Adaptation: Benchmark Data

In many machine learning datasets for classification (e.g., those in UCI), irreducibility is satisfied. [5] Here we manually create datasets that violate irreducibility by uniformly sampling out 90% (or 80%) of the original positive data as the target positive data, with all the remaining data treated as target negative. The target conditional and marginal distributions H, G and F are specified as:

$$
\begin{aligned}
&H \sim p(x|Y = 1) \\
&G \sim \gamma p(x|Y = 1) + (1-\gamma) p(x|Y = 0) \\
&F = (1-\kappa^*)G + \kappa^* H.
\end{aligned}
$$

We draw $m = 1000$ instances from H and $n \in [1000, 4000]$ instances from F (the exact number varies by datasets and is based on number of examples available, see the code provided). The true mixture proportion κ^* is varied in $\{0.1, 0.25, 0.5, 0.75\}$. The proportion γ is determined by the total number of positive and negative data originally available in the dataset, therefore varies case by case.

In addition, we obtain labeled instances following the source distribution, by drawing $\kappa^* n$ data from the target positive and $0.95(1-\kappa^*)n$ data from the target negative distribution. This causes a prior shift that simulates CSPL.

A 2 hidden layer neural network with 512 neurons was trained to predict $P^{sr}(Y = 1|X = x)$. For real-world high-dimensional data, it is hard to know the support. Therefore,

[5] The datasets Mushroom, Landsat, Shuttle and MNIST were chosen for our study because previous empirical research (Ivanov, 2020) showed that the baseline MPE methods perform well on these datasets when the irreducibility assumption holds. Our paper focuses on how to eliminate the estimation bias that arises from the violation of irreducibility. Therefore, we chose to use datasets where the baseline methods perform well, in order to clearly observe and measure the bias introduced by MPE methods when irreducibility is not met.

Table 1. Summarized table of average absolute estimation error, corresponding to testing cases in Sec. 6. Several state-of-the-art MPE algorithms DPL, EN, KM and TIcE are selected. The mean absolute error (avg$[|\widehat{\kappa}^* - \kappa^*|]$) is reported, the smallest error among original, regrouping and subsampling version is bolded. $(+/-)$ denotes that on average the estimator produces positive/negative estimation bias (sgn[avg$(\widehat{\kappa}^* - \kappa^*)$]).

Setup	Dataset	DPL	ReDPL	SuDPL	EN	ReEN	SuEN	KM	ReKM	SuKM	TIcE	ReTIcE	SuTIcE
Unfolding	Gamma Ray	0.045+	0.117−	**0.013+**	0.034+	0.118−	**0.027−**	0.163+	0.076+	**0.042+**	0.095+	0.061+	**0.019+**
Domain Adaptation	Synthetic	0.060+	0.053+	**0.028−**	**0.045−**	0.061−	0.067−	0.063+	0.059+	**0.022−**	0.128+	0.094+	**0.041+**
	Mushroom	0.047+	0.081−	**0.033+**	0.122+	**0.078+**	0.101+	0.067−	0.134−	**0.059−**	0.060+	0.078−	**0.036+**
	Landsat	0.046+	0.042−	**0.017−**	0.141+	0.110+	**0.085+**	0.046+	0.029+	**0.016−**	0.043+	0.044−	**0.035−**
	Shuttle	0.037+	0.138−	**0.015−**	0.090+	0.071+	**0.046+**	0.036+	0.111−	**0.032−**	0.080+	0.086−	**0.038+**
	MNIST17	0.047+	0.085−	**0.028−**	0.231+	0.175+	**0.166+**	0.041+	0.063−	**0.017−**	0.090+	0.073−	**0.048+**
Selected/ Reported at Random	Mushroom	0.047+	0.095−	**0.027+**	0.119+	0.075+	**0.074+**	0.072−	0.134−	**0.064−**	0.047+	0.066−	**0.044−**
	Landsat	0.048+	0.057−	**0.019−**	0.142+	0.108+	**0.092+**	0.046+	0.033+	**0.018+**	0.046+	0.053−	**0.041−**
	Shuttle	0.035+	0.144−	**0.013−**	0.095+	0.073+	**0.056+**	**0.042+**	0.129−	0.044−	0.079+	0.089−	**0.050+**
	MNIST17	0.047+	0.088−	**0.027+**	0.240+	0.173+	**0.164+**	0.038+	0.064−	**0.022+**	0.096+	0.073+	**0.057+**

we choose $A = \{x : \widehat{P}^{sr}(Y = 1|X = x) > 0.5\}$, because an example x with high $\widehat{P}^{sr}(Y = 1|X = x)$ is more likely to lie in the support of H. The acceptance function $\alpha(x)$ was determined according to Eqn. (10). The above procedure was repeated 10 times with different random seeds. Table 4 summarizes the results.

6.4. Selected/Reported at Random: Benchmark Data

Recalling the setting of Sec. 4.2.3, there is a jointly distributed triple (X, Y, Z), where Y indicates whether a condition is reported, and Z indicates whether the condition is actually present. For the experiments below, the data from F, G, and H are generated in the same way as the target distribution of the previous subsection. Instead of observing labeled source data, however, in this subsection we instead observe $\kappa^* n$ instances from $p(x|Z = 1)$, together with their labels Y, matching the setup of Sec. 4.2.3.

A 2 hidden layer neural network with 512 neurons was trained to predict $e(x) = P(Y = 1|X = x, Z = 1)$. We chose $A = \{x : \widehat{e}(x) > 0.6\}$ [6] and $\alpha(x)$ according to Eqn. (11). The above procedure was repeated 10 times with different random seeds. The results are shown in Table 5.

7. Conclusion

This work introduces a more general identifiability condition than irreducibility for mixture proportion estimation. We also propose a subsampling-based framework that achieves bias reduction/elimination for baseline MPE algorithms. Theoretically, our work expands the scope of settings where MPE can be solved. Practically, we illustrate three scenarios where irreducibility fails, and our meta-algorithm successfully improves upon baseline MPE methods.

[6] A needs to be a subset of $\text{supp}(H)$. Note that in population level, $h(x) = p(x|Y = 1) \propto e(x) \cdot p(x|Z = 1)$, thus $\{x : e(x) > 0.6\} \subseteq \{x : e(x) > 0\} = \{x : h(x) > 0\}$. (The last equality holds because $e(x) = P(Y = 1|X = x, Z = 1)$.) Here we replace $e(x)$ with $\widehat{e}(x)$.

Acknowledgements

This work was supported in part by the National Science Foundation under award 2008074 and the Department of Defense, Defense Threat Reduction Agency under award HDTRA1-20-2-0002.

References

Alamaniotis, M., Mattingly, J., and Tsoukalas, L. H. Kernel-based machine learning for background estimation of nai low-count gamma-ray spectra. *IEEE Transactions on Nuclear Science*, 60(3):2209–2221, 2013.

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

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

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

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

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

Blanchard, G., Flaska, M., Handy, G., Pozzi, S., Scott, C., et al. Classification with asymmetric label noise:

Consistency and maximal denoising. *Electronic Journal of Statistics*, 10(2):2780–2824, 2016.

Cai, T. T. and Wei, H. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

Cowan, G. *Statistical data analysis*. Oxford university press, 1998.

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

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

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

Fei, H., Kim, Y., Sahu, S., Naphade, M., Mamidipalli, S. K., and Hutchinson, J. Heat pump detection from coarse grained smart meter data with positive and unlabeled learning. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1330–1338, 2013.

Garg, S., Wu, Y., Smola, A. J., Balakrishnan, S., and Lipton, Z. Mixture proportion estimation and PU learning: A modern approach. *Advances in Neural Information Processing Systems*, 34:8532–8544, 2021.

Gong, C., Wang, Q., Liu, T., Han, B., You, J., Yang, J., and Tao, D. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (8):4163–4177, 2021.

González, P., Castaño, A., Chawla, N. V., and Coz, J. J. D. A review on quantification learning. *ACM Computing Surveys (CSUR)*, 50(5):1–40, 2017.

Gorber, S. C., Schofield-Hurwitz, S., Hardt, J., Levasseur, G., and Tremblay, M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine & tobacco research*, 11(1):12–24, 2009.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4): 5, 2009.

Heckman, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

Ivanov, D. DEDPUL: Difference-of-estimated-densities-based positive-unlabeled learning. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 782–790. IEEE, 2020.

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

Kato, M., Teshima, T., and Honda, J. Learning from positive and unlabeled data with a selection bias. In *International conference on learning representations*, 2018.

Kiryo, R., Niu, G., Du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30, 2017.

Knoll, G. F. *Radiation detection and measurement*. John Wiley & Sons, 2010.

Lawrence, N. and Schölkopf, B. Estimating a kernel Fisher discriminant in the presence of label noise. In *18th International Conference on Machine Learning (ICML 2001)*, pp. 306–306. Morgan Kaufmann, 2001.

Li, F., Gu, Z., Ge, L., Li, H., Tang, X., Lang, X., and Hu, B. Review of recent gamma spectrum unfolding algorithms and their application. *Results in Physics*, 13:102211, 2019.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Maity, S., Dutta, D., Terhorst, J., Sun, Y., and Banerjee, M. A linear adjustment based approach to posterior drift in transfer learning. *Biometrika*, 2023. URL `https://doi.org/10.1093/biomet/asad029`.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

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

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

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

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

Scott, C. A generalized Neyman-Pearson criterion for optimal domain adaptation. In *Algorithmic Learning Theory*, pp. 738–761. PMLR, 2019.

Shanmugam, D. and Pierson, E. Quantifying inequality in underreported medical conditions. *arXiv preprint arXiv:2110.04133*, 2021.

Storkey, A. et al. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.

Werner, C. J., Bull, J., Solomon, C., Brown, F., McKinney, G., Rising, M., Dixon, D., Martz, R., Hughes, H., Cox, L., et al. MCNP 6.2 release notes. *Los Alamos National Laboratory*, 2018.

Yao, Y., Liu, T., Han, B., Gong, M., Niu, G., Sugiyama, M., and Tao, D. Rethinking class-prior estimation for positive-unlabeled learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=aYAA-XHKyk`.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pp. 819–827. PMLR, 2013.

Zhang, Q. and Goldman, S. EM-DD: An improved multiple-instance learning technique. *Advances in neural information processing systems*, 14, 2001.

A. Proofs

A.1. Proof of Proposition 2.4

Proposition. *Under the latent label model,*

$$\operatorname*{ess\,sup}_x P(Y=1|X=x) = \frac{\kappa^*}{\kappa(F|H)}\,,$$

where $0/0 := 0$.

Proof. First consider the trivial case when $\kappa(F|H) = 0$, then from the fact that $\kappa^* \leq \kappa(F|H)$, κ^* must also be 0. According to Eqn. (5), $P(Y=1|X=x) = 0\ \forall x$, therefore the equality holds.

Then for $\kappa(F|H) > 0$, we consider the cases $\kappa^* > 0$ and $\kappa^* = 0$ separately. For the first case, consider two subcases: $h(x) > 0$ and $h(x) = 0$. In the first subcase, we have that $f(x) > 0$, and therefore by the definition of conditional probability,

$$P(Y=1|X=x) = \frac{\kappa^*}{\frac{f(x)}{h(x)}}.$$

Taking the essential supremum over all x with $h(x) > 0$,

$$\begin{aligned}\operatorname*{ess\,sup}_{x:h(x)>0} P(Y=1|X=x) &= \frac{\kappa^*}{\operatorname*{ess\,inf}_{x:h(x)>0} \frac{f(x)}{h(x)}} \\ &= \frac{\kappa^*}{\kappa(F|H)},\end{aligned}$$

where the second equality follows from Proposition 2.2. In the second subcase, $P(Y=1|X=x)$ is zero. Therefore,

$$\begin{aligned}\operatorname*{ess\,sup}_x P(Y=1|X=x) &= \operatorname*{ess\,sup}_{x:h(x)>0} P(Y=1|X=x) \\ &= \frac{\kappa^*}{\kappa(F|H)}.\end{aligned}$$

When $\kappa^* = 0$, $P(Y=1|X=x) = 0$ for all x, and the equality still holds. □

A.2. Proof of Theorem 3.1

Theorem (Identifiability Under Local Supremal Posterior (LSP))**.** *Let A be any non-empty measurable subset of $E_H = \{x : h(x) > 0\}$ and $s = \operatorname*{ess\,sup}_{x\in A} P(Y=1|X=x)$, then*

$$\kappa^* = s \cdot \inf_{S\subseteq A} \frac{F(S)}{H(S)} = s \cdot \operatorname*{ess\,inf}_{x\in A} \frac{f(x)}{h(x)}.$$

Proof. Consider the case of $\kappa^* > 0$ and $\kappa^* = 0$ separately.

If $\kappa^* > 0$, then $E_H = \{x : h(x) > 0\} \subseteq E_F = \{x : f(x) > 0\}$. Recall from Eqn. (5),

$$P(Y=1|X=x) = \kappa^* \cdot \frac{h(x)}{f(x)} \quad \text{when } f(x) > 0.$$

Taking the essential supremum over A and recall the definition of s,

$$s = \operatorname*{ess\,sup}_{x\in A} P(Y=1|X=x) = \kappa^* \cdot \operatorname*{ess\,sup}_{x\in A} \frac{h(x)}{f(x)}.$$

Since $A \subseteq E_H \subseteq E_F$, $f(x)$ and $h(x)$ are both positive for $x \in A$. Rearrange the denominator

$$s = \kappa^* \cdot \frac{1}{\operatorname*{ess\,inf}_{x \in A} \frac{f(x)}{h(x)}},$$

take the denominator to the other side, we get

$$\kappa^* = s \cdot \operatorname*{ess\,inf}_{x \in A} \frac{f(x)}{h(x)} = s \cdot \inf_{S \subseteq A} \frac{F(S)}{H(S)},$$

where the second equality follows from the identity that $\operatorname*{ess\,inf}_{x \in A} \frac{f(x)}{h(x)} = \inf_{S \subseteq A} \frac{F(S)}{H(S)}$.

If $\kappa^* = 0$, then $s = 0$, the above equality still holds. □

A.3. Proof of Theorem 3.2

Theorem (Identifiability Under Tight Posterior Upper Bound)**.** *Consider any non-empty measurable set $A \subseteq E_H = \{x : h(x) > 0\}$, and let $s = \operatorname*{ess\,sup}_{x \in A} P(Y = 1|X = x)$. Let $\alpha(x)$ be any measurable function satisfying*

$$\alpha(x) \in \begin{cases} [P(Y = 1|X = x), s], & x \in A, \\ [P(Y = 1|X = x), 1], & \textit{o.w.} \end{cases} \tag{13}$$

Define a new distribution $\widetilde{F}$ in terms of its density

$$\begin{aligned} \widetilde{f}(x) &= \frac{1}{c} \cdot \alpha(x) \cdot f(x), \\ \textit{where} \quad c &= \int \alpha(x) f(x) dx = \mathbb{E}_{X \sim F} \left[\alpha(x)\right]. \end{aligned} \tag{14}$$

Then

$$\kappa^* = c \cdot \kappa(\widetilde{F}|H).$$

Proof. Write $\kappa(\widetilde{F}|H)$ explicitly according to Proposition 2.2 and Eqn. (14),

$$\begin{aligned} c \cdot \kappa(\widetilde{F}|H) &= c \cdot \operatorname*{ess\,inf}_{x: h(x) > 0} \frac{\widetilde{f}(x)}{h(x)} \quad \text{by definition of } \kappa(\widetilde{F}|H) \\ &= c \cdot \operatorname*{ess\,inf}_{x: h(x) > 0} \frac{\frac{1}{c} \cdot \alpha(x) \cdot f(x)}{h(x)} \quad \text{plug in the expression of } \widetilde{f}(x) \\ &= \operatorname*{ess\,inf}_{x: h(x) > 0} \frac{\alpha(x) \cdot f(x)}{h(x)}. \end{aligned}$$

We will show that it equals to κ^* by proving it is both upper and lower bound of κ^*.

From Eqn. (13), we can conclude that $\alpha(x) \geq P(Y = 1|X = x) \ \forall x$, then

$$\begin{aligned} c \cdot \kappa(\widetilde{F}|H) &\geq \operatorname*{ess\,inf}_{x: h(x) > 0} \frac{P(Y = 1|X = x) \cdot f(x)}{h(x)} \\ &= \operatorname*{ess\,inf}_{x: h(x) > 0} \kappa^* \quad \text{rearrange Eqn. (5)} \\ &= \kappa^*. \end{aligned}$$

Meanwhile, consider $x \in A$

$$\begin{aligned} c \cdot \kappa(\widetilde{F}|H) &= \operatorname*{ess\,inf}_{x:h(x)>0} \frac{\alpha(x) \cdot f(x)}{h(x)} \\ &\leq \operatorname*{ess\,inf}_{x \in A} \frac{\alpha(x) \cdot f(x)}{h(x)} \quad \text{replace } E_H \text{ by } A \\ &\leq \operatorname*{ess\,inf}_{x \in A} \frac{s \cdot f(x)}{h(x)} \quad \text{because } \alpha(x) \leq s, \forall x \in A \\ &= \kappa^* \quad \text{by Theorem 3.1} \end{aligned}$$

This shows that $c \cdot \kappa(\widetilde{F}|H) = \kappa^*$. □

A.4. Proof of Corollary 3.3

Corollary. *Let $\alpha(x)$ be any measurable function with*

$$\alpha(x) \in [P(Y = 1|X = x), 1] \quad \forall x.$$

Define a new distribution $\widetilde{F}$ in terms of its density $\widetilde{f}$, obtained by Eqn. (7)*. Then*

$$\kappa^* \leq c \cdot \kappa(\widetilde{F}|H) \leq \kappa(F|H).$$

Proof. From the proof of Theorem 3.2, we know that

$$c \cdot \kappa(\widetilde{F}|H) = \operatorname*{ess\,inf}_{x:h(x)>0} \frac{\alpha(x) \cdot f(x)}{h(x)}$$

From the fact that $\alpha(x) \geq P(Y = 1|X = x)\ \forall x$, we have

$$\begin{aligned} c \cdot \kappa(\widetilde{F}|H) &\geq \operatorname*{ess\,inf}_{x:h(x)>0} \frac{P(Y = 1|X = x) \cdot f(x)}{h(x)} \\ &= \operatorname*{ess\,inf}_{x:h(x)>0} \kappa^* \\ &= \kappa^*. \end{aligned}$$

What's more, since $\alpha \leq 1\ \forall x$,

$$c \cdot \kappa(\widetilde{F}|H) \leq \operatorname*{ess\,inf}_{x:h(x)>0} \frac{1 \cdot f(x)}{h(x)} = \kappa(F|H).$$

□

A.5. Proof of Theorem 4.1

We now establish a rate of convergence result for estimator $\widehat{\kappa}^* = \widehat{c} \cdot \widehat{\kappa}(\widetilde{F}|H)$ from Algorithm 1.

Theorem. *Assume $\alpha(x)$ satisfies the condition in Eqn.* (6)*. After subsampling, assume the resulting $\widetilde{F}$ and H are such that $\operatorname*{arg\,min}_{S \in \mathcal{A}: H(S)>0} \frac{\widetilde{F}(S)}{H(S)}$ exists. Then there exists a constant $C > 0$ and an existing estimator $\widehat{\kappa}$ such that for m and n sufficiently large, the estimator $\widehat{\kappa}^*$ from Algorithm 1 satisfies*

$$\Pr\left(|\widehat{\kappa}^* - \kappa^*| \leq C\left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \geq 1 - \mathcal{O}\left(\frac{1}{m} + \frac{1}{n}\right).$$

Proof. Recall the setup, we originally have i.i.d. sample [7]

$$\begin{aligned} X_F &:= \{X_1, X_2, \cdots, X_n\} \overset{iid}{\sim} F, \\ X_H &:= \{X_{n+1}, X_{n+2}, \cdots, X_{n+m}\} \overset{iid}{\sim} H. \end{aligned}$$

[7] The notation here is a bit different from Eqn. (1) in that the index of X_i is changed. This allows for more concise notation in the following derivation.

After rejection sampling, we obtain n' i.i.d. sample $X_{\widetilde{F}} \sim \widetilde{F}$ (where $\mathbb{E}[n'] = c \cdot n$ and $c = \int \alpha(x) f(x) dx$), from which we can estimate $\kappa(\widetilde{F}|H)$. Under the assumption that $\underset{S \in \mathcal{A}: H(S)>0}{\arg\min} \frac{\widetilde{F}(S)}{H(S)}$ exists, estimator $\widehat{\kappa}(\widetilde{F}|H)$ by Scott (2015) has rate of convergence

$$\Pr\left(\left|\widehat{\kappa}(\widetilde{F}|H) - \kappa(\widetilde{F}|H)\right| \leq C_1 \left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n'}{n'}}\right]\right) \geq 1 - \frac{2}{m} - \frac{2}{n'}. \tag{15}$$

Now n' is a random variable here, and we want to establish a rate of convergence result involving n. This can be done by applying a concentration inequality for n'.

Theorem. *(Hoeffding's Inequality) Let $Z_1, \cdots Z_n$ be independent random variables on $\mathbb{R}$ that take values in $[0, 1]$. Denote $Z = \frac{1}{n}\sum_{i=1}^n Z_i$, then for all $\epsilon > 0$,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z]\right| \leq \epsilon\right) \geq 1 - 2\exp(-2n\epsilon^2).$$

Let $\delta = 2\exp(-2n\epsilon^2)$, the theorem can be restated as:

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z]\right| \leq \sqrt{\frac{\log(1/\delta)}{2n}}\right) \geq 1 - \delta.$$

Take $Z_i = \mathbb{1}_{\{V_i \leq \alpha(X_i)\}}$, where $i \in \{1, 2, \cdots, n\}$, V_i denotes the i-th independent draw from Uniform$(0, 1)$ [8] and X_i denote the i-th draw from F. Then

$$n' = \sum_{i=1}^n Z_i,$$

$$\mathbb{E}[Z] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i] = \mathbb{E}_{(X_i,V_i)}[Z_i] = \mathbb{E}_{X_i}\left[\mathbb{E}_{V_i|X_i}\left[\mathbb{1}_{\{V_i \leq \alpha(X_i)\}}\right]\right] = \mathbb{E}_{X_i}[\alpha(X_i)] = \int \alpha(x) f(x) dx = c.$$

Plug into Hoeffding's Inequality and setting $\epsilon = \frac{c}{2}$, we have

$$\Pr\left(|n' - c \cdot n| \leq \frac{c}{2} \cdot n\right) \geq 1 - 2\exp\left(-\frac{c}{2}n\right), \tag{16}$$

which allows us to bound n' by a constant times n.

Now we can establish a rate of convergence result of $\widehat{\kappa}(\widetilde{F}|H)$ w.r.t. n

$$\begin{aligned}
&\Pr\left(\left|\widehat{\kappa}(\widetilde{F}|H) - \kappa(\widetilde{F}|H)\right| \leq C_2 \left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \\
\geq\ &\Pr\left(\left(\left|\widehat{\kappa}(\widetilde{F}|H) - \kappa(\widetilde{F}|H)\right| \leq C_2 \left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \text{and} \left(\frac{c}{2} \cdot n \leq n' \leq \frac{3c}{2} \cdot n\right)\right) \quad \because \Pr(A) \geq \Pr(A \cap B) \\
=\ &\Pr\left(\left(\left|\widehat{\kappa}(\widetilde{F}|H) - \kappa(\widetilde{F}|H)\right| \leq C_2 \left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \middle| \left(\frac{c}{2} \cdot n \leq n' \leq \frac{3c}{2} \cdot n\right)\right) \cdot \Pr\left(\frac{c}{2} \cdot n \leq n' \leq \frac{3c}{2} \cdot n\right) \\
\geq\ &\left(1 - \mathcal{O}\left(\frac{1}{m} + \frac{1}{n}\right)\right) \cdot \left(1 - 2\exp\left(-\frac{c}{2}n\right)\right) \quad \because n \text{ and } n' \text{ can be used interchangeably in the first probability term} \\
\geq\ &1 - \mathcal{O}\left(\frac{1}{m} + \frac{1}{n}\right) - 2\exp\left(-\frac{c}{2}n\right) \quad \because (1-a)(1-b) \geq 1 - a - b \\
\geq\ &1 - \mathcal{O}\left(\frac{1}{m} + \frac{1}{n}\right) \quad \because \exp(-n) \text{ decays faster}
\end{aligned}$$

[8] V_i is used in rejection sampling (Algorihm 4).

As for the estimator of $c = \mathbb{E}_{X\sim F}[\alpha(x)]$, the rate of convergence can also be shown via Hoeffding's Inequality,

$$\widehat{c} = \frac{n'}{n} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad c = \mathbb{E}[Z].$$

Plug into Hoeffding's Inequality and let the confidence $\delta = 1/n$, we have

$$\Pr\left(|\widehat{c} - c| \le \sqrt{\frac{\log n}{2n}}\right) \ge 1 - 1/n. \tag{17}$$

Note that by triangle inequality

$$\begin{aligned}
\left|\widehat{c}\cdot\widehat{\kappa}(\widetilde{F}|H) - c\cdot\kappa(\widetilde{F}|H)\right| &= \left|\widehat{c}\cdot\widehat{\kappa}(\widetilde{F}|H) - \widehat{c}\cdot\kappa(\widetilde{F}|H) + \widehat{c}\cdot\kappa(\widetilde{F}|H) - c\cdot\kappa(\widetilde{F}|H)\right| \\
&\le \left|\widehat{c}\cdot\widehat{\kappa}(\widetilde{F}|H) - \widehat{c}\cdot\kappa(\widetilde{F}|H)\right| + \left|\widehat{c}\cdot\kappa(\widetilde{F}|H) - c\cdot\kappa(\widetilde{F}|H)\right| \\
&= |\widehat{c}|\cdot\left|\widehat{\kappa}(\widetilde{F}|H) - \kappa(\widetilde{F}|H)\right| + \left|\kappa(\widetilde{F}|H)\right|\cdot|\widehat{c} - c| \\
&\le \left|\widehat{\kappa}(\widetilde{F}|H) - \kappa(\widetilde{F}|H)\right| + |\widehat{c} - c|.
\end{aligned}$$

Finally, combine all previous results

$$\begin{aligned}
&\Pr\left(|\widehat{\kappa}^* - \kappa^*| \le C\left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \\
=\ &\Pr\left(\left|\widehat{c}\cdot\widehat{\kappa}(\widetilde{F}|H) - c\cdot\kappa(\widetilde{F}|H)\right| \le C\left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \\
\ge\ &\Pr\left(\left|\widehat{\kappa}(\widetilde{F}|H) - \kappa(\widetilde{F}|H)\right| + |\widehat{c} - c| \le C\left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \\
\ge\ &\Pr\left(\left(\left|\widehat{\kappa}(\widetilde{F}|H) - \kappa(\widetilde{F}|H)\right| \le \frac{C}{2}\left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right) \text{and} \left(|\widehat{c} - c| \le \frac{C}{2}\left[\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\right)\right) \\
\ge\ &1 - \mathcal{O}\left(\frac{1}{m} + \frac{1}{n}\right),
\end{aligned}$$

then we can conclude that $\widehat{\kappa}^* \to \kappa^*$ with rate of convergence $O\left(\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right)$.

□

A.6. Proof of Proposition 5.2

Proposition. *For $\kappa(F|H')$ obtained from ReMPE-2:*

$$\kappa(F|H') < \kappa(F|H).$$

Proof. From the fact that $B = \arg\min_{S\in\mathfrak{S}} \frac{F(S)}{H(S)}$, we have

$$\kappa(F|H) = \inf_{S\in\mathcal{A}: H(S)>0} \frac{F(S)}{H(S)} = \frac{F(B)}{H(B)}.$$

After regrouping, $H'(B) > H(B)$. Therefore,

$$\begin{aligned}\kappa(F|H') &= \inf_{S\in\mathcal{A}:H'(S)>0} \frac{F(S)}{H'(S)} \\ &= \frac{F(B)}{H'(B)} \\ &< \frac{F(B)}{H(B)} \\ &= \kappa(F|H).\end{aligned}$$

□

B. More about Subsampling MPE

B.1. Intuition

Theorem 3.2 and Corollary 3.3 have already justified the use of subsampling. Here, we explain in another perspective (in terms of distributions, similar to the analysis in Yao et al. (2022)). The idea is that, since the original G may violate irreducibility assumption, we modify G such that the resulting latent component distribution is less likely to violate the assumption.

Write the unknown distribution G itself as a mixture $G = \gamma G_1 + (1-\gamma)G_2,\ \gamma \geq 0$. Then F becomes:

$$\begin{aligned}F &= (1-\kappa^*)G + \kappa^* H \\ &= (1-\kappa^*)\left[\gamma G_1 + (1-\gamma)G_2\right] + \kappa^* H.\end{aligned}$$

Switch G_1 to the other side (i.e., discarding the probability mass from G_1)

$$F - (1-\kappa^*)\gamma G_1 = (1-\kappa^*)(1-\gamma)G_2 + \kappa^* H,$$

the left hand side can be re-written as

$$\underbrace{[1-(1-\kappa^*)\gamma]}_{:=c,\text{ normalizing const.}} \underbrace{\frac{1}{1-(1-\kappa^*)\gamma}\left[F - (1-\kappa^*)\gamma G_1\right]}_{:=\widetilde{F},\ \textbf{after subsampling}}.$$

Then the resulting distribution $\widetilde{F}$ is

$$\begin{aligned}\widetilde{F} &= \frac{(1-\kappa^*)(1-\gamma)}{c} G_2 + \frac{\kappa^*}{c} H \\ &=: (1-\tilde{\kappa}^*)\widetilde{G} + \tilde{\kappa}^* H.\end{aligned}$$

By discarding a portion of probability mass from G, which is done by subsampling in practice, the resulting latent component distribution $\widetilde{G}$ is less likely to violate the irreducibility assumption. The new mixture proportion $\tilde{\kappa}^* = \kappa^*/c$.

In the following, we provide justification of the above claim, which can also be seen as a reformulation of Corollary 3.3.

Proposition B.1. *Given some probability mass from G being dropped out, $c \cdot \kappa(\widetilde{F}|H)$ is always bounded by:*

$$\kappa^* \leq c \cdot \kappa(\widetilde{F}|H) \leq \kappa(F|H).$$

Furthermore, bias is strictly reduced when $(1-\gamma)\kappa(G_2|H) < \kappa(G|H)$.

Proof. Observe that

$$\begin{aligned}c \cdot \kappa(\widetilde{F}|H) &= \inf_{S\in\mathcal{A}:H(S)>0} \frac{c\cdot\widetilde{F}(S)}{H(S)} \\ &= \kappa^* + (1-\kappa^*) \inf_{S\in\mathcal{A}:H(S)>0} \frac{(1-\gamma)G_2(S)}{H(S)},\end{aligned}$$

therefore $c \cdot \kappa(\widetilde{F}|H) \geq \kappa^*$.

From the fact that $G = \gamma G_1 + (1-\gamma)G_2$, we have $G \geq (1-\gamma)G_2$. Then

$$\begin{aligned} c \cdot \kappa(\widetilde{F}|H) &\leq \kappa^* + (1-\kappa^*) \inf_{S \in \mathcal{A}: H(S)>0} \frac{G(S)}{H(S)} \\ &= \kappa(F|H). \end{aligned}$$

To have bias reduction, we need $c \cdot \kappa(\widetilde{F}|H) < \kappa(F|H)$, where both quantities can be represented as

$$\begin{aligned} c \cdot \kappa(\widetilde{F}|H) &= \kappa^* + (1-\kappa^*)(1-\gamma)\kappa(G_2|H) \\ \kappa(F|H) &= \kappa^* + (1-\kappa^*)\kappa(G|H). \end{aligned}$$

Then we can get the result of $(1-\gamma)\kappa(G_2|H) < \kappa(G|H)$ by direct comparison. □

Based on the above proof, we claim that $c \cdot \kappa(\widetilde{F}|H)$ leads to no worse estimation bias than $\kappa(F|H)$. To be specific, when G is irreducible w.r.t. H, then $\kappa^* = c \cdot \kappa(\widetilde{F}|H) = \kappa(F|H)$. When G is not irreducible w.r.t. H, then $\kappa^* \leq c \cdot \kappa(\widetilde{F}|H) \leq \kappa(F|H)$. The key difference compared to Yao et al. (2022)'s approach is that we are modifying G, thus equivalently subsampling F, rather than regrouping on H.

In summary, without making assumptions about irreducibility, as long as we remove some contribution from G in F by subsampling, the resulting identifiable quantity $c \cdot \kappa(\widetilde{F}|H)$ will be at least no worse than the maximum proportion $\kappa(F|H)$. Furthermore, with the knowledge of set A and $\operatorname{ess\,sup}_{x \in A} P(Y=1|X=x)$, $c \cdot \kappa(\widetilde{F}|H)$ will equal to κ^*.

B.2. Rejection Sampling

Rejection sampling is a Monte Carlo method that aims to generate sample following a new distribution $\widetilde{Q}$ based on sample from distribution Q, which are characterised by the densities $\widetilde{q}$ and q. An instance x drawn from $q(x)$ is kept with acceptance probability $\beta(x) \in [0,1]$, and rejected otherwise. Algorithm 4 shows the detailed procedure.

Algorithm 4 Rejection Sampling

Input:
  sample $[x_i] = \{x_1, \cdots, x_n\} \sim q(x)$,
  acceptance function $\beta(x)$
Output:
  sample $[z_j] \sim \widetilde{q}(x) = \frac{1}{c} \cdot \beta(x) \cdot q(x)$,
    where $c = \int \beta(x) q(x) dx = \mathbb{E}_{X \sim q}[\beta(X)]$
Initialize $j \leftarrow 1$
for $i = 1$ **to** n **do**
  $v \sim \text{Uniform}(0,1)$
  if $v \leq \beta(x)$ **then**
    $z_j \leftarrow x_i$
    $j \leftarrow j + 1$
  end if
end for
return $[z_j]$

B.3. Gamma Spectrum Unfolding

In gamma spectrum unfolding, a gamma ray detector measures the energies of incoming gamma ray particles. The gamma rays were emitted either by a source of interest or from the background. The measurement is represented as a histogram $f(x)$, where the bins correspond to a quantization of energy. The histogram $h(x)$ of measurements from the source of interest is also known.

The goal is to obtain a lower bound $\rho(x)$ of the quantity $(1-\kappa^*)g(x)$ on a certain set $A \subset \text{supp}(H)$. We specify A to be the energy bins near the main peak of $h(x)$ (aka, full-energy peak). For $x \in \text{supp}(F) \setminus \text{supp}(H)$, $\rho(x) = f(x)$ because the gamma rays must come from the background in these regions. Typically, $\text{supp}(F) \setminus \text{supp}(H)$ contains two (or more) intervals, therefore we know the value of $\rho(x)$ on either side of set A. Then $\rho(x)\ \forall x \in A$ can be estimated using linear interpolation (Knoll, 2010; Alamaniotis et al., 2013). The above procedure is illustrated in Figure 1 and the acceptance function is chosen to be

$$\alpha(x) = \begin{cases} 1 - \frac{\rho(x)}{f(x)}, & x \in A, \\ 1, & \text{o.w.} \end{cases} \tag{18}$$

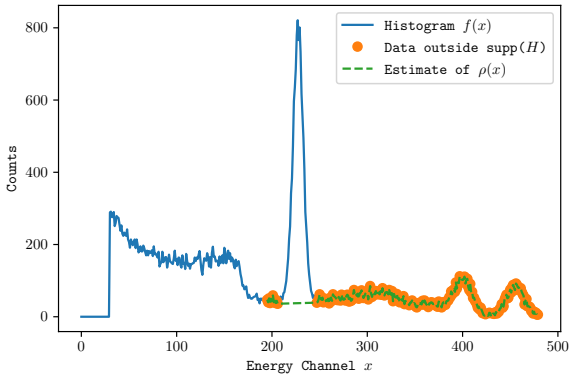

Figure 1. Visual illustration of how to estimate $\rho(x)$

C. Detailed Experimental Result in Sec. 6

This section shows four tables corresponding to four experimental setups in Sec. 6.

Table 2. Average absolute estimation error on gamma ray spectra data, corresponding to the unfolding scenario. Several state-of-the-art MPE algorithms DPL (Ivanov, 2020), EN (Elkan & Noto, 2008), KM (Ramaswamy et al., 2016) and TIcE (Bekker & Davis, 2018) are selected. The mean absolute error (avg$[|\widehat{\kappa}^* - \kappa^*|]$) is reported, the smallest error among original, regrouping and subsampling versions is bolded. $(+/-)$ denotes that on average the estimator produces positive/negative estimation bias (sgn[avg$(\widehat{\kappa}^* - \kappa^*)$]).

κ^*	DPL	ReDPL	SuDPL	EN	ReEN	SuEN	KM	ReKM	SuKM	TIcE	ReTIcE	SuTIcE
0.1	0.063+	0.053+	**0.008+**	0.057+	0.048+	**0.012−**	0.226+	0.146+	**0.036+**	0.129+	0.115+	**0.019+**
0.25	0.055+	**0.008+**	0.011+	0.040+	**0.004−**	0.017−	0.216+	**0.020+**	0.059+	0.112+	0.065+	**0.017+**
0.5	0.029+	0.123−	**0.019−**	**0.017+**	0.129−	0.032−	0.130+	**0.043+**	0.049+	0.082+	**0.018+**	0.019+
0.75	0.034+	0.284−	**0.015+**	**0.022−**	0.289−	0.047−	0.078+	0.094−	**0.022+**	0.055+	0.045−	**0.021+**
average	0.045+	0.117−	**0.013+**	0.034+	0.118−	**0.027−**	0.163+	0.076+	**0.042+**	0.095+	0.061+	**0.019+**

Table 3. Average absolute estimation error on synthetic data, corresponding to domain adaptation scenario. Several state-of-the-art MPE algorithms DPL (Ivanov, 2020), EN (Elkan & Noto, 2008), KM (Ramaswamy et al., 2016) and TIcE (Bekker & Davis, 2018) are selected. The mean absolute error (avg$[|\widehat{\kappa}^* - \kappa^*|]$) is reported, and the smallest error among the original, regrouping and subsampling versions is bolded. $(+/-)$ denotes that on average the estimator produces positive/negative estimation bias (sgn[avg$(\widehat{\kappa}^* - \kappa^*)$]).

κ^*	DPL	ReDPL	SuDPL	EN	ReEN	SuEN	KM	ReKM	SuKM	TIcE	ReTIcE	SuTIcE
0.1	0.089+	0.075+	**0.014−**	0.059+	0.051+	**0.027−**	0.102+	0.091+	**0.013−**	0.150+	0.140+	**0.038+**
0.25	0.083+	0.060+	**0.016−**	0.037+	**0.016+**	0.051−	0.081+	0.059+	**0.018−**	0.137+	0.108+	**0.040+**
0.5	0.053+	0.028+	**0.024−**	**0.022−**	0.057−	0.077−	0.052+	0.037+	**0.020−**	0.114+	0.068+	**0.035+**
0.75	**0.016−**	0.050−	0.056−	**0.063−**	0.118−	0.114−	**0.018+**	0.050−	0.036−	0.112+	0.058+	**0.051+**
average	0.060+	0.053+	**0.028−**	**0.045−**	0.061−	0.067−	0.063+	0.059+	**0.022−**	0.128+	0.094+	**0.041+**

Table 4. Average absolute estimation error on benchmark data, corresponding to domain adaptation scenario. Several state-of-the-art MPE algorithms DPL (Ivanov, 2020), EN (Elkan & Noto, 2008), KM (Ramaswamy et al., 2016) and TIcE (Bekker & Davis, 2018) are selected. The mean absolute error is reported, and the smallest error among the original, regrouping and subsampling versions is bolded. $+/-/\cdot$ denotes that on average the estimator produces positive/negative/no estimation bias.

Dataset	κ^*	DPL	ReDPL	SuDPL	EN	ReEN	SuEN	KM	ReKM	SuKM	TIcE	ReTIcE	SuTIcE
Mushroom	0.1	0.075+	0.069+	**0.053+**	0.121+	0.105+	**0.099+**	0.061+	0.054+	**0.039+**	0.082+	0.086+	**0.064+**
	0.25	0.063+	**0.023+**	0.040+	0.139+	**0.101+**	0.109+	0.053+	0.041+	**0.024+**	0.057+	0.034+	**0.024+**
	0.5	0.035+	0.084−	**0.025+**	0.132+	**0.074+**	0.112+	**0.057−**	0.189−	0.063−	**0.026+**	0.058−	0.028−
	0.75	0.015−	0.149−	**0.013−**	0.097+	**0.033+**	0.082+	**0.096−**	0.253−	0.108−	0.076−	0.134−	**0.027−**
	avg	0.047+	0.081−	**0.033+**	0.122+	**0.078+**	0.101+	0.067−	0.134−	**0.059−**	0.060+	0.078−	**0.036+**
Landsat	0.1	0.066+	0.053+	**0.014+**	0.168+	0.139+	**0.115+**	0.065+	0.047+	**0.017+**	0.070+	0.074+	**0.031+**
	0.25	0.060+	0.010−	**0.005+**	0.161+	0.119+	**0.100+**	0.053+	0.027+	**0.007+**	0.034+	0.028+	**0.012−**
	0.5	0.037+	0.034−	**0.010−**	0.131+	0.097+	**0.084+**	0.037+	0.018+	**0.016−**	0.033+	0.031−	**0.030−**
	0.75	**0.022+**	0.071−	0.037−	0.102+	0.085+	**0.041+**	0.028+	**0.023−**	0.024−	**0.036−**	0.042−	0.067−
	avg	0.046+	0.042−	**0.017−**	0.141+	0.110+	**0.085+**	0.046+	0.029+	**0.016−**	0.043+	0.044−	**0.035−**
Shuttle	0.1	0.055+	0.048+	**0.006+**	0.105+	0.096+	**0.053+**	0.044+	0.033+	**0.006−**	0.083+	0.072+	**0.023+**
	0.25	0.045+	0.010−	**0.007−**	0.098+	0.082+	**0.038+**	0.027+	0.018−	**0.016−**	0.074+	0.030+	**0.020+**
	0.5	0.025+	0.102−	**0.016−**	0.093+	0.066+	**0.050+**	**0.017−**	0.099−	0.030−	0.074+	0.065−	**0.041+**
	0.75	**0.022−**	0.393−	0.033−	0.063+	**0.041+**	0.044+	**0.058−**	0.294−	0.075−	0.089+	0.179−	**0.068+**
	avg	0.037+	0.138−	**0.015−**	0.090+	0.071+	**0.046+**	0.036+	0.111−	**0.032−**	0.080+	0.086−	**0.038+**
MNIST17	0.1	0.076+	0.078+	**0.035+**	0.184+	0.157+	**0.116+**	0.065+	0.057+	**0.025+**	0.101+	0.093+	**0.052+**
	0.25	0.058+	0.047+	**0.008−**	0.218+	0.175+	**0.138+**	0.053+	0.017+	**0.009−**	0.098+	0.084+	**0.031+**
	0.5	0.032+	0.047−	**0.022−**	0.275+	**0.180+**	0.191+	0.029+	0.030−	**0.017−**	0.080+	**0.021+**	0.047+
	0.75	**0.023−**	0.169−	0.046−	0.250+	**0.189+**	0.217+	**0.016+**	0.146−	**0.016−**	0.081+	0.094−	**0.060+**
	avg	0.047+	0.085−	**0.028−**	0.231+	0.175+	**0.166+**	0.041+	0.063−	**0.017−**	0.090+	0.073−	**0.048+**
Overall	avg	0.044+	0.087−	**0.023−**	0.146+	0.109+	**0.099+**	0.047+	0.084−	**0.031−**	0.068+	0.070−	**0.039+**

Table 5. Average absolute estimation error on benchmark data, corresponding to the selected/reported at random scenario. Several state-of-the-art MPE algorithms DPL (Ivanov, 2020), EN (Elkan & Noto, 2008), KM (Ramaswamy et al., 2016) and TIcE (Bekker & Davis, 2018) are selected. The mean absolute error is reported, the smallest error among the original, regrouping and subsampling versions is bolded. $+/-/\cdot$ denotes that on average the estimator produces positive/negative/no estimation bias.

Dataset	κ^*	DPL	ReDPL	SuDPL	EN	ReEN	SuEN	KM	ReKM	SuKM	TIcE	ReTIcE	SuTIcE
Mushroom	0.1	0.073+	0.067+	**0.058+**	0.119+	0.105+	**0.099+**	0.059+	0.051+	**0.040+**	0.084+	0.082+	**0.057+**
	0.25	0.060+	0.023+	**0.005−**	0.134+	0.096+	**0.054+**	0.054+	0.034+	**0.010−**	0.049+	0.028+	**0.008+**
	0.5	0.032+	0.087−	**0.018−**	0.125+	**0.068+**	0.076+	**0.073−**	0.190−	0.082−	**0.027+**	0.057−	0.052−
	0.75	**0.023−**	0.203−	0.028−	0.096+	**0.030+**	0.066+	**0.100−**	0.262−	0.123−	**0.028−**	0.096−	0.058−
	avg	0.047+	0.095−	**0.027+**	0.119+	0.075+	**0.074+**	0.072−	0.134−	**0.064−**	0.047+	0.066−	**0.044−**
Landsat	0.1	0.066+	0.053+	**0.034+**	0.167+	0.141+	**0.121+**	0.066+	0.049+	**0.028+**	0.068+	0.074+	**0.036+**
	0.25	0.052+	0.011−	**0.007·**	0.159+	0.115+	**0.085+**	0.061+	0.035+	**0.011+**	0.034+	0.020+	**0.011−**
	0.5	0.043+	0.065−	**0.011−**	0.137+	0.103+	**0.089+**	0.033+	0.020+	**0.016−**	0.040+	**0.029−**	0.048−
	0.75	0.032+	0.097−	**0.023−**	0.103+	**0.071+**	**0.071+**	0.023+	0.027−	**0.017·**	**0.041−**	0.087−	0.068−
	avg	0.048+	0.057−	**0.019−**	0.142+	0.108+	**0.092+**	0.046+	0.033+	**0.018+**	0.046+	0.053−	**0.041−**
Shuttle	0.1	0.056+	0.048+	**0.015+**	0.112+	0.097+	**0.060+**	0.049+	0.035+	**0.016−**	0.080+	0.066+	**0.032+**
	0.25	0.044+	0.018−	**0.007−**	0.106+	0.083+	**0.045+**	0.029+	0.020−	**0.018−**	0.077+	0.024+	**0.037+**
	0.5	0.026+	0.162−	**0.008−**	0.098+	0.081+	**0.068+**	**0.016−**	0.165−	0.051−	0.083+	0.099−	**0.052+**
	0.75	**0.012+**	0.348−	0.023−	0.065+	0.029+	**0.050+**	**0.072−**	0.296−	0.089−	0.075+	0.169−	**0.078+**
	avg	0.035+	0.144−	**0.013−**	0.095+	0.073+	**0.056+**	**0.042+**	0.129−	0.044−	0.079+	0.089−	**0.050+**
MNIST17	0.1	0.077+	0.078+	**0.055+**	0.183+	0.157+	**0.134+**	0.060+	0.052+	**0.046+**	0.096+	0.083+	**0.076+**
	0.25	0.063+	0.052+	**0.005+**	0.229+	0.179+	**0.113+**	0.060+	0.027+	**0.009−**	0.099+	0.087+	**0.034+**
	0.5	0.035+	0.042−	**0.014+**	0.300+	**0.198+**	0.208+	0.022+	0.024−	**0.011−**	0.099+	**0.028+**	0.053+
	0.75	**0.014−**	0.178−	0.033−	0.248+	**0.161+**	0.200+	**0.008+**	0.153−	0.021−	0.090+	0.095−	**0.066+**
	avg	0.047+	0.088−	**0.027+**	0.240+	0.173+	**0.164+**	0.038+	0.064−	**0.022+**	0.096+	0.073+	**0.057+**
Overall	avg	0.044+	0.096−	**0.022−**	0.149+	0.107+	**0.097+**	0.050+	0.090−	**0.037−**	0.067+	0.070+	**0.048+**

D. When Irreducibility Holds

In theory, when irreducibility holds, baseline MPE methods shall be asymptotically unbiased estimators of the mixture proportion κ^*, regrouping may introduce negative bias, subsampling should not introduce bias. Here we run some synthetically generated experiments (in a controlled setting) to verify the theoretical claim.

We specify the distributions H, G and F as:

$$H \sim \mathcal{N}(\mu_1 = 0, \sigma_1 = 1)$$
$$G \sim \mathcal{N}(\mu_2 = 2, \sigma_2 = 1)$$
$$F = (1 - \kappa^*)G + \kappa^* H.$$

where G is irreducible w.r.t. H. We draw $m = 500, n = 1500$ instances from both H and F. The true mixture proportion $\kappa^* = \kappa(F|H)$ is varied in $\{0.1, 0.25, 0.5, 0.75\}$.

We draw 2000 labeled instances from F. A 1 hidden layer neural network with 16 neurons was trained to predict $P(Y = 1|X = x)$. The acceptance function $\alpha(x)$ used in Subsampling MPE is chosen to be

$$\alpha(x) = \begin{cases} \widehat{P}(Y = 1|X = x), & x \in A, \\ 1, & \text{o.w.}, \end{cases}$$

where $A = \{x : \widehat{P}(Y = 1|X = x) > 0.6\}$. As for ReMPE, 10% of the sample from F is chosen to be regrouped to the sample from H, as suggested by Yao et al. (2022). The above procedure was repeated 10 times with different random seeds. Results where shown in Table 6, where SuMPE never performs the worst [9] and ReMPE may introduce extremely high bias.

Table 6. Bias of the estimators on synthetic data. Several state-of-the-art MPE algorithms DPL, EN, KM and TIcE are selected. For each of them, we compute the bias (avg$[\widehat{\kappa}^* - \kappa^*]$) of the original, regrouping and subsampling version. The biggest absolute bias among the original, regrouping and subsampling versions is in bold.

κ^*	DPL	ReDPL	SuDPL	EN	ReEN	SuEN	KM	ReKM	SuKM	TIcE	ReTIcE	SuTIcE
0.1	**+0.029**	+0.021	+0.010	**+0.020**	+0.005	+0.008	**+0.020**	+0.007	+0.006	+0.085	**+0.089**	+0.063
0.25	+0.012	**−0.062**	−0.017	−0.023	**−0.080**	−0.043	−0.016	**−0.074**	−0.031	**+0.049**	+0.010	+0.024
0.5	+0.038	**−0.128**	−0.003	−0.045	**−0.156**	−0.067	−0.001	**−0.133**	−0.018	**+0.059**	−0.048	+0.032
0.75	+0.012	**−0.282**	−0.015	−0.071	**−0.278**	−0.088	0.000	**−0.278**	−0.012	+0.089	**−0.199**	+0.037

[9] In some cases, subsampling slightly increases the bias compared to original version, this is partly due to $P(Y|X)$ being estimated.