Unified Risk Analysis for Weakly Supervised Learning

Chao-Kai Chiang *chaokai@k.u-tokyo.ac.jp*
Masashi Sugiyama *sugi@k.u-tokyo.ac.jp*
Department of Complexity Science and Engineering
Graduate School of Frontier Sciences
The University of Tokyo
5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8561, Japan

Reviewed on OpenReview: *https://openreview.net/forum?id=RGsdAwWuu6*

Abstract

Among the flourishing research of weakly supervised learning (WSL), we recognize the lack of a unified interpretation of the mechanism behind the weakly supervised scenarios, let alone a systematic treatment of the risk rewrite problem, a crucial step in the empirical risk minimization approach. In this paper, we introduce a framework providing a comprehensive understanding and a unified methodology for WSL. The formulation component of the framework, leveraging a contamination perspective, provides a unified interpretation of how weak supervision is formed and subsumes fifteen existing WSL settings. The induced reduction graphs offer comprehensive connections over WSLs. The analysis component of the framework, viewed as a decontamination process, provides a systematic method of conducting risk rewrite. In addition to the conventional inverse matrix approach, we devise a novel strategy called marginal chain aiming to decontaminate distributions. We justify the feasibility of the proposed framework by recovering existing rewrites reported in the literature.

1 Introduction

Accurate labels allow one to generalize to unseen data via empirical risk minimization (ERM) and analyze the generalization error in terms of the classification risk. In practice, there are various situations in which acquiring accurate labels is hard or even impossible. One obstacle preventing us from acquiring accurate labels is labeling restrictions, such as imperfect supervision due to imperceptibility, time constraints, annotation costs, and even data sensitivity. Another obstacle is the disruption by unavoidable noise from the environment.

To address the first obstacle of restrictions, various formulations have been studied under the notion of weakly supervised learning (WSL) (Zhou, 2018; Sugiyama et al., 2022). Based on various types of available label information, it evolves to thriving topics, including the conventional settings (Lu et al., 2019; 2020; 2021; Elkan & Noto, 2008; du Plessis et al., 2014; 2015; Niu et al., 2016; Kiryo et al., 2017; Sansone et al., 2019) that investigating the potential of unlabeled data, complementary-label learning (Ishida et al., 2017; 2019; Yu et al., 2018; Feng et al., 2020a; Katsura & Uchida, 2020; Chou et al., 2020), partial-label learning (Cour et al., 2011; Wang et al., 2019; Lv et al., 2020; Feng et al., 2020b; Wu et al., 2023), learning with confidence information (Ishida et al., 2018; Cao et al., 2021a;b; Berthon et al., 2021; Ishida et al., 2023), and learning with comparative information (Bao et al., 2018; Shimada et al., 2021; Feng et al., 2021; Cao et al., 2021b). Developing to resolve the second obstacle of noise, learning with noisy labels (LNL) can be categorized into two major formulations; one is called mutually contaminated distributions (MCD) (Scott et al., 2013; Menon et al., 2015; Katz-Samuels et al., 2019) in which class-conditional distributions contaminate each other, and the other is named class-conditional random label noise (CCN) (Natarajan et al., 2013; 2017) where a label is flipped by random noise.

Despite fruitful results and tremendous impact, we recognize a lack of global understanding and systematic treatment of WSL. From the perspective of *formulation*, there are only scattered links among WSLs. Lu et al. (2019) and Feng et al. (2021) showed that parameter substitution could reduce unlabeled-unlabeled to similar-unlabeled and positive-unlabeled settings. Figure 1 in Wu et al. (2023) showed relationships among four WSLs of partial- and complementary-labels. A similar observation can be found in the intersection of WSLs and LNLs. Several WSLs were shown to be special cases of the MCD model, and some other WSLs are special cases of the CCN model. For details, please refer to the discussions in Sections 8.2.3 and 9.2.4 of Sugiyama et al. (2022). These connections encourage us to consider the possibility that there exists a unique interpretation that explains the mechanism behind WSL. Luckily, from the *methodological* viewpoint, most of the existing WSL research adopted certain forms of the ERM approach. A crucial shared step is to perform the risk rewrite, a way of rephrasing the uncomputable risk to a computable one in terms of the data-generating distributions. A successful rewrite is the starting point of many downstream tasks, including but not limited to the following: Devising a practical or robust objective for training, comparing the strengths and properties of loss functions, proving the consistency, and analyzing generalization error bounds. However, many rewrite forms (summarized in Tables 4 and 5) look independent as if they are tailored to fit each problem's unique form of supervision and are not adaptable to each other. These seemingly non-adaptable estimators post a practical challenge: When facing a new form of weak (or noisy) supervision, we do not have a guideline or general strategy to leverage developed methods to address the new situation.

These observations raise the following questions we aim to answer in this paper: What is the essence of WSL? From a formulation perspective, can a unique interpretation be found to explain the mechanism behind WSL? Does a methodology exist to address as many WSLs as possible?

This paper proposes a framework with the following contributions to answer the research questions.

1. To the best of our knowledge, the framework is the first systematic attempt to address how and why WSLs are related. The framework consists of a formulation component and an analysis component, and subsumes fifteen weakly supervised scenarios. Table 10 summarizes the results obtained from our framework. This paper brings forth the next two new insights.

2. The formulation component, modeling from a *contamination* perspective, provides a coherent interpretation of the weakly supervised data-generating processes. It produces a comprehensive relationship graph, shown in Figure 1, consisting of Tables 7, 8, and 9. Figure1 summarizes the WSLs and reveals connections between scenarios that were previously unknown to the community. Figure 1 is our answer to the second research question. Figure 1 also unveils a distinctive confidence-based type of WSLs that do not belong to the prominent MCD or CCN categories.

3. The analysis component, leveraging the *decontamination* concept, establishes a generic methodology for conducting risk rewrites for all WSLs discussed in this paper. Thus, rewrite derivations that previously seemed irrelevant can now be systematically analyzed, and the final research question is answered. In addition, the analysis component also applies to new scenarios.

4. Regarding the technical contributions, the proposed framework distinguishes two approaches, the inversion method and the marginal chain method presented in Theorems 1 and 2, to implement the decontamination idea. The marginal chain method is a newly developed invertibility-free loss correction approach. We also illustrate the subtle adjustments to develop simplified and intuitive proofs for the existing risk rewrites. These alternative proofs have their own logic stemming from the proposed framework.

The idea of decontamination has been widely implemented and investigated. There are two major approaches, loss correction, and label correction, in LNL. Closest to the current paper, Cid-Sueiro (2012), van Rooyen & Williamson (2017), Katz-Samuels et al. (2019), Patrini et al. (2017), and van Rooyen & Williamson (2015) exploited the inverse matrix, sometimes known as the backward method (Patrini et al., 2017), to construct a corrected training loss to obtain an unbiased estimator. There were deep learning methods leveraging the contamination assumption, sometimes called the forward method (Patrini et al., 2017), to train a classifier (Patrini et al., 2017; Yu et al., 2018; Sukhbaatar & Fergus, 2015; Goldberger & Ben-Reuven, 2017; Berthon

et al., 2021). Besides modifying the loss function, one has two other strategies to manipulate the corrupted labels. The (iterative) pseudo-label method modified the labels for training (Ma et al., 2018; Tanaka et al., 2018; Reed et al., 2015). Filtering clean data points for training is the other option (Northcutt et al., 2017; 2021; Jiang et al., 2018; Han et al., 2018; Yu et al., 2019). Apart from classification, a different research branch studies conditions and methods for recovering the base distributions (Katz-Samuels et al., 2019; Blanchard & Scott, 2014; Blanchard et al., 2016).

The current work is close to the loss correction approach in LNL. Most previous loss correction methods exploited invertibility to construct the corrected losses. In contrast, the marginal chain approach we propose in this paper adopts the conditional probability formula to build the corrected losses. Many of the existing work targeted either the MCD or the CCN models. Scott & Zhang (2020), Berthon et al. (2021), Patrini et al. (2017), Goldberger & Ben-Reuven (2017), Sukhbaatar & Fergus (2015), Yu et al. (2018), Natarajan et al. (2013), Natarajan et al. (2017), Northcutt et al. (2017), and Northcutt et al. (2021) were based on the CCN model, and Katz-Samuels et al. (2019), Blanchard & Scott (2014), and Blanchard et al. (2016) were based on the MCD model. Menon et al. (2015), van Rooyen & Williamson (2017), and Katz-Samuels et al. (2019) studied multiple noise models at the same time. However, the current paper investigates the connections between MCD, CCN, and confidence-based settings simultaneously through the lens of matrix decontamination as broadly as possible to identify a generic methodology for WSLs. Different from the current paper aiming for risk minimization, research also studied various performance measures, such as the balanced error rate (Scott & Zhang, 2020; 2019; Menon et al., 2015; du Plessis et al., 2013), the area under the receiver operating characteristic curve (Charoenphakdee et al., 2019; Sakai et al., 2018; Menon et al., 2015), and cost-sensitive measures (Charoenphakdee et al., 2021; Natarajan et al., 2017). We choose the classification risk as the only measure due to the focus of this paper.

The remaining sections are organized as follows. Section 2 reviews ERM in supervised learning, the risk rewrite problem, and the existing results. Section 3 presents the proposed framework. We show that the proposed framework provides a unified way to formulate diverse weakly supervised scenarios in Section 4. Section 5 demonstrates how to instantiate the framework to conduct risk rewrite. We demonstrate the applicability of the proposed framework to new scenarios in Section 6. Finally, we conclude the paper and discuss outlooks in Section 7. The current organization of Sections 4 and 5 aims at connecting multiple WSLs under one framework. We note that this paper can serve multiple purposes for the study of WSLs. A summary of possible use cases of the paper is provided in Appendix A.

2 Preliminaries

Let (y, x) be a data example where the instance x X and the label y Y. For binary classification, the label space Y is $\{\mathrm{p}, \mathrm{n}\}$, and for multiclass classification with K classes, $\mathsf{Y} = \{1, 2, \ldots, K\} := [K]$. The joint distribution is $\mathsf{P}(Y, X)$, the class prior is $\mathsf{P}(Y)$, the class-conditional distribution is $\mathsf{P}(X|Y)$, and the class probability function is $\mathsf{P}(Y|X)$. Given a space of hypotheses G, we denote the loss of a hypothesis g G on predicting y of (y, x) as $\ell_{Y=y}(g(x))$. To accommodate concise expressions and readability for all WSLs considered in this paper simultaneously, we use alias notations when the context is unambiguous. Table 1 provides a set of common notations used in this paper.

We use (y, x) instead of the convention (x, y) to represent a data example because, in the current paper, we focus on discussing different types of supervision. Placing the label before the instance emphasizes the type of supervision under investigation in theorems and derivations.

2.1 Supervised Learning and the ERM Method

In supervised learning with K classes, the observed data is of the form

$$\{x_i^y\}_{i=1}^{n_y} \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{X|Y=y}, \quad y \quad [K].$$

Table 1: Alias of Common Notations.

Name of the notation	Expression	Aliases
Binary classes	$\{\mathrm{p}, \mathrm{n}\}$	
Multiple classes	$\{1, \ldots, K\}$	$[K]$
Compound set of $[K]$	$2^{[K]} \setminus \{\ , [K]\}$	S
Joint distribution	$\mathsf{P}(Y = y, X = x)$	$P_{Y=y,x}$, $P_{Y=y,X}$, or $P_{Y,X}$
Hypothesis and its space	$g \quad \mathsf{G}$	
Loss of g	$\ell_{Y=y}(g(x))$	ℓ_y, $\ell_y(X)$, or $\ell_Y(g(X))$
Classification risk	$\mathsf{E}_{Y,X}\left[\ell_Y(g(X))\right]$	$R(g)$
The j-th entry of vector V	$(V)_j$	V_j
Class prior	$\mathsf{P}(Y = y)$	π_y
Marginal	$\mathsf{P}(X)$	P_X
Class-conditional	$\mathsf{P}(X = x \mid Y = y)$	$P_{X\mid Y}$, $P_{X\mid Y=y}$, or $P_{x\mid Y=y}$
Confidence	$\mathsf{P}(Y = y \mid X = x)$	$r_y(X)$, $r_y(x)$, or $r(X)$ if $y = \mathrm{p}$

Notation x_i^y denotes the shorthand of (y, x_i). The goal of learning is to find a classifier $g \quad \mathsf{G}$ that minimizes the classification risk

$$R(g) := \mathsf{E}_{Y,X}\left[\ell_Y(g(X))\right] = \sum_{y=1}^{K} \int_{x \quad \mathsf{X}} P_{Y=y,x}\, \ell_{Y=y}(g(x))\, \mathrm{d}x. \tag{1}$$

To find such a classifier, ERM first constructs an empirical risk estimator with the data in hand:

$$\hat{R}(g) = \sum_{y=1}^{K} \frac{1}{n_y} \sum_{i=1}^{n_y} \pi_y \ell_{Y=y}(g(x_i^y)). \tag{2}$$

The estimator approximates $R(g)$ consistently since it can be shown that (2) approaches (1) as $N \to \infty$ (Tewari & Bartlett, 2014; Kiryo et al., 2017) and (Sugiyama et al., 2022, Chapter 3). Then, ERM takes $\hat{R}(g)$ as the training objective and optimizes it to find the optimal classifier

$$g \quad = \arg\min_{g \quad \mathsf{G}} \hat{R}(g) \tag{3}$$

in the hypothesis space G as the output of ERM.

2.2 The Risk Rewrite Problem and Existing Results

Sections 2.2.1 to 2.2.17 review the learning scenarios including WSLs, MCD, and CCN that will be discussed in this paper. A knowledgeable reader may refer directly to summary Tables 2 through 6 and proceed to Section 3.

In every WSL scenario, the goal of learning is the same as supervised learning. However, the observed data is no longer as perfectly labeled as in supervised learning. That said, there are differences in the formulations of the observed data and the ways of estimating the classification risk. We begin with reviewing WSLs derived from binary classes. For $K = 2$, we denote $\mathsf{Y} := \{\mathrm{p}, \mathrm{n}\}$.

2.2.1 Positive-Unlabeled (PU) learning

The observed data in PU learning (du Plessis et al., 2015) is of the form

$$\begin{aligned} \{x_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p}}} &\overset{\text{i.i.d.}}{\sim} P_{\mathrm{P}} := P_{X\mid Y=\mathrm{p}}, \\ \{x_j^{\mathrm{u}}\}_{j=1}^{n_{\mathrm{u}}} &\overset{\text{i.i.d.}}{\sim} P_{\mathrm{U}} := \pi_{\mathrm{p}}\, P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{n}}\, P_{X\mid Y=\mathrm{n}}, \end{aligned} \tag{4}$$

where x_j^{u} is viewed as the shorthand of (u, x_j) symbolizing the unlabeled data[1]. The unlabeled data set $\{x_j^{\mathrm{u}}\}_j$ consists of a mixture of samples from $P_{X|Y=\mathrm{p}}$ and $P_{X|Y=\mathrm{n}}$ with proportion π_{p}. Since the information of negatively sampled data is unavailable, (2) is uncomputable, causing directly optimizing (3) infeasibility. Therefore, to make ERM applicable, the ***risk rewrite problem*** (Sugiyama et al., 2022) asks:

Can one rephrase the classification risk $R(g)$ (1) in terms of the given data formulation?

du Plessis et al. (2015) rewrote the classification risk in terms of the data-generating distributions P_{P} and P_{U} as

$$R(g) = \mathsf{E}_{\mathrm{P}}\left[\pi_{\mathrm{p}}\ell_{\mathrm{p}} - \pi_{\mathrm{p}}\ell_{\mathrm{n}}\right] + \mathsf{E}_{\mathrm{U}}\left[\ell_{\mathrm{n}}\right]. \tag{5}$$

2.2.2 Positive-confidence (Pconf) Learning Learning

The observed data in Pconf learning (Ishida et al., 2018) is of the form

$$\{x_i, r(x_i)\}_{i=1}^{n},$$

where

$$\begin{aligned} x_i &\overset{\text{i.i.d.}}{} P_{\mathrm{P}} := P_{X|Y=\mathrm{p}}, \\ r(x_i) &:= P_{Y=\mathrm{p}|X=x_i}. \end{aligned} \tag{6}$$

The function $r(x)$ represents how confident an example x would be positively labeled. Ishida et al. (2018) rewrote the classification risk as

$$R(g) = \pi_{\mathrm{p}}\mathsf{E}_{\mathrm{P}}\left[\ell_{\mathrm{p}} + \frac{1 - r(X)}{r(X)}\ell_{\mathrm{n}}\right]. \tag{7}$$

2.2.3 Unlabeled-Unlabeled (UU) learning

The observed data in UU learning (Lu et al., 2019) is of the form

$$\begin{aligned} \{x_i^{\mathrm{u}_1}\}_{i=1}^{n_{\mathrm{u}_1}} &\overset{\text{i.i.d.}}{} P_{\mathrm{U}_1} := (1 - \gamma_1)\, P_{X|Y=\mathrm{p}} + \gamma_1\, P_{X|Y=\mathrm{n}}, \\ \left\{x_j^{\mathrm{u}_2}\right\}_{j=1}^{n_{\mathrm{u}_2}} &\overset{\text{i.i.d.}}{} P_{\mathrm{U}_2} := \gamma_2\, P_{X|Y=\mathrm{p}} + (1 - \gamma_2)\, P_{X|Y=\mathrm{n}}, \end{aligned} \tag{8}$$

where $x_i^{\mathrm{u}_1}$ (resp. $x_j^{\mathrm{u}_2}$) being the shorthand of (u_1, x_i) (resp. (u_2, x_j)) represents x_i (resp. x_j) belonging to the unlabeled data whose mixture parameter is γ_1 (resp. γ_2). Notice a difference that the mixture proportion of the unlabeled data in PU learning is π_{p}. Lu et al. (2019) rewrote the classification risk in terms of the data-generating distributions P_{U_1} and P_{U_2} as follows: Assume $\gamma_1 + \gamma_2 = 1$. Then,

$$R(g) = \mathsf{E}_{\mathrm{U}_1}\left[\frac{(1 - \gamma_2)\pi_{\mathrm{p}}}{1 - \gamma_1 - \gamma_2}\ell_{\mathrm{p}} + \frac{-\gamma_2\pi_{\mathrm{n}}}{1 - \gamma_1 - \gamma_2}\ell_{\mathrm{n}}\right] + \mathsf{E}_{\mathrm{U}_2}\left[\frac{-\gamma_1\pi_{\mathrm{p}}}{1 - \gamma_1 - \gamma_2}\ell_{\mathrm{p}} + \frac{(1 - \gamma_1)\pi_{\mathrm{n}}}{1 - \gamma_1 - \gamma_2}\ell_{\mathrm{n}}\right]. \tag{9}$$

2.2.4 Similar-Unlabeled (SU) learning

The observed data in SU learning (Bao et al., 2018) is of the form

$$\begin{aligned} \left\{\left(x_i^{\mathrm{s}}, x_i^{\mathrm{s}}\right)\right\}_{i=1}^{n_{\mathrm{s}}} &\overset{\text{i.i.d.}}{} P_{\mathrm{S}} := \frac{\pi_{\mathrm{p}}^2 P_{X|Y=\mathrm{p}} P_{X\ |Y=\mathrm{p}} + \pi_{\mathrm{n}}^2 P_{X|Y=\mathrm{n}} P_{X\ |Y=\mathrm{n}}}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}, \\ \left\{x_j^{\mathrm{u}}\right\}_{j=1}^{n_{\mathrm{u}}} &\overset{\text{i.i.d.}}{} P_{\mathrm{U}} := \pi_{\mathrm{p}}\, P_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}\, P_{X|Y=\mathrm{n}}. \end{aligned} \tag{10}$$

The word "similar" means the examples in every $(x^{\mathrm{s}}, x^{\mathrm{s}})$ pair have the same label; either both are positive, or both are negative. Under the assumption $\pi_{\mathrm{p}} = \pi_{\mathrm{n}}$, Bao et al. (2018) rewrote the classification risk as

$$R(g) = \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\mathsf{E}_{\mathrm{S}}\left[\frac{L(X) + L(X)}{2}\right] + \mathsf{E}_{\mathrm{U}}\left[L_{-}(X)\right], \tag{11}$$

[1]Seemingly being redundant, but it is helpful to use (u, x_j) to distinguish it from the positively labeled instance (p, x_i).

where

$$\begin{aligned} \mathsf{L}(X) &:= \frac{1}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) - \frac{1}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X), \\ \mathsf{L}_{-}(X) &:= -\frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) + \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X). \end{aligned}$$

2.2.5 Dissimilar-Unlabeled (DU) learning

The observed data in DU learning (Shimada et al., 2021) is of the form

$$\begin{aligned} &\left\{\left(x_i^{\mathrm{d}}, x_i^{\mathrm{d}}\right)\right\}_{i=1}^{n_{\mathrm{d}}} \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathrm{D}} := \frac{\mathsf{P}_{X|Y=\mathrm{p}}\mathsf{P}_{X\,|Y=\mathrm{n}} + \mathsf{P}_{X|Y=\mathrm{n}}\mathsf{P}_{X\,|Y=\mathrm{p}}}{2}, \\ &\left\{x_j^{\mathrm{u}}\right\}_{j=1}^{n_{\mathrm{u}}} \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathrm{U}} := \pi_{\mathrm{p}}\,\mathsf{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}\,\mathsf{P}_{X|Y=\mathrm{n}}. \end{aligned} \tag{12}$$

The word "dissimilar" means the examples in every $(x^{\mathrm{d}}, x^{\mathrm{d}})$ pair have distinct labels. Under the assumption $\pi_{\mathrm{p}} = \pi_{\mathrm{n}}$, Shimada et al. (2021) rewrote the classification risk as

$$R(g) = 2\pi_{\mathrm{p}}\pi_{\mathrm{n}}\mathsf{E}_{\mathrm{D}}\left[-\frac{\mathsf{L}(X) + \mathsf{L}(X')}{2}\right] + \mathsf{E}_{\mathrm{U}}\left[\mathsf{L}_{+}(X)\right], \tag{13}$$

where

$$\begin{aligned} \mathsf{L}(X) &= \frac{1}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) - \frac{1}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X), \\ \mathsf{L}_{+}(X) &:= \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) - \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X). \end{aligned}$$

Note that $\mathsf{L}(X)$ has been defined in the SU setting. We repeat it here for clarity.

2.2.6 Similar-Dissimilar (SD) learning

The observed data in SD learning (Shimada et al., 2021) is of the form

$$\begin{aligned} &\left\{\left(x_i^{\mathrm{s}}, x_i^{\mathrm{s}}\right)\right\}_{i=1}^{n_{\mathrm{s}}} \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathrm{S}} := \frac{\pi_{\mathrm{p}}^2\mathsf{P}_{X|Y=\mathrm{p}}\mathsf{P}_{X\,|Y=\mathrm{p}} + \pi_{\mathrm{n}}^2\mathsf{P}_{X|Y=\mathrm{n}}\mathsf{P}_{X\,|Y=\mathrm{n}}}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}, \\ &\left\{\left(x_i^{\mathrm{d}}, x_i^{\mathrm{d}}\right)\right\}_{i=1}^{n_{\mathrm{d}}} \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathrm{D}} := \frac{\mathsf{P}_{X|Y=\mathrm{p}}\mathsf{P}_{X\,|Y=\mathrm{n}} + \mathsf{P}_{X|Y=\mathrm{n}}\mathsf{P}_{X\,|Y=\mathrm{p}}}{2}. \end{aligned} \tag{14}$$

Under the assumption $\pi_{\mathrm{p}} = \pi_{\mathrm{n}}$, Shimada et al. (2021) rewrote the classification risk as

$$R(g) = \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\mathsf{E}_{\mathrm{S}}\left[\frac{\mathsf{L}_{+}(X) + \mathsf{L}_{+}(X')}{2}\right] + 2\pi_{\mathrm{p}}\pi_{\mathrm{n}}\mathsf{E}_{\mathrm{D}}\left[\frac{\mathsf{L}_{-}(X) + \mathsf{L}_{-}(X')}{2}\right], \tag{15}$$

where

$$\begin{aligned} \mathsf{L}_{+}(X) &= \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) - \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X), \\ \mathsf{L}_{-}(X) &= -\frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) + \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X). \end{aligned}$$

Note that $\mathsf{L}_{+}(X)$ and $\mathsf{L}_{-}(X)$ have been defined in the DU and SU settings. We repeat them here for clarity.

2.2.7 Pairwise Comparison (Pcomp) Learning

The observed data in Pcomp learning (Feng et al., 2021) is of the form

$$\left\{\left(x_i^{\mathrm{pc}}, x_i^{\mathrm{pc}}\right)\right\}_{i=1}^{n_{\mathrm{pc}}} \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathrm{PC}} := \frac{\pi_{\mathrm{p}}^2\mathsf{P}_{X|Y=\mathrm{p}}\mathsf{P}_{X\,|Y=\mathrm{p}} + \pi_{\mathrm{p}}\pi_{\mathrm{n}}\mathsf{P}_{X|Y=\mathrm{p}}\mathsf{P}_{X\,|Y=\mathrm{n}} + \pi_{\mathrm{n}}^2\mathsf{P}_{X|Y=\mathrm{n}}\mathsf{P}_{X\,|Y=\mathrm{n}}}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{p}}\pi_{\mathrm{n}} + \pi_{\mathrm{n}}^2}. \tag{16}$$

The pairwise comparison encodes a meaning that each x^{pc} "can not be more negative" than x^{pc} in the $(x^{\mathrm{pc}}, x^{\mathrm{pc}})$ pair. That is, the labels in $(x^{\mathrm{pc}}, x^{\mathrm{pc}})$ are of the form (p, p), (p, n), or (n, n). Feng et al. (2021) rewrote the classification risk as

$$R(g) = \mathsf{E}_{\mathrm{Sup}}\left[\ell_{\mathrm{p}} - \pi_{\mathrm{p}}\ell_{\mathrm{n}}\right] + \mathsf{E}_{\mathrm{Inf}}\left[-\pi_{\mathrm{n}}\ell_{\mathrm{p}} + \ell_{\mathrm{n}}\right], \tag{17}$$

where the expectations are computed over the following distributions

$$\begin{aligned} P_{\mathrm{Sup}} &:= \int_{x} {}_{X} P_{\mathrm{PC}}\, \mathrm{d}x\,, \\ P_{\mathrm{Inf}} &:= \int_{x} {}_{X} P_{\mathrm{PC}}\, \mathrm{d}x. \end{aligned}$$

2.2.8 Similarity-Confidence Learning (Sconf) Learning

The observed data in Sconf learning (Cao et al., 2021b) is of the form

$$\left\{x_i^{\mathrm{sc}}, x_i^{\mathrm{sc}}\,, r\left(x_i^{\mathrm{sc}}, x_i^{\mathrm{sc}}\right)\right\}_{i=1}^{n},$$

where

$$\begin{aligned} &x_i^{\mathrm{sc}} \overset{\text{i.i.d.}}{\sim} P_X := \pi_{\mathrm{p}}\, P_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}\, P_{X|Y=\mathrm{n}}, \\ &x_i^{\mathrm{sc}} \overset{\text{i.i.d.}}{\sim} P_X := \pi_{\mathrm{p}}\, P_{X\,|Y=\mathrm{p}} + \pi_{\mathrm{n}}\, P_{X\,|Y=\mathrm{n}}, \\ &r\left(x_i^{\mathrm{sc}}, x_i^{\mathrm{sc}}\right) := P_{Y=y_i^{\mathrm{sc}}=Y\,=y_i^{\mathrm{sc}}\,|X=x_i^{\mathrm{sc}},X\,=x_i^{\mathrm{sc}}}\,. \end{aligned} \tag{18}$$

Cao et al. (2021b) rewrote the classification risk as

$$R(g) = \mathsf{E}_{X,X}\left[\frac{r(X,X) - \pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} L_{\mathrm{p}}(X,X) + \frac{\pi_{\mathrm{p}} - r(X,X)}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} L_{\mathrm{n}}(X,X)\right], \tag{19}$$

where

$$\begin{aligned} L_{\mathrm{p}}(X,X) &:= \frac{\ell_{\mathrm{p}}(X) + \ell_{\mathrm{p}}(X)}{2}, \\ L_{\mathrm{n}}(X,X) &:= \frac{\ell_{\mathrm{n}}(X) + \ell_{\mathrm{n}}(X)}{2}. \end{aligned}$$

2.2.9 Complementary-Label (CL) Learning

One can also formulate weak supervision from multiclass classification. For K classes, we denote $Y := [K]$.

The observed data in CL learning (Ishida et al., 2019) is of the form

$$\{(\bar{s}_i, x_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{\bar{S},X} := \frac{1}{K-1}\sum_{Y=\bar{S}} P_{Y,X}. \tag{20}$$

As is named "complementary," $\bar{s} \quad [K]$ represents that the true label y of x cannot be $\bar{s}$. Ishida et al. (2019) rewrote the classification risk as

$$R(g) = \mathsf{E}_{\bar{S},X}\left[\sum_{y=1}^{K} \ell_y - (K-1)\ell_{\bar{S}}\right]. \tag{21}$$

2.2.10 Multi-Complementary-Label (MCL) Learning

The observed data in MCL learning (Feng et al., 2020a) is of the form

$$\{(\bar{s}_i, x_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{\bar{S},X} := \begin{cases} \sum_{d=1}^{K-1} P_{|\bar{S}|=d} \cdot \frac{1}{\binom{K-1}{|\bar{S}|}} \sum_{Y \not\in \bar{S}} P_{Y,X}, & \text{if } |\bar{S}| = d, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

Generalized from CL, $\bar{s} \subset [K]$ in MCL is a set of classes of size $d \subset [K-1]$, representing multiple exclusions. In other words, CL is the special case of MCL with $d = 1$. Feng et al. (2020a) rewrote the classification risk as

$$R(g) = \sum_{d=1}^{K-1} P_{|\bar{S}|=d} \mathsf{E}_{\bar{S},X||\bar{S}|=d} \left[\sum_{y \not\in \bar{S}} \ell_y - \frac{K-1-|\bar{S}|}{|\bar{S}|} \sum_{\bar{s} \subset \bar{S}} \ell_{\bar{s}} \right]. \tag{23}$$

2.2.11 Provably Consistent Partial-Label (PCPL) Learning

The observed data in PCPL learning (Feng et al., 2020b) is of the form

$$\{(s_i, x_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{S,X} := \frac{1}{2^{K-1}-1} \sum_{Y \subset S} P_{Y,X}. \tag{24}$$

A partial-label $s \subset [K]$ is a set of classes containing the true label y of x. Feng et al. (2020b) rewrote the classification risk as

$$R(g) = \frac{1}{2} \mathsf{E}_{S,X} \left[\sum_{y=1}^{K} \frac{P_{Y=y|X}}{\sum_{a \subset S} P_{Y=a|X}} \ell_y \right]. \tag{25}$$

2.2.12 Proper Partial-Label (PPL) Learning

The observed data in PPL learning (Wu et al., 2023) is of the form

$$\{(s_i, x_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{S,X} := C(S,X) \sum_{Y \subset S} P_{Y,X}. \tag{26}$$

The weight $\frac{1}{2^{K-1}-1}$ in PCPL is generalized to $C(S,X)$, a function of the partial-label and the instance, allowing one to characterize the "properness" of a partial-label. Wu et al. (2023) rewrote the classification risk as

$$R(g) = \mathsf{E}_{S,X} \left[\sum_{y \subset S} \frac{P_{Y=y|X}}{\sum_{a \subset S} P_{Y=a|X}} \ell_y \right]. \tag{27}$$

2.2.13 Single-Class Confidence (SC-Conf) Learning

The observed data in SC-Conf learning (Cao et al., 2021a) is of the form

$$\{x_i, r_1(x_i), \ldots, r_K(x_i)\}_{i=1}^{n},$$

where

$$\begin{aligned} &x_i \overset{\text{i.i.d.}}{\sim} P_{X|Y=y_{\mathrm{s}}} \text{ with } y_{\mathrm{s}} \subset [K], \\ &r_k(x_i) := P_{Y=k|X=x_i} \text{ for each } k \subset [K]. \end{aligned} \tag{28}$$

The constraint of SC-Conf is that the examples are sampled from a specific class y_{s}. The key to risk rewrite is the availability of confident information $r_k(x)$ about each class. Cao et al. (2021a) rewrote the classification risk as

$$R(g) = \pi_{y_{\mathrm{s}}} \mathsf{E}_{X|Y=y_{\mathrm{s}}} \left[\sum_{y=1}^{K} \frac{r_y(X)}{r_{y_{\mathrm{s}}}(X)} \ell_y \right]. \tag{29}$$

2.2.14 Subset Confidence (Sub-Conf) Learning

The observed data in Sub-Conf learning (Cao et al., 2021a) is of the form

$$\{x_i, r_1(x_i), \ldots, r_K(x_i)\}_{i=1}^{n},$$

where

$$\begin{aligned} &x_i \overset{\text{i.i.d.}}{\sim} P_{X|Y \ \ Y_{\mathrm{s}}} \text{ with } Y_{\mathrm{s}} \quad [K], \\ &r_k(x_i) := P_{Y=k|X=x_i} \text{ for each } k \quad [K]. \end{aligned} \tag{30}$$

Sub-Conf is a relaxed setting of SC-Conf where the samples come from a set of classes Y_{s}. Cao et al. (2021a) rewrote the classification risk as

$$R(g) = \pi_{Y_{\mathrm{s}}} \mathsf{E}_{X|Y \ \ Y_{\mathrm{s}}} \left[\sum_{y=1}^{K} \frac{r_y(X)}{r_{Y_{\mathrm{s}}}(X)} \ell_y \right], \tag{31}$$

where $\pi_{Y_{\mathrm{s}}} := \sum_{j \ \ Y_{\mathrm{s}}} \pi_j$, and $r_{Y_{\mathrm{s}}}(X) := P_{Y \ \ Y_{\mathrm{s}}|X} = \sum_{j \ \ Y_{\mathrm{s}}} P_{Y=j|X}$.

2.2.15 Soft-Label Learning

Ishida et al. (2023) formulated soft-label learning under the binary setting, in which the observed data is of the form

$$\{x_i, r(x_i)\}_{i=1}^{n},$$

where

$$\begin{aligned} &x_i \overset{\text{i.i.d.}}{\sim} P_X := P_{Y=\mathrm{p},X} + P_{Y=\mathrm{n},X}, \\ &r(x_i) := P_{Y=\mathrm{p}|X=x_i}. \end{aligned} \tag{32}$$

It is straightforward to obtain a corresponding formulation under the multiclass setting:

$$\{x_i, r_1(x_i), \ldots, r_K(x_i)\}_{i=1}^{n},$$

where

$$\begin{aligned} &x_i \overset{\text{i.i.d.}}{\sim} P_X := \sum_{k=1}^{K} P_{Y=k,X}, \\ &r_k(x_i) := P_{Y=k|X=x_i} \text{ for each } k \quad [K]. \end{aligned} \tag{33}$$

The difference between SC-Conf and multiclass soft-label (resp. the difference between Pconf and binary soft-label) is the sample distribution of x. We rewrote the classification risk as

$$R(g) = \mathsf{E}_X \left[\sum_{y=1}^{K} r_y(X) \ell_y \right]. \tag{34}$$

2.2.16 Summary of Existing WSL Formulations and Risk Rewrites

We summarize the weakly supervised scenarios discussed and their risk rewrite results. The formulations are divided into the binary classification settings in Table 2 and the multiclass classification settings in Table 3. We list the formulations in chronological order, according to their publication order. Tables 4 and 5 are the corresponding rewrites.

Table 2: Binary WSL formulations.

WSL	Formulation	Equation
PU	$\{x_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p}}} \overset{\text{i.i.d.}}{} P_{\mathrm{P}} := P_{X\mid Y=\mathrm{p}}$, $\{x_j^{\mathrm{u}}\}_{j=1}^{n_{\mathrm{u}}} \overset{\text{i.i.d.}}{} P_{\mathrm{U}} := \pi_{\mathrm{p}} P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{n}} P_{X\mid Y=\mathrm{n}}$.	(4)
Pconf	$\{x_i, r(x_i)\}_{i=1}^{n}$, where $x_i \overset{\text{i.i.d.}}{} P_{\mathrm{P}} := P_{X\mid Y=\mathrm{p}}$, $r(x_i) := P_{Y=\mathrm{p}\mid X=x_i}$.	(6)
UU	$\{x_i^{\mathrm{u}_1}\}_{i=1}^{n_{\mathrm{u}_1}} \overset{\text{i.i.d.}}{} P_{\mathrm{U}_1} := (1-\gamma_1) P_{X\mid Y=\mathrm{p}} + \gamma_1 P_{X\mid Y=\mathrm{n}}$, $\{x_j^{\mathrm{u}_2}\}_{j=1}^{n_{\mathrm{u}_2}} \overset{\text{i.i.d.}}{} P_{\mathrm{U}_2} := \gamma_2 P_{X\mid Y=\mathrm{p}} + (1-\gamma_2) P_{X\mid Y=\mathrm{n}}$.	(8)
SU	$\left\{\left(x_i^{\mathrm{s}}, x_i^{\mathrm{s}}\right)\right\}_{i=1}^{n_{\mathrm{s}}} \overset{\text{i.i.d.}}{} P_{\mathrm{S}} := \dfrac{\pi_{\mathrm{p}}^2 P_{X\mid Y=\mathrm{p}} P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{n}}^2 P_{X\mid Y=\mathrm{n}} P_{X\mid Y=\mathrm{n}}}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}$, $\left\{x_j^{\mathrm{u}}\right\}_{j=1}^{n_{\mathrm{u}}} \overset{\text{i.i.d.}}{} P_{\mathrm{U}} := \pi_{\mathrm{p}} P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{n}} P_{X\mid Y=\mathrm{n}}$.	(10)
DU	$\left\{\left(x_i^{\mathrm{d}}, x_i^{\mathrm{d}}\right)\right\}_{i=1}^{n_{\mathrm{d}}} \overset{\text{i.i.d.}}{} P_{\mathrm{D}} := \dfrac{P_{X\mid Y=\mathrm{p}} P_{X\mid Y=\mathrm{n}} + P_{X\mid Y=\mathrm{n}} P_{X\mid Y=\mathrm{p}}}{2}$, $\left\{x_j^{\mathrm{u}}\right\}_{j=1}^{n_{\mathrm{u}}} \overset{\text{i.i.d.}}{} P_{\mathrm{U}} := \pi_{\mathrm{p}} P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{n}} P_{X\mid Y=\mathrm{n}}$.	(12)
SD	$\left\{\left(x_i^{\mathrm{s}}, x_i^{\mathrm{s}}\right)\right\}_{i=1}^{n_{\mathrm{s}}} \overset{\text{i.i.d.}}{} P_{\mathrm{S}} := \dfrac{\pi_{\mathrm{p}}^2 P_{X\mid Y=\mathrm{p}} P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{n}}^2 P_{X\mid Y=\mathrm{n}} P_{X\mid Y=\mathrm{n}}}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}$, $\left\{\left(x_i^{\mathrm{d}}, x_i^{\mathrm{d}}\right)\right\}_{i=1}^{n_{\mathrm{d}}} \overset{\text{i.i.d.}}{} P_{\mathrm{D}} := \dfrac{P_{X\mid Y=\mathrm{p}} P_{X\mid Y=\mathrm{n}} + P_{X\mid Y=\mathrm{n}} P_{X\mid Y=\mathrm{p}}}{2}$.	(14)
Pcomp	$\left\{\left(x_i^{\mathrm{pc}}, x_i^{\mathrm{pc}}\right)\right\}_{i=1}^{n_{\mathrm{pc}}} \overset{\text{i.i.d.}}{} P_{\mathrm{PC}}$ $:= \dfrac{\pi_{\mathrm{p}}^2 P_{X\mid Y=\mathrm{p}} P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{p}}\pi_{\mathrm{n}} P_{X\mid Y=\mathrm{p}} P_{X\mid Y=\mathrm{n}} + \pi_{\mathrm{n}}^2 P_{X\mid Y=\mathrm{n}} P_{X\mid Y=\mathrm{n}}}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{p}}\pi_{\mathrm{n}} + \pi_{\mathrm{n}}^2}$.	(16)
Sconf	$\left\{x_i^{\mathrm{sc}}, x_i^{\mathrm{sc}}, r\left(x_i^{\mathrm{sc}}, x_i^{\mathrm{sc}}\right)\right\}_{i=1}^{n_{\mathrm{sc}}}$, where $x_i^{\mathrm{sc}} \overset{\text{i.i.d.}}{} P_X := \pi_{\mathrm{p}} P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{n}} P_{X\mid Y=\mathrm{n}}$, $x_i^{\mathrm{sc}} \overset{\text{i.i.d.}}{} P_X := \pi_{\mathrm{p}} P_{X\mid Y=\mathrm{p}} + \pi_{\mathrm{n}} P_{X\mid Y=\mathrm{n}}$, $r\left(x_i^{\mathrm{sc}}, x_i^{\mathrm{sc}}\right) := P_{Y=y_i^{\mathrm{sc}}=Y=y_i^{\mathrm{sc}}\mid X=x_i^{\mathrm{sc}}, X=x_i^{\mathrm{sc}}}$.	(18)

Table 3: Multiclass WSL formulations.

WSL	Formulation	Equation
CL	$\{(\bar{s}_i, x_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{\bar{S},X} := \frac{1}{K-1} \sum_{Y=\bar{S}} P_{Y,X}$.	(20)
MCL	$\{(\bar{s}_i, x_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{\bar{S},X} := \begin{cases} \sum_{d=1}^{K-1} P_{\lvert\bar{S}\rvert=d} \cdot \frac{1}{\binom{K-1}{\lvert\bar{S}\rvert}} \sum_{Y \not\in \bar{S}} P_{Y,X}, & \text{if } \lvert\bar{S}\rvert = d, \\ 0, & \text{otherwise.} \end{cases}$	(22)
PCPL	$\{(s_i, x_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{S,X} := \frac{1}{2^{K-1}-1} \sum_{Y \quad S} P_{Y,X}$.	(24)
PPL	$\{(s_i, x_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{S,X} := C(S,X) \sum_{Y \quad S} P_{Y,X}$.	(26)
SC-Conf	$\{x_i, r_1(x_i), \ldots, r_K(x_i)\}_{i=1}^n$, where $x_i \overset{\text{i.i.d.}}{\sim} P_{X\lvert Y=y_s}$ with $y_s \quad [K]$, $r_k(x_i) := P_{Y=k\lvert X=x_i}$ for each $k \quad [K]$.	(28)
Sub-Conf	$\{x_i, r_1(x_i), \ldots, r_K(x_i)\}_{i=1}^n$, where $x_i \overset{\text{i.i.d.}}{\sim} P_{X\lvert Y \quad Y_s}$ with $Y_s \quad [K]$, $r_k(x_i) := P_{Y=k\lvert X=x_i}$ for each $k \quad [K]$.	(30)
Soft-label	$\{x_i, r_1(x_i), \ldots, r_K(x_i)\}_{i=1}^n$, where $x_i \overset{\text{i.i.d.}}{\sim} P_X$, $r_k(x_i) := P_{Y=k\lvert X=x_i}$ for each $k \quad [K]$.	(33)

Table 4: Risk rewrites for binary WSLs.

WSL	Risk rewrite for $R(g) = \mathsf{E}_{Y,X}\left[\ell_Y(g(X))\right]$ (1)	Equation
PU	$R(g) = \mathsf{E}_{\mathrm{P}}\left[\pi_{\mathrm{p}}\ell_{\mathrm{p}} - \pi_{\mathrm{p}}\ell_{\mathrm{n}}\right] + \mathsf{E}_{\mathrm{U}}\left[\ell_{\mathrm{n}}\right].$	(5)
Pconf	$R(g) = \pi_{\mathrm{p}}\mathsf{E}_{\mathrm{P}}\left[\ell_{\mathrm{p}} + \dfrac{1 - r(X)}{r(X)}\ell_{\mathrm{n}}\right].$	(7)
UU	$R(g) = \mathsf{E}_{\mathrm{U}_1}\left[\dfrac{(1-\gamma_2)\pi_{\mathrm{p}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{p}} + \dfrac{-\gamma_2\pi_{\mathrm{n}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{n}}\right] + \mathsf{E}_{\mathrm{U}_2}\left[\dfrac{-\gamma_1\pi_{\mathrm{p}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{p}} + \dfrac{(1-\gamma_1)\pi_{\mathrm{n}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{n}}\right].$	(9)
SU	$R(g) = \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\mathsf{E}_{\mathrm{S}}\left[\dfrac{L(X) + L(X\,)}{2}\right] + \mathsf{E}_{\mathrm{U}}\left[L_{-}(X)\right],$ where $L(X) := \dfrac{1}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) - \dfrac{1}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X),$ $L_{-}(X) := -\dfrac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) + \dfrac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X).$	(11)
DU	$R(g) = 2\pi_{\mathrm{p}}\pi_{\mathrm{n}}\mathsf{E}_{\mathrm{D}}\left[-\dfrac{L(X) + L(X\,)}{2}\right] + \mathsf{E}_{\mathrm{U}}\left[L_{+}(X)\right],$ where $L(X)$ is defined in the SU setting, and $L_{+}(X) := \dfrac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) - \dfrac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X).$	(13)
SD	$R(g) = \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\mathsf{E}_{\mathrm{S}}\left[\dfrac{L_{+}(X) + L_{+}(X\,)}{2}\right] + 2\pi_{\mathrm{p}}\pi_{\mathrm{n}}\mathsf{E}_{\mathrm{D}}\left[\dfrac{L_{-}(X) + L_{-}(X\,)}{2}\right],$ where $L_{+}(X)$ and $L_{-}(X\,)$ are defined in the SU and DU settings.	(15)
Pcomp	$R(g) = \mathsf{E}_{\mathrm{Sup}}\left[\ell_{\mathrm{p}} - \pi_{\mathrm{p}}\ell_{\mathrm{n}}\right] + \mathsf{E}_{\mathrm{Inf}}\left[-\pi_{\mathrm{n}}\ell_{\mathrm{p}} + \ell_{\mathrm{n}}\right],$ where $P_{\mathrm{Sup}} := \int_x {}_X P_{\mathrm{PC}}\,\mathrm{d}x\,,$ $P_{\mathrm{Inf}} := \int_x {}_X P_{\mathrm{PC}}\,\mathrm{d}x.$	(17)
Sconf	$R(g) = \mathsf{E}_{X,X}\left[\dfrac{r(X, X\,) - \pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\dfrac{\ell_{\mathrm{p}}(X) + \ell_{\mathrm{p}}(X\,)}{2} + \dfrac{\pi_{\mathrm{p}} - r(X, X\,)}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\dfrac{\ell_{\mathrm{n}}(X) + \ell_{\mathrm{n}}(X\,)}{2}\right].$	(19)

Table 5: Risk rewrites for multiclass WSLs.

WSL	Risk rewrite for $R(g) = \mathsf{E}_{Y,X}\left[\ell_Y(g(X))\right]$ (1)	Equation
CL	$R(g) = \mathsf{E}_{\bar{S},X}\left[\sum_{y=1}^{K} \ell_y - (K-1)\ell_{\bar{S}}\right].$	(21)
MCL	$R(g) = \sum_{d=1}^{K-1} \boldsymbol{P}_{\lvert\bar{S}\rvert=d}\, \mathsf{E}_{\bar{S},X\lvert\lvert\bar{S}\rvert=d}\left[\sum_{y/\bar{S}} \ell_y - \frac{K-1-\lvert\bar{S}\rvert}{\lvert\bar{S}\rvert}\sum_{\bar{s}\ \bar{S}} \ell_{\bar{s}}\right].$	(23)
PCPL	$R(g) = \frac{1}{2}\mathsf{E}_{S,X}\left[\sum_{y=1}^{K} \frac{\boldsymbol{P}_{Y=y\lvert X}}{\sum_{a\ \ S} \boldsymbol{P}_{Y=a\lvert X}}\ell_y\right].$	(25)
PPL	$R(g) = \mathsf{E}_{S,X}\left[\sum_{y\ \ S} \frac{\boldsymbol{P}_{Y=y\lvert X}}{\sum_{a\ \ S} \boldsymbol{P}_{Y=a\lvert X}}\ell_y\right].$	(27)
SC-Conf	$R(g) = \pi_{y_{\mathrm{s}}}\mathsf{E}_{X\lvert Y=y_{\mathrm{s}}}\left[\sum_{y=1}^{K} \frac{r_y(X)}{r_{y_{\mathrm{s}}}(X)}\ell_y\right].$	(29)
Sub-Conf	$R(g) = \pi_{Y_{\mathrm{s}}}\mathsf{E}_{X\lvert Y\ \ Y_{\mathrm{s}}}\left[\sum_{y=1}^{K} \frac{r_y(X)}{r_{Y_{\mathrm{s}}}(X)}\ell_y\right].$	(31)
Soft-label	$R(g) = \mathsf{E}_X\left[\sum_{y=1}^{K} r_y(X)\ell_y\right].$	(34)

From the above tables, finding a way to reexpress the classification risk $R(g)$ (1) in terms of the data-generating distributions becomes the crux when applying ERM for most WSL studies. The rewrites also replace loss functions ℓ_Y defining (1) with various modified losses (shown inside the expectations). These modified loss functions are sometimes called corrected losses, which is why the approach is also called loss correction. Proposing a generic methodology that finds properly corrected losses to achieve risk rewrite in different scenarios is a main topic we would like to elaborate on in this paper.

2.2.17 Learning with Noisy Labels (LNL) Formulations

Next, we review two related formulations in LNL, the MCD and CCN settings, in Table 6. The observed instances in MCD and CCN are still labeled by $\{\mathrm{p}, \mathrm{n}\}$ but are polluted by certain noise models. We use $\bar{Y}$ to represent a polluted label, compared to an unpolluted Y. In MCD, a small portion of the negatively labeled data $\gamma_{\mathrm{p}}\boldsymbol{P}_{X|Y=\mathrm{n}}$ contaminates the positively labeled data $\boldsymbol{P}_{X|Y=\mathrm{p}}$. Likewise, a small portion of the positive data $\gamma_{\mathrm{n}}\boldsymbol{P}_{X|Y=\mathrm{p}}$ contaminates the negatively labeled data $\boldsymbol{P}_{X|Y=\mathrm{n}}$ (Scott et al., 2013). In the CCN setting, a label Y is flipped to become $\bar{Y}$ with probability $\boldsymbol{P}_{\bar{Y}|Y,X}$ (Natarajan et al., 2013). Although they are formulated for the study of noisy labels, their formulations share similar structures with many WSLs above. In Section 4, we will use the similarities to categorize WSLs and provide a bird's eye view to reveal connections among WSLs.

Table 6: MCD and CCN formulations.

Scenario	Formulation
MCD	$\{x_i^{\bar{\mathrm{p}}}\}_{i=1}^{n_{\bar{\mathrm{p}}}} \overset{\text{i.i.d.}}{\sim} \boldsymbol{P}_{X\mid\bar{Y}=\mathrm{p}} := (1-\gamma_{\mathrm{p}})\,\boldsymbol{P}_{X\mid Y=\mathrm{p}} + \gamma_{\mathrm{p}}\,\boldsymbol{P}_{X\mid Y=\mathrm{n}}$. $\{x_j^{\bar{\mathrm{n}}}\}_{j=1}^{n_{\bar{\mathrm{n}}}} \overset{\text{i.i.d.}}{\sim} \boldsymbol{P}_{X\mid\bar{Y}=\mathrm{n}} := \gamma_{\mathrm{n}}\,\boldsymbol{P}_{X\mid Y=\mathrm{p}} + (1-\gamma_{\mathrm{n}})\,\boldsymbol{P}_{X\mid Y=\mathrm{n}}$.
CCN	$\{(\bar{y}_i, x_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \boldsymbol{P}_{\bar{Y}=\bar{y}_i,X} := \sum_{k\ \{\mathrm{p},\mathrm{n}\}} \boldsymbol{P}_{\bar{Y}=\bar{y}_i\mid Y=k,X}\boldsymbol{P}_{Y=k,X},\quad \bar{y}_i \quad \{\mathrm{p},\mathrm{n}\}$.

3 A Framework for Risk Rewrite

We illustrate the proposed framework in this section. Its job is to provide a unified treatment and understanding of WSL. It consists of a formulation component and an analysis component. The analysis component suggests a generic methodology to solve the risk rewrite problem. Moreover, diving into the formulation component's logic, we can interpret multiple WSL formulations and the diverse risk rewrites from a single perspective.

Before introducing the framework, we first define several abstract notations that will be used throughout the paper. There are three main characters and one supporting character in the framework. The main characters are the vector of data-generating distributions $\bar{P}$, the vector of risk-defining distributions P, and the vector of base distributions B. The supporting character is the vector of loss functions L. The reason for using vectorized pseudonyms is that the proposed framework uses matrix multiplication as a basic mathematical operation. $\bar{P}$ contains distributions that produce the observational data. For instance, $\bar{P} = \binom{\boldsymbol{P}_{\mathrm{P}}}{\boldsymbol{P}_{\mathrm{U}}}$ in PU learning (Table 2) and $\boldsymbol{P}_{\bar{S}=k,X}$ is the k-th entry of $\bar{P}$ in CL learning (Table 3). We use classification risk (1) to illustrate our framework. So P consists of joint distributions $\boldsymbol{P}_{X,Y}$. We can look at Tables 2 and 3 and see that there are basic elements that define a data-generating distribution. These are class-conditionals $\boldsymbol{P}_{X|Y}$ in Table 2 and joint distributions $\boldsymbol{P}_{X,Y}$ in Table 3. Since these basic elements do not necessarily coincide with the entries of P, we denote them as B. L consists of loss functions and its k-th entry is $\ell_{Y=k}(g(X))$[2].

3.1 The Formulation Component of the Framework

The construction of the formulation component is to study the connections among WSLs and provide a foundation for developing the generic methodology. We draw inspiration from Section 2.2. Each WSL formulation represents a type of weaken information of the joint distribution $\boldsymbol{P}_{Y,X}$ in supervised learning. For instance, unlabeled data discards the label information (Lu et al., 2020; 2021), the complementary-label is a label that cannot be the ground truth (Ishida et al., 2017; Yu et al., 2018), and the similarity encodes a comparative relationship of two ground truth labels (Bao et al., 2018; Shimada et al., 2021; Cao et al., 2021b). Thus, we are motivated to search for a general way to link data-generating distributions with the joint distribution.

We start by linking the data-generating distributions $\bar{P}$ and the base distributions B. This involves finding matrix correspondences to Tables 2 and 3. We assume that a matrix M_{corr} formalizes the link:

$$\bar{P} = M_{\mathrm{corr}} B. \tag{35}$$

Taking PU learning (4) for example, M_{corr} aims to connect $\bar{P} = \binom{\boldsymbol{P}_{\mathrm{P}}}{\boldsymbol{P}_{\mathrm{U}}}$ with $B = \binom{\boldsymbol{P}_{X\mid Y=\mathrm{p}}}{\boldsymbol{P}_{X\mid Y=\mathrm{n}}}$. To keep the framework as abstract as possible, we would like to defer the discussion of all other $\bar{P}$ and B until we realize their corresponding M_{corr} in Section 4.

The matrix formulation has two advantages. First, it provides a unified way to characterize a wide range of WSL settings. By studying the entries of a matrix, we can easily link one WSL scenario to another to form

[2]We reserve $\bar{P}$, P, and B for vectors of distributions and L and $\bar{L}$ for vectors of loss functions. We address them as "the distributions" and "the losses" to avoid the verbose "the vector of distributions/losses."

reduction graphs of WSLs. As the first main topic of this work, Section 4 shows, for a given WSL setting, how to find the corresponding matrix M_{corr}, and Tables 7 – 9 summarize fifteen WSL settings covered by our matrix formulation and depict a reduction graph rooted from M_{corr}. The following subsection illustrates the second advantage of aiding the construction of a generic methodology for conducting risk rewrite.

3.2 The Analysis Component of the Framework

Note that the conventional expression (1) can be simplified, by the inner product, to be $R(g) = \int_{x \ \mathsf{X}} L \ \ P\mathrm{d}x$. It is immediately possible to rewrite the risk under data-generating distributions $\bar{P}$ by showing $L \ \ P = \bar{L} \ \ \bar{P}$, where $\bar{L}$ is called the vector of corrected losses and its role will be clarified later. Therefore, it is imperative to establish the connection between $\bar{P}$ and P. The goal can be achieved by linking B and P since we have assumed that $\bar{P} = M_{\text{corr}}B$ in the previous subsection. Recall from the beginning of this section that the base distributions B are of the forms $\mathsf{P}_{X|Y}$ or $\mathsf{P}_{X,Y}$, and the risk-defining distributions P are of the form $\mathsf{P}_{X,Y}$. Given their label-relevant nature (i.e., they are either the joint distribution or the class-conditionals), we assume that there exists a transformation matrix M_{trsf} that satisfies $B = M_{\text{trsf}}P$. Thus, (35) becomes

$$\bar{P} = M_{\text{corr}}M_{\text{trsf}}P. \tag{36}$$

Having the freedom to choose M_{trsf} allows the framework to handle different base distributions B and to adapt to various performance measures that define P, as we will discuss in Sections 5.1.1, 5.2.1, and 6.1, respectively.

The reason why connecting P with $\bar{P}$ (36) helps the construction of the corrected losses is that if we manage to find a way to compensate for the combined effect of M_{corr} and M_{trsf}, we can implement the compensation mechanism on the "corrected" losses $\bar{L}$. Specifically, suppose there exists a matrix $M_{\text{corr}}^{\dagger}$ satisfying

$$P = M_{\text{corr}}^{\dagger}\bar{P}. \tag{37}$$

Then, the corrected losses defined by

$$\bar{L} \ \ := L \ \ M_{\text{corr}}^{\dagger} \tag{38}$$

allows us to rephrase the classification risk as

$$\begin{aligned} \int_{x \ \mathsf{X}} \bar{L} \ \ \bar{P}\,\mathrm{d}x \ \ &= \ \int_{x \ \mathsf{X}} L \ \ M_{\text{corr}}^{\dagger}\bar{P}\,\mathrm{d}x \\ &= \ \int_{x \ \mathsf{X}} L \ \ P\,\mathrm{d}x \ = R(g), \end{aligned} \tag{39}$$

providing a rewrite for $R(g)$ with respect to $\bar{P}$.

The above procedure describes a generic methodology for the risk rewrite problem. As the second main topic, we instantiate the framework by presenting the corresponding matrices $M_{\text{corr}}^{\dagger}$ and M_{trsf} for each learning scenario in Section 5 to demonstrate its applicability.

3.3 Intuition of the Framework

The key equations discussed in Sections 3.1 and 3.2 are

$$\begin{gathered} \bar{P} \overset{(35)}{=} M_{\text{corr}}B \overset{(36)}{=} M_{\text{corr}}M_{\text{trsf}}P, \\ \bar{L} \ \ \bar{P} \overset{(38)}{=} L \ \ M_{\text{corr}}^{\dagger}M_{\text{corr}}M_{\text{trsf}}P \overset{(37)}{=} L \ \ P. \end{gathered}$$

The logic behind them is succinct and interpretive. First, from a formulation perspective, viewing matrix M_{corr} as a contamination matrix that corrupts the base B to become the contaminated $\bar{P}$ (35), we interpret this *contamination mechanism* as sacrificing certain information in exchange for certain saved costs or privacy, reflecting the essence underlying WSL formulations. In addition to formulating the data-generating

mechanism, the link between B and the risk-defining distributions P (36) connects $\bar{P}$ and P to motivate the methodology design. This novel viewpoint of connecting the data distributions via the explicit two-stage formulation facilitates the unification work in this paper.

Second, regarding the methodological design, it becomes easier to devise a countermeasure when the connection between $\bar{P}$ and P is in good shape. The design of $\bar{L} = L\ M^{\dagger}_{\text{corr}}$ involves $M^{\dagger}_{\text{corr}}$, which captures a common idea behind risk rewrite: Restoration of the risk-defining distributions and the original loss functions is accomplished by the *decontamination* (37) provided by $\bar{L}$. Furthermore, the instantiations of $\bar{L} = L\ M^{\dagger}_{\text{corr}}$ (38) justify that the apparently different forms of corrected losses reported in the literature (i.e., referred papers that contribute to Tables 4 and 5, and those referred to as recoveries in Section 5) essentially stem from $M^{\dagger}_{\text{corr}}$. In summary, the proposed framework is abstract and flexible enough that we use it in the current paper to formulate the contamination mechanisms and provide a generic methodology for a wide range of WSLs.

3.4 Building Blocks: The Inversion and the Marginal Chain Approaches

We describe two building blocks, the inversion method and the marginal chain method, that will be used to devise $M^{\dagger}_{\text{corr}}$ that satisfies (37) in each scenario we study later.

Proposition 1 (The inversion method)**.** *Let P and $\bar{P}$ be vectors. Suppose $\bar{P} = MP$ holds for an invertible matrix M. Then, choosing $M^{\dagger}_{\text{corr}} = M^{-1}$, we have $P = M^{\dagger}_{\text{corr}}\bar{P}$.*

Proof. For any invertible M, it is easy to see that, by assigning $M^{\dagger}_{\text{corr}} = M^{-1}$, one has

$$M^{\dagger}_{\text{corr}}\bar{P} = M^{-1}\bar{P} = M^{-1}MP = P.$$

□

We remark that this simple strategy was adopted in many LNL works. A handful of related papers are Cid-Sueiro (2012), Blanchard & Scott (2014), Menon et al. (2015), van Rooyen & Williamson (2015), Patrini et al. (2017), van Rooyen & Williamson (2017), and Katz-Samuels et al. (2019). Hence, it can be applied to WSLs that are special cases of certain LNL scenarios.

Proposition 2 (The marginal chain method)**.** *Let $Y = k \quad [K]$ be a class label, where $[K]$ is the set of classes associated with the classification risk. Let $S = \{s_1, s_2, \ldots, s_{|S|}\} \quad 2^{[K]}$ be the set of class sets and S be the random variable of the observational outcome. Denote*

$$P = \begin{pmatrix} P_{Y=1,X} \\ \vdots \\ P_{Y=K,X} \end{pmatrix} \text{ and } \bar{P} = \begin{pmatrix} P_{S=s_1,X} \\ \vdots \\ P_{S=s_{|S|},X} \end{pmatrix}.$$

Then,

$$M = \begin{pmatrix} P_{S=s_1|Y=1,X} & P_{S=s_1|Y=2,X} & \cdots & P_{S=s_1|Y=K,X} \\ P_{S=s_2|Y=1,X} & P_{S=s_2|Y=2,X} & \cdots & P_{S=s_2|Y=K,X} \\ \vdots & \vdots & \ddots & \vdots \\ P_{S=s_{|S|}|Y=1,X} & P_{S=s_{|S|}|Y=2,X} & \cdots & P_{S=s_{|S|}|Y=K,X} \end{pmatrix} \tag{40}$$

satisfies $\bar{P} = MP$, and

$$M^{\dagger}_{\text{corr}} = \begin{pmatrix} P_{Y=1|S=s_1,X} & P_{Y=1|S=s_2,X} & \cdots & P_{Y=1|S=s_{|S|},X} \\ P_{Y=2|S=s_1,X} & P_{Y=2|S=s_2,X} & \cdots & P_{Y=2|S=s_{|S|},X} \\ \vdots & \vdots & \ddots & \vdots \\ P_{Y=K|S=s_1,X} & P_{Y=K|S=s_2,X} & \cdots & P_{Y=K|S=s_{|S|},X} \end{pmatrix} \tag{41}$$

satisfies $P = M_{\text{corr}}^{\dagger}\bar{P}$.

The role of S is to represent a weak supervision that encodes some combinatorial information about the unobservable true label Y. We will discuss this concept in detail in Sections 4.2 and 5.2.

Proof. It suffices to show $\left(MP\right)_j = \bar{P}_j$ for any $j \quad [|\mathsf{S}|]$. Taking the inner product of the j-th row of M and P, we have

$$\sum_{k=1}^{K} \mathsf{P}_{S=s_j|Y=k,X}\mathsf{P}_{Y=k,X} = \sum_{k=1}^{K} \mathsf{P}_{S=s_j,Y=k,X} = \mathsf{P}_{S=s_j,X}$$

that verifies (40).

Next, we prove $P = M_{\text{corr}}^{\dagger}\bar{P}$ by showing $\left(M_{\text{corr}}^{\dagger}\bar{P}\right)_i = P_i$. For each $i \quad [K]$,

$$\begin{aligned}
\left(M_{\text{corr}}^{\dagger}\bar{P}\right)_i = \left(M_{\text{corr}}^{\dagger}MP\right)_i \quad &= \quad \sum_{j=1}^{|\mathsf{S}|} \mathsf{P}_{Y=i|S=s_j,X} \sum_{k=1}^{K} \mathsf{P}_{S=s_j|Y=k,X}\mathsf{P}_{Y=k,X} \\
&\overset{\text{(a)}}{=} \quad \sum_{j=1}^{|\mathsf{S}|} \mathsf{P}_{Y=i|S=s_j,X}\mathsf{P}_{S=s_j,X} \\
&\overset{\text{(b)}}{=} \quad \mathsf{P}_{Y=i,X} = P_i. \qquad (42)
\end{aligned}$$

□

Besides finding the inverse matrix, we propose a new approach called the *marginal chain* to achieve (37). The development of this approach begins with the observation that $\mathsf{P}_{S=s_j,X}$ in $\bar{P} = MP$ is a distribution where Y is marginalized out. It inspires an idea that one could perform another marginalization to restore the original distribution $\mathsf{P}_{Y,X}$; specifically, by marginalizing out S. The design of $M_{\text{corr}}^{\dagger}$ in (41) aims to carry out the idea. As shown by (a) and (b) in the proof, two consecutive marginalization steps on Y and then S give the name of the marginal chain.

Both the inversion and marginal chain methods have strengths and weaknesses. The inversion method only requires P as a real vector but needs the invertible assumption on the contamination matrix M. In contrast, the marginal chain method exploits that P, in fact, is a distributional vector, allowing it to find a decontamination matrix $M_{\text{corr}}^{\dagger}$ even for a non-invertible M. A restriction of the marginal chain method is that the construction of $M_{\text{corr}}^{\dagger}$ is regulated by probability equations.

We are ready to justify the proposed framework through the following two sections. Section 4 discusses weakly supervised scenarios that can be subsumed by the formulation component (35). Section 5 verifies the analysis component by instantiating (38) to conduct the risk rewrite for each scenario mentioned in Section 4. In both sections, we divide the scenarios into three categories. The first two are WSLs that can be viewed as special cases in either the prevalent MCD or CCN settings. The third category contains confidence-based scenarios. The notations listed in Table 1 will still be functional. For all notations and their abbreviations required in the coming sections, please refer to Appendix B.

4 Contamination as Weak Supervision

In this section, we instantiate the contamination matrix for each weakly supervised scenario listed in Table 2 and Table 3. Tables 7 – 9 summarize the contamination matrices developed in this section. Each table also represents a reduction graph of WSL settings. These reduction graphs cluster WSL settings into three main categories, providing a hierarchy of relationships. With this hierarchy, we can understand, compare with, and relate to different settings or even grow the hierarchy by adding new branches. Next are the notations

for reading the graphs. For two contamination mechanisms, U and V, we use $M_{\mathrm{U}} \to M_{\mathrm{V}}$ to denote "M_{U} is reduced to M_{V}" or "M_{U} is realized as M_{V}", and $M_{\mathrm{U}} \quad M_{\mathrm{V}}$ means "M_{U} is generalized to M_{V}".

The proposed framework provides a generic strategy for formulating multiple weakly supervised scenarios. Thus, the proofs will have a certain degree of similarity. To avoid repeating similar proofs, we provide proofs that appear for the first time. For auxiliary lemmas and results whose proofs are similar to the previous ones, we refer to the omitted proofs in Appendix C. In particular, the omitted proofs in Section 4.1 can be found in Appendix C.1, and those in Section 4.2 can be found in Appendix C.2.

Table 7: Contamination matrices of MCD category in Section 4.1.

WSLs	Entry Parameter	Contamination Matrix	Reduction path
MCD	γ_{p}, γ_{n}	M_{MCD} (45)	$M_{\mathrm{corr}} \to M_{\mathrm{MCD}}$
UU	γ_1, γ_2	M_{UU} (49)	$M_{\mathrm{corr}} \to M_{\mathrm{UU}} \approx M_{\mathrm{MCD}}$
PU	$\gamma_1 = 0$, $\gamma_2 = \pi_{\mathrm{p}}$	M_{PU} (50)	$M_{\mathrm{UU}} \to M_{\mathrm{PU}}$
SU	$\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2}$, $\gamma_2 = \pi_{\mathrm{p}}$	M_{SU} (51)	$M_{\mathrm{UU}} \to M_{\mathrm{SU}}$
Pcomp	$\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2}$, $\gamma_2 = \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}}$	M_{Pcomp} (52)	$M_{\mathrm{UU}} \to M_{\mathrm{Pcomp}}$
DU	$\gamma_1 = 1/2$, $\gamma_2 = \pi_{\mathrm{p}}$	M_{DU} (53)	$M_{\mathrm{UU}} \to M_{\mathrm{DU}}$
SD	$\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2}$, $\gamma_2 = 1/2$	M_{SD} (54)	$M_{\mathrm{UU}} \to M_{\mathrm{SD}}$
Sconf	–	M_{Sconf} (55)	$M_{\mathrm{corr}} \to M_{\mathrm{Sconf}}$

Table 8: Contamination matrices of CCN category in Section 4.2.

WSLs	Entry Parameter	Contamination Matrix	Reduction path
CCN	$\mathsf{P}_{\bar{Y}\mid Y,X}$ (59)	M_{CCN} (60)	$M_{\mathrm{corr}} \to M_{\mathrm{CCN}}$
Generalized CCN	$\mathsf{P}_{S\mid Y,X}$ (61)	M_{gCCN} (64)	$M_{\mathrm{corr}} \to M_{\mathrm{CCN}} \quad M_{\mathrm{gCCN}}$
PPL	$C(S,X)\mathsf{I}\,[Y \quad S]$ (65)	M_{PPL} (66)	$M_{\mathrm{gCCN}} \to M_{\mathrm{PPL}}$
PCPL	$\frac{1}{2^{K-1}-1}\mathsf{I}\,[Y \quad S]$	M_{PCPL} (68)	$M_{\mathrm{gCCN}} \to M_{\mathrm{PPL}} \to M_{\mathrm{PCPL}}$
MCL	$\frac{q_{\lvert\bar{S}\rvert}}{\binom{K-1}{\lvert\bar{S}\rvert}}\mathsf{I}\left[Y \;/\; \bar{S}\right]$ (75)	M_{MCL} (71)	$M_{\mathrm{gCCN}} \to M_{\mathrm{PPL}} \to M_{\mathrm{MCL}}$
CL	$\lvert S\rvert = 1$, $\frac{1}{K-1}\mathsf{I}\,[Y \quad S]$	M_{CL} (76)	$M_{\mathrm{gCCN}} \to M_{\mathrm{PPL}} \to M_{\mathrm{MCL}} \to M_{\mathrm{CL}}$

Table 9: Contamination matrices of confidence-based category in Section 4.3.

WSLs	Entry Parameter	Contamination Matrix	Reduction path
Sub-Conf	$\frac{\mathsf{P}_{Y \quad Y_{\mathrm{s}}\mid X}}{\mathsf{P}_{Y=k\mid X}}$	M_{Sub} (80)	$M_{\mathrm{corr}} \to M_{\mathrm{Sub}}$
SC	$\mathsf{Y}_{\mathrm{s}} = \{y_{\mathrm{s}}\}$ in M_{Sub}	M_{SC} (81)	$M_{\mathrm{Sub}} \to M_{\mathrm{SC}}$
Pconf	$K = 2$, $y_{\mathrm{s}} = \mathrm{p}$ in M_{SC}	M_{Pconf} (82)	$M_{\mathrm{Sub}} \to M_{\mathrm{SC}} \to M_{\mathrm{Pconf}}$
Soft	$\frac{1}{\mathsf{P}_{Y=k\mid X}}$	M_{Soft} (84)	$M_{\mathrm{Sub}} \to M_{\mathrm{Soft}}$

4.1 MCD Scenarios

As listed in Table 6, in binary classification, the MCD model (Menon et al., 2015) corrupts the clean class-conditionals $\boldsymbol{P}_{X|Y=\mathrm{p}}$ and $\boldsymbol{P}_{X|Y=\mathrm{n}}$ via parameters γ_{p} and γ_{n} as follows:

$$\begin{aligned} \boldsymbol{P}_{X|\bar{Y}=\mathrm{p}} &:= (1 - \gamma_{\mathrm{p}})\, \boldsymbol{P}_{X|Y=\mathrm{p}} + \gamma_{\mathrm{p}}\, \boldsymbol{P}_{X|Y=\mathrm{n}}, \\ \boldsymbol{P}_{X|\bar{Y}=\mathrm{n}} &:= \gamma_{\mathrm{n}}\, \boldsymbol{P}_{X|Y=\mathrm{p}} + (1 - \gamma_{\mathrm{n}})\, \boldsymbol{P}_{X|Y=\mathrm{n}}, \end{aligned} \tag{43}$$

where $\gamma_{\mathrm{p}}, \gamma_{\mathrm{n}} \quad [0, 1]$ and $\gamma_{\mathrm{p}} + \gamma_{\mathrm{n}} < 1$. Viewing the contamination targets $\boldsymbol{P}_{X|Y=\mathrm{p}}$ and $\boldsymbol{P}_{X|Y=\mathrm{n}}$ as the base distributions

$$B := \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix}$$

and denoting the vector of data-generating distributions as

$$\bar{P} := \begin{pmatrix} \boldsymbol{P}_{X|\bar{Y}=\mathrm{p}} \\ \boldsymbol{P}_{X|\bar{Y}=\mathrm{n}} \end{pmatrix},$$

we can express (43) in the following matrix form

$$\begin{pmatrix} \boldsymbol{P}_{X|\bar{Y}=\mathrm{p}} \\ \boldsymbol{P}_{X|\bar{Y}=\mathrm{n}} \end{pmatrix} = \begin{pmatrix} 1 - \gamma_{\mathrm{p}} & \gamma_{\mathrm{p}} \\ \gamma_{\mathrm{n}} & 1 - \gamma_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix}. \tag{44}$$

Comparing (44) with $\bar{P} = M_{\mathrm{corr}} B$ (35), we find that the contamination matrix M_{corr} is realized as

$$M_{\mathrm{MCD}} := \begin{pmatrix} 1 - \gamma_{\mathrm{p}} & \gamma_{\mathrm{p}} \\ \gamma_{\mathrm{n}} & 1 - \gamma_{\mathrm{n}} \end{pmatrix} \tag{45}$$

in the MCD setting.

4.1.1 Unlabeled-Unlabeled (UU) Learning (Lu et al., 2019)

Next, we show how to characterize UU learning by a contamination matrix. Naming

$$\pi_{\mathrm{p}}\, \boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}\, \boldsymbol{P}_{X|Y=\mathrm{n}}$$

as $\boldsymbol{P}_{\mathrm{U}}$ is feasible since $\pi_{\mathrm{p}}\, \boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}\, \boldsymbol{P}_{X|Y=\mathrm{n}} = \boldsymbol{P}_X$ generates data that statistically equals to data sampled from $\boldsymbol{P}_{Y,X}$ with labels removed. Viewing π_{p} as the mixture rate of samples from $\boldsymbol{P}_{X|Y=\mathrm{p}}$ and $\boldsymbol{P}_{X|Y=\mathrm{n}}$, $\boldsymbol{P}_{\mathrm{U}}$ is parameterized by π_{p}. Therefore, we can interpret (8), recalled as follows, as formulating two unlabeled data distributions w.r.t. mixture rates $(1 - \gamma_1)$ and γ_2, respectively:

$$\begin{aligned} \boldsymbol{P}_{\mathrm{U}_1} &= (1 - \gamma_1)\, \boldsymbol{P}_{X|Y=\mathrm{p}} + \gamma_1\, \boldsymbol{P}_{X|Y=\mathrm{n}}, \\ \boldsymbol{P}_{\mathrm{U}_2} &= \gamma_2\, \boldsymbol{P}_{X|Y=\mathrm{p}} + (1 - \gamma_2)\, \boldsymbol{P}_{X|Y=\mathrm{n}}. \end{aligned}$$

Taking the class-conditionals as the base distributions

$$B := \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix} \tag{46}$$

and converting (8) to the matrix form, we express the data-generating distributions of UU learning

$$\bar{P} := \begin{pmatrix} \boldsymbol{P}_{\mathrm{U}_1} \\ \boldsymbol{P}_{\mathrm{U}_2} \end{pmatrix} \tag{47}$$

as

$$\begin{pmatrix} \boldsymbol{P}_{\mathrm{U}_1} \\ \boldsymbol{P}_{\mathrm{U}_2} \end{pmatrix} = \begin{pmatrix} 1 - \gamma_1 & \gamma_1 \\ \gamma_2 & 1 - \gamma_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix}, \tag{48}$$

and we arrive at the following lemma.

Lemma 3. *Let B (46) be the base distributions and $\bar{P}$ (47) be the data-generating distributions. For $\gamma_1, \gamma_2 \quad [0, 1]$ such that $\gamma_1 + \gamma_2 = 1$, the contamination matrix*

$$M_{\mathrm{UU}} := \begin{pmatrix} 1 - \gamma_1 & \gamma_1 \\ \gamma_2 & 1 - \gamma_2 \end{pmatrix} \tag{49}$$

characterizes the data-generating process of UU learning (8) via (48).

Comparing (48) with the formulation framework $\bar{P} = M_{\mathrm{corr}} B$ (35), we see that in UU learning, M_{corr} is realized as M_{UU}:

$$M_{\mathrm{corr}} \rightarrow M_{\mathrm{UU}}.$$

Like MCD, we assume $\gamma_1 + \gamma_2 = 1$. Our assumption is equivalent to that of MCD since the case of swapping P_{corr} and Q_{corr} in Menon et al. (2015) corresponds to $\gamma_1 + \gamma_2 > 1$ in our case. For details, refer to the discussion in Section 2.2 of Menon et al. (2015). The need for $\gamma_1 + \gamma_2 = 1$ can be explained by examining the entries in M_{UU}. The constraint $\gamma_1 + \gamma_2 = 1$ guarantees distinct rows in M_{UU}, implying the observed data sets are sampled from two distinct distributions. On the contrary, allowing $\gamma_1 + \gamma_2 = 1$ ends up observing one unlabeled data set (i.e., $\boldsymbol{P}_{\mathrm{U}_1} = \boldsymbol{P}_{\mathrm{U}_2}$) since $1 - \gamma_1 = \gamma_2$. Lu et al. (2019) proved in Section 3 that it is impossible to conduct a risk rewrite if one only observes one unlabeled data set.

Assigning $\gamma_1 = \gamma_{\mathrm{p}}$ and $\gamma_2 = \gamma_{\mathrm{n}}$ implies that MCD and UU have essentially the same data-generating process from the contamination perspective, as (44) and (48) have the identical right-hand sides (i.e., the same contamination targets and the same contamination matrix). However, they bear different meanings in respective research topics (i.e., distinct notions on the left-hand sides of the equations): In MCD, one still observes data with labels, nonetheless noisy, while in the UU setting, one observes two distinct unlabeled data sets. We use "$\approx$" to denote their relation in the UU row of Table 7.

Connecting UU learning with MCD, and later the generalized CCN with CCN in Section 4.2.1, allows us to categorize WSLs from the LNL perspective into Sections 4.1 and 4.2. In the rest of this subsection, we collect WSLs whose base distributions are class-conditionals and show M_{UU} instantiates their formulations via respective assignments of γ_1 and γ_2.

4.1.2 Positive-Unlabeled (PU) Learning (Kiryo et al., 2017)

Recall from (4) that $\boldsymbol{P}_{\mathrm{P}} = \boldsymbol{P}_{X|Y=\mathrm{p}}$ and $\boldsymbol{P}_{\mathrm{U}} = \boldsymbol{P}_X$. The following lemma describes the contamination matrix of PU learning.

Lemma 4. *Let B (46) be the base distributions and*

$$\bar{P} := \begin{pmatrix} \boldsymbol{P}_{\mathrm{P}} \\ \boldsymbol{P}_{\mathrm{U}} \end{pmatrix}$$

be the data-generating distributions. Define the contamination matrix

$$M_{\mathrm{PU}} := \begin{pmatrix} 1 & 0 \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix}. \tag{50}$$

Then, $\bar{P} = M_{\mathrm{PU}} B$, and M_{PU} characterizes the data-generating process of PU learning (4).

Proof. We apply the same proof strategy as in Lemma 3. By definitions,

$$M_{\mathrm{PU}}B = \begin{pmatrix} 1 & 0 \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix} = \begin{pmatrix} 1 \cdot \boldsymbol{P}_{X|Y=\mathrm{p}} + 0 \cdot \boldsymbol{P}_{X|Y=\mathrm{n}} \\ \pi_{\mathrm{p}} \cdot \boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}} \cdot \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix}.$$

Since $1 \cdot \boldsymbol{P}_{X|Y=\mathrm{p}} + 0 \cdot \boldsymbol{P}_{X|Y=\mathrm{n}} = \boldsymbol{P}_{X|Y=\mathrm{p}}$ and $\pi_{\mathrm{p}} \cdot \boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}} \cdot \boldsymbol{P}_{X|Y=\mathrm{n}} = \boldsymbol{P}_X$, we obtain $M_{\mathrm{PU}}B = \bar{P}$. □

Comparing with (35), we see that the contamination matrix M_{corr} is instantiated as M_{PU} (50) in PU learning. Further, M_{PU} can be obtained by assigning $\gamma_1 = 0$ and $\gamma_2 = \pi_{\mathrm{p}}$ in M_{UU} (49), and hence, we obtain the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{UU}} \to M_{\mathrm{PU}}.$$

4.1.3 Similar-Unlabeled (SU) Learning (Bao et al., 2018)

Recall $\boldsymbol{P}_{\mathrm{S}}$ (10) generates the pair of data points (x, x') who share the same label. In addition to the pairwise distribution $\boldsymbol{P}_{\mathrm{S}}$, a pointwise distribution

$$\boldsymbol{P}_{\tilde{\mathrm{S}}} = \frac{\pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|Y=\mathrm{p}}}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}$$

is also defined for single data point x (Bao et al., 2018, Lemma 1). Therefore, we choose $\boldsymbol{P}_{\tilde{\mathrm{S}}}$ as the data-generating distribution when constructing the contamination matrix in the following lemma.

Lemma 5. *Let B (46) be the base distributions and*

$$\bar{P} := \begin{pmatrix} \boldsymbol{P}_{\tilde{\mathrm{S}}} \\ \boldsymbol{P}_{\mathrm{U}} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{SU}} := \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix}, \tag{51}$$

which satisfies $\bar{P} = M_{\mathrm{SU}}B$, characterizes the data-generating distributions $\bar{P}$.

Further, M_{SU} can be obtained by assigning $\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2}$ and $\gamma_2 = \pi_{\mathrm{p}}$ in M_{UU} (49), and hence, we obtain the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{UU}} \to M_{\mathrm{SU}}.$$

4.1.4 Pairwise Comparison (Pcomp) Learning (Feng et al., 2021)

In SU learning, we formulate the pointwise data-generating distributions $\boldsymbol{P}_{\tilde{\mathrm{S}}}$ and $\boldsymbol{P}_{\mathrm{U}}$; likewise, we use the following pointwise distributions of $\boldsymbol{P}_{\mathrm{PC}}$ (16) to formulate Pcomp learning:

$$\begin{aligned} \boldsymbol{P}_{\mathrm{Sup}} &:= \int_x \int_{\mathcal{X}} \boldsymbol{P}_{\mathrm{PC}} \,\mathrm{d}x = \frac{\pi_{\mathrm{p}} \boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|Y=\mathrm{n}}}{\pi_{\mathrm{p}} + \pi_{\mathrm{n}}^2}, \\ \boldsymbol{P}_{\mathrm{Inf}} &:= \int_x \int_{\mathcal{X}} \boldsymbol{P}_{\mathrm{PC}} \,\mathrm{d}x = \frac{\pi_{\mathrm{p}}^2 \boldsymbol{P}_{X\,|Y=\mathrm{p}} + \pi_{\mathrm{n}} \boldsymbol{P}_{X\,|Y=\mathrm{n}}}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}}. \end{aligned}$$

Lemma 6. *Let B (46) be the base distributions and*

$$\bar{P} := \begin{pmatrix} P_{\mathrm{Sup}} \\ P_{\mathrm{Inf}} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{Pcomp}} := \begin{pmatrix} \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2} \\ \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}} & \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}} \end{pmatrix}, \tag{52}$$

which satisfies $\bar{P} = M_{\mathrm{Pcomp}} B$, characterizes the data-generating distributions $\bar{P}$.

Further, M_{Pcomp} can be obtained by assigning $\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2}$ and $\gamma_2 = \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}}$ in M_{UU} (49), and hence, we obtain the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{UU}} \to M_{\mathrm{Pcomp}}.$$

4.1.5 Similar-dissimilar-unlabeled (SDU) Learning (Shimada et al., 2021)

Dissimilar-unlabeled (DU) learning and similar-dissimilar (SD) learning are two critical components of SDU learning. Hence, we present the matrix formulations of M_{DU} and M_{SD}. Similar to the strategy taken in Sections 4.1.3 and 4.1.4, we use the following pointwise distribution

$$P_{\tilde{\mathrm{D}}} = \int_x P_{\mathrm{D}} \, \mathrm{d}x \; = \frac{P_{x|Y=\mathrm{p}} + P_{x|Y=\mathrm{n}}}{2}$$

(Shimada et al., 2021, (36) in Appendix A.1) in the following formulations.

We formulate the contamination matrix of DU learning via the following lemma.

Lemma 7. *Let B (46) be the base distributions and*

$$\bar{P} = \begin{pmatrix} P_{\tilde{\mathrm{D}}} \\ P_{\mathrm{U}} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{DU}} = \begin{pmatrix} 1/2 & 1/2 \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix}, \tag{53}$$

which satisfies $\bar{P} = M_{\mathrm{DU}} B$, characterizes the data-generating distributions $\bar{P}$.

Furthermore, since M_{UU} (49) reduces to M_{DU} by assigning $\gamma_1 = 1/2$ and $\gamma_2 = \pi_{\mathrm{p}}$, we have the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{UU}} \to M_{\mathrm{DU}}.$$

The next lemma formulates the contamination matrix of SD learning.

Lemma 8. *Let B (46) be the base distributions and*

$$\bar{P} = \begin{pmatrix} P_{\tilde{\mathrm{S}}} \\ P_{\tilde{\mathrm{D}}} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{SD}} = \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} \\ 1/2 & 1/2 \end{pmatrix} \tag{54}$$

which satisfies $\bar{P} = M_{\mathrm{SD}}B$, characterizes the data-generating distributions $\bar{P}$.

Moreover, because M_{UU} (49) reduces to M_{SD} via $\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2}$ and $\gamma_2 = 1/2$, we obtain the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{UU}} \to M_{\mathrm{SD}}.$$

4.1.6 Similarity-Confidence (Sconf) Learning (Cao et al., 2021b)

Recall from the Sconf setting (18) that (x, x') is a pair of data sampled i.i.d. from $P_{X,X'} := P_X P_{X'}$. On seeing P_X, one might wonder if it is sufficient to express the data-generating distribution simply as $P_X = P_{Y=\mathrm{p},X} + P_{Y=\mathrm{n},X}$. This approach, however correct, does not consider all available information in the Sconf setting. Similar to M_{UU} that uses parameters γ_1 and γ_2 to characterize the data-generating process in UU learning, we use the following lemma that includes the confidence $r(x, x') := P_{y=y'|x,x'}$ to characterize Sconf learning. Let us abbreviate $r(X, X')$ as r, $P_{X|Y=\mathrm{p}}$ as $P_{X|\mathrm{p}}$, and $P_{X'|Y=\mathrm{n}}$ as $P_{X'|\mathrm{n}}$.

Lemma 9. *Assume $\pi_{\mathrm{p}} = 1/2$. Let B (46) be the base distributions and*

$$\bar{P} := \begin{pmatrix} P_X P_{X'} \\ P_X P_{X'} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{Sconf}} := \begin{pmatrix} \frac{\pi_{\mathrm{p}}\left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}}\right)}{r-\pi_{\mathrm{n}}} & \frac{\pi_{\mathrm{p}}\left(\pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{p}}\right)}{r-\pi_{\mathrm{n}}} \\ \frac{\pi_{\mathrm{n}}\left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}}\right)}{\pi_{\mathrm{p}}-r} & \frac{\pi_{\mathrm{n}}\left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}}\right)}{\pi_{\mathrm{p}}-r} \end{pmatrix} \tag{55}$$

which satisfies $\bar{P} = M_{\mathrm{Sconf}}B$, characterizes the data-generating distributions $\bar{P}$.

Proof. Note that once

$$\left(\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}}\right) P_X P_{X'} = \left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}}\right) P_{X|\mathrm{p}} + \left(\pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{p}}\right) P_{X|\mathrm{n}} \tag{56}$$

and

$$\left(\frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{n}}}\right) P_X P_{X'} = \left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}}\right) P_{X|\mathrm{p}} + \left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}}\right) P_{X|\mathrm{n}}, \tag{57}$$

is obtained, reorganizing the terms gives

$$\begin{pmatrix} P_X P_{X'} \\ P_X P_{X'} \end{pmatrix} = \begin{pmatrix} \frac{\pi_{\mathrm{p}}\left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}}\right)}{r-\pi_{\mathrm{n}}} & \frac{\pi_{\mathrm{p}}\left(\pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{p}}\right)}{r-\pi_{\mathrm{n}}} \\ \frac{\pi_{\mathrm{n}}\left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}}\right)}{\pi_{\mathrm{p}}-r} & \frac{\pi_{\mathrm{n}}\left(\pi_{\mathrm{p}}^2 P_{X'|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{X'|\mathrm{n}}\right)}{\pi_{\mathrm{p}}-r} \end{pmatrix} \begin{pmatrix} P_{X|\mathrm{p}} \\ P_{X|\mathrm{n}} \end{pmatrix} \tag{58}$$

and finishes the proof.

Therefore, we will focus on proving (56) and (57). According to (2) of Cao et al. (2021b), the confidence $r(X, X')$, measuring how likely X and X' share the same label, is shown to be

$$r = r(X, X') = \frac{\pi_{\mathrm{p}}^2 P_{X|\mathrm{p}} P_{X'|\mathrm{p}} + \pi_{\mathrm{n}}^2 P_{X|\mathrm{n}} P_{X'|\mathrm{n}}}{P_X P_{X'}}.$$

It implies

$$r\boldsymbol{P}_X\boldsymbol{P}_X = \pi_{\mathrm{p}}^2\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{p}} + \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{n}}$$

and

$$(1-r)\boldsymbol{P}_X\boldsymbol{P}_X = \pi_{\mathrm{p}}\pi_{\mathrm{n}}\left(\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{n}} + \boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{p}}\right).$$

If $\pi_{\mathrm{p}} = 1/2$, $\pi_{\mathrm{p}} - r = 0$ and $r - \pi_{\mathrm{n}} = 0$. As a result, (56) is achieved as follows

$$\begin{aligned}
\left(\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}}\right)\boldsymbol{P}_X\boldsymbol{P}_X &= \left(r - \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}}(1-r)\right)\boldsymbol{P}_X\boldsymbol{P}_X \\
&= \pi_{\mathrm{p}}^2\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{p}} + \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{n}} - \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}}\pi_{\mathrm{p}}\pi_{\mathrm{n}}\left(\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{n}} + \boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{p}}\right) \\
&= \pi_{\mathrm{p}}^2\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{p}} - \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{n}} + \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{n}} - \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{p}}.
\end{aligned}$$

Also, (57) is achieved by having

$$\begin{aligned}
\left(\frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{n}}}\right)\boldsymbol{P}_X\boldsymbol{P}_X &= \left(\frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{n}}}(1-r) - r\right)\boldsymbol{P}_X\boldsymbol{P}_X \\
&= \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{n}}}\pi_{\mathrm{p}}\pi_{\mathrm{n}}\left(\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{n}} + \boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{p}}\right) - \pi_{\mathrm{p}}^2\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{p}} - \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{n}} \\
&= \pi_{\mathrm{p}}^2\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{n}} - \pi_{\mathrm{p}}^2\boldsymbol{P}_{X|\mathrm{p}}\boldsymbol{P}_{X|\mathrm{p}} + \pi_{\mathrm{p}}^2\boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{p}} - \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|\mathrm{n}}\boldsymbol{P}_{X|\mathrm{n}}.
\end{aligned}$$

□

The equality (58) implies that the inner product of the first row (resp. the second row) of M_{Sconf} and B represents a way (resp. another way) of obtaining $\boldsymbol{P}_X\boldsymbol{P}_X$. Although one might suspect that it is redundant to formulate $\boldsymbol{P}_X\boldsymbol{P}_X$ twice, we show in Section 5.1.6 this expression is crucial to rewrite the classification risk via the proposed framework. Furthermore, comparing $\bar{P} = M_{\mathrm{Sconf}}B$ (58) with $\bar{P} = M_{\mathrm{corr}}B$ (35), we have the reduction path

$$M_{\mathrm{corr}} \twoheadrightarrow M_{\mathrm{Sconf}}.$$

Note that M_{Sconf} does not fit the intuition of mutual contamination perfectly; we list Sconf learning in this subsection as all settings share the same base distributions B (46).

4.2 CCN Scenarios

The formulation component (35) also applies to the CCN model. Unlike MCD contaminating class-conditionals (distributions of X), CCN corrupts class probability functions (labeling distributions). Next, we show how to formulate CCN via (35) and extend the formulation to characterize diverse weakly supervised settings.

In binary classification, CCN (Natarajan et al., 2013; 2017) corrupts the labels by flipping the positive (resp. negative) labels with probability $\boldsymbol{P}_{\bar{Y}=\mathrm{n}|Y=\mathrm{p},X}$ (resp. $\boldsymbol{P}_{\bar{Y}=\mathrm{p}|Y=\mathrm{n},X}$). Specifically,

$$\begin{aligned}
\boldsymbol{P}_{\bar{Y}=\mathrm{p}|X} &:= \boldsymbol{P}_{\bar{Y}=\mathrm{p}|Y=\mathrm{p},X}\,\boldsymbol{P}_{Y=\mathrm{p}|X} + \boldsymbol{P}_{\bar{Y}=\mathrm{p}|Y=\mathrm{n},X}\,\boldsymbol{P}_{Y=\mathrm{n}|X}, \\
\boldsymbol{P}_{\bar{Y}=\mathrm{n}|X} &:= \boldsymbol{P}_{\bar{Y}=\mathrm{n}|Y=\mathrm{p},X}\,\boldsymbol{P}_{Y=\mathrm{p}|X} + \boldsymbol{P}_{\bar{Y}=\mathrm{n}|Y=\mathrm{n},X}\,\boldsymbol{P}_{Y=\mathrm{n}|X}
\end{aligned} \tag{59}$$

define the contaminated class probability functions. Taking the contamination targets $\boldsymbol{P}_{Y=\mathrm{p}|X}$ and $\boldsymbol{P}_{Y=\mathrm{n}|X}$ as the base distributions

$$B := \begin{pmatrix} \boldsymbol{P}_{Y=\mathrm{p}|X} \\ \boldsymbol{P}_{Y=\mathrm{n}|X} \end{pmatrix}$$

and denoting the label-generating distributions $P_{\bar{Y}=\mathrm{p}|X}$ and $P_{\bar{Y}=\mathrm{n}|X}$ as

$$\bar{P} := \begin{pmatrix} P_{\bar{Y}=\mathrm{p}|X} \\ P_{\bar{Y}=\mathrm{n}|X} \end{pmatrix},$$

we express (59) in the matrix form as follows

$$\begin{pmatrix} P_{\bar{Y}=\mathrm{p}|X} \\ P_{\bar{Y}=\mathrm{n}|X} \end{pmatrix} = \begin{pmatrix} P_{\bar{Y}=\mathrm{p}|Y=\mathrm{p},X} & P_{\bar{Y}=\mathrm{p}|Y=\mathrm{n},X} \\ P_{\bar{Y}=\mathrm{n}|Y=\mathrm{p},X} & P_{\bar{Y}=\mathrm{n}|Y=\mathrm{n},X} \end{pmatrix} \begin{pmatrix} P_{Y=\mathrm{p}|X} \\ P_{Y=\mathrm{n}|X} \end{pmatrix}.$$

Comparing with the abstract form $\bar{P} = M_{\mathrm{corr}} B$ (35), we see that

$$M_{\mathrm{CCN}} := \begin{pmatrix} P_{\bar{Y}=\mathrm{p}|Y=\mathrm{p},X} & P_{\bar{Y}=\mathrm{p}|Y=\mathrm{n},X} \\ P_{\bar{Y}=\mathrm{n}|Y=\mathrm{p},X} & P_{\bar{Y}=\mathrm{n}|Y=\mathrm{n},X} \end{pmatrix} \tag{60}$$

instantiates the contamination matrix M_{corr} in the CCN setting.

4.2.1 Generalized CCN

The concept of contaminating a *single* label can be extended to generating a *compound* label in the multiclass classification setting. Let 2^Y be the power set of the label space $Y = [K]$. Define $S := 2^Y \setminus \{\ , Y\}$ as the observable space of compound labels [3]. Since a compound label S S consists of an arbitrary number of class indices, one can view the probability of observing S for a given X is governed by several class probabilities $P_{Y=k|X}$. Therefore, generalizing the CCN formulation (59), we define the label-generating process of a compound label S as

$$P_{S|X} = \sum_{k=1}^{K} P_{S|Y=k,X} P_{Y=k|X},$$

where the role of $P_{S|Y,X}$ is the probability of converting a single label Y to a compound label S S. Moreover, in CCN, the distribution P_X is not contaminated. Thus, by multiplying P_X on both sides, we obtain the data-generating distribution

$$P_{S,X} = \sum_{k=1}^{K} P_{S|Y=k,X} P_{Y=k,X}, \tag{61}$$

Viewing $P_{S|Y,X}$ as a contamination probability, we arrange $P_{S=s|Y=k,X}$ into a matrix in the following lemma to formulate the contamination matrix for our generalized CCN (gCCN) setting.

Lemma 10. *Let the elements in S be $\{s_1, s_2, \ldots s_{|S|}\}$. For the gCCN setting, denote the data-generating distributions as*

$$\bar{P} := \begin{pmatrix} P_{S=s_1,X} \\ \vdots \\ P_{S=s_{|S|},X} \end{pmatrix} \tag{62}$$

and the base distributions as

$$B := \begin{pmatrix} P_{Y=1,X} \\ \vdots \\ P_{Y=K,X} \end{pmatrix} = P. \tag{63}$$

[3] Removing and Y is that the empty set does not contain any label information and Y is a trivial case.

Define

$$M_{\mathrm{gCCN}} := \begin{pmatrix} P_{S=s_1|Y=1,X} & P_{S=s_1|Y=2,X} & \cdots & P_{S=s_1|Y=K,X} \\ P_{S=s_2|Y=1,X} & P_{S=s_2|Y=2,X} & \cdots & P_{S=s_2|Y=K,X} \\ \vdots & \vdots & \ddots & \vdots \\ P_{S=s_{|\mathsf{S}|}|Y=1,X} & P_{S=s_{|\mathsf{S}|}|Y=2,X} & \cdots & P_{S=s_{|\mathsf{S}|}|Y=K,X} \end{pmatrix}. \tag{64}$$

Then, $\bar{P} = M_{\mathrm{gCCN}} B$.

The lemma implies that M_{gCCN} is the contamination matrix characterizing $\bar{P}$ of the gCCN setting. Also, note that $\bar{P} = M_{\mathrm{gCCN}} B$ is essentially the matrix form of (61). Moreover, M_{gCCN} generalizes M_{CCN} (60) by extending the label spaces: Both the clean label Y and the contaminated label $\bar{Y}$ belong to $\{\mathrm{p}, \mathrm{n}\}$ in CCN, while in the gCCN setting, the clean label $Y \quad \{1, \cdots, K\}$ and the compound label $S \quad \{s_1, \cdots, s_{|\mathsf{S}|}\}$.

Proof. For each $j \quad [|\mathsf{S}|]$, we have

$$\left(M_{\mathrm{gCCN}} B\right)_j = \sum_{k=1}^{K} P_{S=s_j|Y=k,X} P_{Y=k,X} = \sum_{k=1}^{K} P_{S=s_j,Y=k,X} = P_{S=s_j,X} = \bar{P}_j.$$

□

Comparing $\bar{P} = M_{\mathrm{gCCN}} B$ with the formulation framework $\bar{P} = M_{\mathrm{corr}} B$ (35), we have the reduction path

$$M_{\mathrm{corr}} \ \rightarrow \ M_{\mathrm{CCN}} \quad M_{\mathrm{gCCN}}.$$

Similar to M_{UU} (49), which induces multiple contamination matrices as special cases of the MCD model, M_{gCCN} also derives several contamination matrices formulating partial- or complementary-label settings, as we will show in the rest of this subsection.

4.2.2 Proper Partial-Label (PPL) Learning (Wu et al., 2023)

For a given example (y, x) and a compound label $s \quad \mathsf{S}$, we call s a partial-label of x if $y \quad s$. Statistically speaking, we assume $P_{Y \quad S|S,X} = 1$. Formally, according to Definition 1 of Wu et al. (2023), if the contamination probability can be defined as

$$P_{S|Y,X} := C(S, X) \mathsf{I}\left[Y \quad S\right], \tag{65}$$

via a function $C : \mathsf{S} \times \mathsf{X} \ \rightarrow \ \mathsf{R}$, we call such a partial-label scenario proper.

Since the discussion above only involves specifying $P_{S|Y,X}$, we replace the entries of M_{gCCN} (64) according to (65) to construct M_{PPL}:

$$\begin{pmatrix} C(s_1, X)\mathsf{I}\left[Y = 1 \quad s_1\right] & C(s_1, X)\mathsf{I}\left[Y = 2 \quad s_1\right] & \cdots & C(s_1, X)\mathsf{I}\left[Y = K \quad s_1\right] \\ C(s_2, X)\mathsf{I}\left[Y = 1 \quad s_2\right] & C(s_2, X)\mathsf{I}\left[Y = 2 \quad s_2\right] & \cdots & C(s_2, X)\mathsf{I}\left[Y = K \quad s_2\right] \\ \vdots & \vdots & \ddots & \vdots \\ C(s_{|\mathsf{S}|}, X)\mathsf{I}\left[Y = 1 \quad s_{|\mathsf{S}|}\right] & C(s_{|\mathsf{S}|}, X)\mathsf{I}\left[Y = 2 \quad s_{|\mathsf{S}|}\right] & \cdots & C(s_{|\mathsf{S}|}, X)\mathsf{I}\left[Y = K \quad s_{|\mathsf{S}|}\right] \end{pmatrix}. \tag{66}$$

The following lemma justifies M_{PPL} as the contamination matrix for PPL learning.

Lemma 11. *Let the elements in* S *be* $\left\{s_1, s_2, \ldots s_{|\mathsf{S}|}\right\}$. *For each* $j \quad [|\mathsf{S}|]$, *let the* j*-th entry of* $\bar{P}$ *be*

$$\bar{P}_j = P_{S=s_j,X} := C(S = s_j, X) \sum_{k \quad s_j} P_{Y=k,X},$$

which denotes the data-generating distribution of (s_j, X). *Assume the base distributions* B *and the contamination matrix* M_{PPL} *are given by (63) and (66), respectively. Then,* M_{PPL} *satisfies* $\bar{P} = M_{\mathrm{PPL}} B$ *and characterizes PPL learning (26).*

The entry replacement that converts (64) to (66) through (65) gives the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{gCCN}} \to M_{\mathrm{PPL}}.$$

4.2.3 Provably Consistent Partial-Label (PCPL) Learning (Feng et al., 2020b)

In PCPL, the probability of each partial-label is assumed to be sampled uniformly from all feasible partial-labels. Since there are $2^{K-1}-1$ feasible partial-labels for every y, the label-converting probability $P_{S=s|Y=y,X}$ is $\frac{1}{2^{K-1}-1}$ if $y \quad s$[4]. It corresponds to assign $C(S,X) = \frac{1}{2^{K-1}-1}$ in (65). Hence, we obtain

$$C(S,X)\mathsf{I}\left[Y \quad S\right] := \frac{1}{2^{K-1}-1}\mathsf{I}\left[Y \quad S\right], \tag{67}$$

which reduces the label-converting process of PPL to that of PCPL and recovers (5) of Feng et al. (2020b). Then, replacing entries in (66) via (67), we obtain the contamination matrix of PCPL learning

$$M_{\mathrm{PCPL}} := \frac{1}{2^{K-1}-1}\begin{pmatrix} \mathsf{I}\left[Y=1 \quad s_1\right] & \mathsf{I}\left[Y=2 \quad s_1\right] & \cdots & \mathsf{I}\left[Y=K \quad s_1\right] \\ \mathsf{I}\left[Y=1 \quad s_2\right] & \mathsf{I}\left[Y=2 \quad s_2\right] & \cdots & \mathsf{I}\left[Y=K \quad s_2\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{I}\left[Y=1 \quad s_{|\mathsf{S}|}\right] & \mathsf{I}\left[Y=2 \quad s_{|\mathsf{S}|}\right] & \cdots & \mathsf{I}\left[Y=K \quad s_{|\mathsf{S}|}\right] \end{pmatrix} \tag{68}$$

and the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{gCCN}} \to M_{\mathrm{PPL}} \to M_{\mathrm{PCPL}}.$$

M_{PCPL} characterizing the data-generating process of PCPL is justified by the following lemma, whose proof follows the same steps as that for Lemma 11.

Lemma 12. *Let the elements in* S *be* $\{s_1, s_2, \ldots s_{|\mathsf{S}|}\}$. *For each* $j \quad [|\mathsf{S}|]$, *let the* j*-th entry of* $\bar{P}$ *be*

$$\bar{P}_j = P_{S=s_j,X} := \frac{1}{2^{K-1}-1}\sum_{k \quad s_j} P_{Y=k,X},$$

which denotes the data-generating distribution of (s_j, X). *Assume the base distributions* B *and the contamination matrix* M_{PCPL} *are given by (63) and (68), respectively. Then,* M_{PCPL} *satisfies* $\bar{P} = M_{\mathrm{PCPL}} B$ *and characterizes PCPL learning (24).*

4.2.4 Multi-Complementary-Label (MCL) Learning (Feng et al., 2020a)

Recall the discussions in Sections 2.2.9 and 2.2.10 that a complementary-label contains the exclusion information of a true label. That is, for a data example (y, x), we call a set of class indices $\bar{s} \quad \mathsf{S} = 2^{\mathsf{Y}} \setminus \{\ , \mathsf{Y}\}$ an MCL of x if $\bar{s}$ does not contain y.

Denote $\mathsf{S} := \{\bar{s}_1, \bar{s}_2, \ldots \bar{s}_N\}$. The equivalence

$$\sum_{d=1}^{K-1} P_{|\bar{S}|=d} \cdot \frac{1}{\binom{K-1}{|\bar{S}|}} \sum_{Y \,/\, \bar{S}} P_{Y,X}\mathsf{I}\left[|\bar{S}| = d\right] = \begin{cases} \sum_{d=1}^{K-1} P_{|\bar{S}|=d} \cdot \frac{1}{\binom{K-1}{|\bar{S}|}} \sum_{Y \,/\, \bar{S}} P_{Y,X}, & \text{if } |\bar{S}| = d, \\ 0, & \text{otherwise} \end{cases}$$

[4]There are $2^{\mathsf{Y}\setminus\{y\}} \setminus \{\mathsf{Y}\setminus\{y\}\} = 2^{K-1} - 1$ combinations whose union with $\{y\}$ are partial-labels of y.

allows us to define the data-generating distribution of MCL (22) as

$$\bar{P} := \begin{pmatrix} P_{\bar{S}=\bar{s}_1,X} \\ \vdots \\ P_{\bar{S}=\bar{s}_N,X} \end{pmatrix}, \tag{69}$$

where for each j $[N]$,

$$P_{\bar{S}=\bar{s}_j,X} := \sum_{d=1}^{K-1} \frac{P_{|\bar{s}_j|=d}}{\binom{K-1}{|\bar{s}_j|}} \sum_{Y / \bar{s}_j} P_{Y,X} \mathbb{I}\left[|\bar{s}_j| = d\right]. \tag{70}$$

Let (y, x) be fixed. The data-generating process of MCL proposed by Feng et al. (2020a) is that one first samples a size d with probability $P_{|\bar{S}|=d}$, and then samples a $\bar{s}$ uniformly at random from $\{\bar{s}_{d,1}, \bar{s}_{d,2}, \ldots, \bar{s}_{d,N_d}\}$ S, where $\bar{s}_{d,\cdot}$ means a set of size d excluding y and N_d is the total number of those sets. Note that $N_d = \binom{K-1}{d}$ since we remove y from Y and then choose a set of size d to form a $\bar{s}_{d,\cdot}$. Furthermore, we need a more complicated lower index system to distinguish $\{\bar{s}_{d,1}, \bar{s}_{d,2}, \ldots, \bar{s}_{d,N_d}\}$ from $S = \{\bar{s}_1, \bar{s}_2, \ldots \bar{s}_N\}$ since d ranges from 1 to $K-1$ and $\sum_{d=1}^{K-1} N_d = \sum_{d=1}^{K-1} \binom{K-1}{d} = 2^K - 2 = |S|$. According to this mechanism, we construct M_{MCL}:

$$\begin{pmatrix} \frac{P_{|\bar{S}|=|\bar{s}_1|}}{\binom{K-1}{|\bar{s}_1|}} \mathbb{I}\left[Y = 1 \ / \ \bar{s}_1\right] & \frac{P_{|\bar{S}|=|\bar{s}_1|}}{\binom{K-1}{|\bar{s}_1|}} \mathbb{I}\left[Y = 2 \ / \ \bar{s}_1\right] & \cdots & \frac{P_{|\bar{S}|=|\bar{s}_1|}}{\binom{K-1}{|\bar{s}_1|}} \mathbb{I}\left[Y = K \ / \ \bar{s}_1\right] \\ \frac{P_{|\bar{S}|=|\bar{s}_2|}}{\binom{K-1}{|\bar{s}_2|}} \mathbb{I}\left[Y = 1 \ / \ \bar{s}_2\right] & \frac{P_{|\bar{S}|=|\bar{s}_2|}}{\binom{K-1}{|\bar{s}_2|}} \mathbb{I}\left[Y = 2 \ / \ \bar{s}_2\right] & \cdots & \frac{P_{|\bar{S}|=|\bar{s}_2|}}{\binom{K-1}{|\bar{s}_2|}} \mathbb{I}\left[Y = K \ / \ \bar{s}_2\right] \\ \vdots & \vdots & \ddots & \vdots \\ \frac{P_{|\bar{S}|=|\bar{s}_N|}}{\binom{K-1}{|\bar{s}_N|}} \mathbb{I}\left[Y = 1 \ / \ \bar{s}_N\right] & \frac{P_{|\bar{S}|=|\bar{s}_N|}}{\binom{K-1}{|\bar{s}_N|}} \mathbb{I}\left[Y = 2 \ / \ \bar{s}_N\right] & \cdots & \frac{P_{|\bar{S}|=|\bar{s}_N|}}{\binom{K-1}{|\bar{s}_N|}} \mathbb{I}\left[Y = K \ / \ \bar{s}_N\right] \end{pmatrix}. \tag{71}$$

The following lemma justifies M_{MCL} as the contamination matrix for MCL learning.

Lemma 13. *Suppose the base distributions B, the contamination matrix M_{MCL}, and the data-generating distributions $\bar{P}$ are given by (63), (71), and (70), respectively. Then, M_{MCL} satisfies $\bar{P} = M_{\mathrm{MCL}}B$ and characterizes MCL (22).*

At the first sight, M_{MCL} (71) does not resemble M_{PCPL} (68) or M_{PPL} (66). The subtle connection can be established via a relation between partial-label and complementary-label. Recall

$$2^Y \setminus \{\ , Y\} = S = \{\bar{s}_1, \bar{s}_2, \ldots \bar{s}_N\},$$

where $\bar{s}_j$ is a MCL. From the partial-label perspective, we can establish the following set equality relationship:

$$2^Y \setminus \{\ , Y\} = S = \{\bar{s}_1, \bar{s}_2, \ldots \bar{s}_N\} = \{s_1 := Y \setminus \bar{s}_1, s_2 := Y \setminus \bar{s}_2, \ldots, s_N := Y \setminus \bar{s}_N\}. \tag{72}$$

This is because for every $\bar{s}$ S, there is a s S such that $s := Y \setminus \bar{s}$. The intuition behind (72) is that if $\bar{s}$ is an MCL of x, then $s := Y \setminus \bar{s}$ must be a partial-label of x. Therefore, we can also use the set of partial-labels $\{s_1, s_2, \ldots, s_N\}$ to denote S. The following lemma exploits this relation to show that M_{MCL} is indeed a special case of M_{PPL}.

Lemma 14. *Assign each (s, k) entry of M_{PPL} (66) with*

$$C(s, X)\mathbb{I}\left[Y = k \quad s\right] := \frac{P_{|S|=|s|}}{\binom{K-1}{|s|-1}} \mathbb{I}\left[Y = k \quad s\right]. \tag{73}$$

Then, the resulting matrix

$$M_{\mathrm{MCL}} = \begin{pmatrix} \frac{P_{|S|=|s_1|}}{\binom{K-1}{|s_1|-1}} \mathbb{I}[Y=1 \quad s_1] & \frac{P_{|S|=|s_1|}}{\binom{K-1}{|s_1|-1}} \mathbb{I}[Y=2 \quad s_1] & \cdots & \frac{P_{|S|=|s_1|}}{\binom{K-1}{|s_1|-1}} \mathbb{I}[Y=K \quad s_1] \\ \frac{P_{|S|=|s_2|}}{\binom{K-1}{|s_2|-1}} \mathbb{I}[Y=1 \quad s_2] & \frac{P_{|S|=|s_2|}}{\binom{K-1}{|s_2|-1}} \mathbb{I}[Y=2 \quad s_2] & \cdots & \frac{P_{|S|=|s_2|}}{\binom{K-1}{|s_2|-1}} \mathbb{I}[Y=K \quad s_2] \\ \vdots & \vdots & \ddots & \vdots \\ \frac{P_{|S|=|s_N|}}{\binom{K-1}{|s_N|-1}} \mathbb{I}[Y=1 \quad s_N] & \frac{P_{|S|=|s_N|}}{\binom{K-1}{|s_N|-1}} \mathbb{I}[Y=2 \quad s_N] & \cdots & \frac{P_{|S|=|s_N|}}{\binom{K-1}{|s_N|-1}} \mathbb{I}[Y=K \quad s_N] \end{pmatrix} \tag{74}$$

is equivalent to M_{MCL} (71) under the relationship (72).

Proof. Note that for every $j \in [N]$, $s_j = Y \setminus \bar{s}_j$. This implies $P_{|\bar{S}|=|\bar{s}_j|} = P_{|S|=|s_j|}$, $\binom{K-1}{|s_j|-1} = \binom{K-1}{|\bar{s}_j|}$, and $\mathbb{I}[Y \in s_j] = \mathbb{I}[Y \notin \bar{s}_j]$ hold for every $j \in [N]$. Therefore, for each (j,k) entry in M_{MCL} and M_{MCL} (71), we have

$$\frac{P_{|S|=|s_j|}}{\binom{K-1}{|s_j|-1}} \mathbb{I}[Y=k \in s_j] = \frac{P_{|\bar{S}|=|\bar{s}_j|}}{\binom{K-1}{|\bar{s}_j|}} \mathbb{I}[Y=k \notin \bar{s}_j]. \tag{75}$$

□

The assignment rule (73) implies the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{gCCN}} \to M_{\mathrm{PPL}} \to M_{\mathrm{MCL}}.$$

Comparing (73) with (67) of PCPL, we see that MCL and PCPL can be viewed as different ways of composing $P_{Y,X}$ to generate a partial-label, with weights $\frac{P_{|S|=|s|}}{\binom{K-1}{|s|-1}}$ and $\frac{1}{2^{K-1}-1}$, respectively.

4.2.5 Complementary-Label (CL) Learning (Ishida et al., 2019)

As a special case of MCL (Section 2.2.10), we can construct the contamination matrix M_{CL} from M_{MCL}. The set of all CLs is composed of MCL with size 1: $\{1\}, \ldots, \{K\}$. Therefore, we assign values in M_{MCL} (71) as follows. For each $\bar{s} \in S$, $P_{|\bar{S}|=|\bar{s}|} = 1$ if $|\bar{s}| = 1$ and $P_{|\bar{S}|=|\bar{s}|} = 0$ if $|\bar{s}| > 1$. Dropping all-zero rows, we obtain from (71) the contamination matrix of CL learning

$$\begin{aligned} M_{\mathrm{CL}} &:= \begin{pmatrix} \frac{1}{K-1} \mathbb{I}[Y=1 \notin \{1\}] & \frac{1}{K-1} \mathbb{I}[Y=2 \notin \{1\}] & \cdots & \frac{1}{K-1} \mathbb{I}[Y=K \notin \{1\}] \\ \frac{1}{K-1} \mathbb{I}[Y=1 \notin \{2\}] & \frac{1}{K-1} \mathbb{I}[Y=2 \notin \{2\}] & \cdots & \frac{1}{K-1} \mathbb{I}[Y=K \notin \{2\}] \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{K-1} \mathbb{I}[Y=1 \notin \{K\}] & \frac{1}{K-1} \mathbb{I}[Y=2 \notin \{K\}] & \cdots & \frac{1}{K-1} \mathbb{I}[Y=K \notin \{K\}] \end{pmatrix} \\ &= \frac{1}{K-1} \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix} \end{aligned} \tag{76}$$

and the reduction path

$$M_{\mathrm{corr}} \to M_{\mathrm{gCCN}} \to M_{\mathrm{PPL}} \to M_{\mathrm{MCL}} \to M_{\mathrm{CL}}.$$

Furthermore, it is easy to verify that given B (63), for any $j \in [K]$,

$$\left(M_{\mathrm{CL}} B\right)_j = \sum_{Y=j} \frac{1}{K-1} P_{Y,X} = P_{\bar{S}=j,X},$$

which corresponds to formulation (20). Hence, we have the following.

Lemma 15. M_{CL} *(76) is the contamination matrix characterizing the data-generating distribution* $P_{\bar{S},X}$ *(20) of CL learning.*

4.3 Confidence-based Scenarios

At first sight, there seems to be no connection between "contamination" and single-class classification (Cao et al., 2021a). However, the following derivation

$$\frac{P_{Y=y_{\mathrm{s}}|X}}{P_{Y=j|X}} \cdot P_{Y=j,X} = \frac{P_{Y=y_{\mathrm{s}}|X}}{P_{Y=j|X}} \cdot P_{Y=j|X} P_X = P_{Y=y_{\mathrm{s}},X} \tag{77}$$

reveals a way to *contaminate* a clean joint probability $P_{Y=j,X}$ to the joint probability $P_{Y=y_{\mathrm{s}},X}$ of a designated class y_{s} via confidence weighting $\frac{P_{Y=y_{\mathrm{s}}|X}}{P_{Y=j|X}}$. As we will see in the rest of this subsection, the confidence weights are the key elements in formulating the contamination matrices for the confidence-based WSL settings.

4.3.1 Subset Confidence (Sub-Conf) Learning (Cao et al., 2021a)

Let $Y_{\mathrm{s}} \subseteq [K]$ be a subset of classes. Viewing Y_{s} as a "superclass", such that every instance x of (y, x) will be labeled Y_{s} if $y \in Y_{\mathrm{s}}$, we can define its class prior as $P_{Y \in Y_{\mathrm{s}}} = \pi_{Y_{\mathrm{s}}} := \sum_{y \in Y_{\mathrm{s}}} \pi_y$ and its class probability function as $P_{Y \in Y_{\mathrm{s}}|X} := \sum_{y \in Y_{\mathrm{s}}} P_{Y=y|X}$. Substituting the designated class y_{s} in (77) with the superclass Y_{s},

$$\frac{P_{Y \in Y_{\mathrm{s}}|X}}{P_{Y=j|X}} \cdot P_{Y=j,X} = \frac{P_{Y \in Y_{\mathrm{s}}|X}}{P_{Y=j|X}} \cdot P_{Y=j|X} P_X = P_{Y \in Y_{\mathrm{s}},X} \tag{78}$$

shows that no matter what joint distribution $P_{Y=j,X}$ to begin with, the confidence weight $\frac{P_{Y \in Y_{\mathrm{s}}|X}}{P_{Y=j|X}}$ twists that joint distribution so that every observed data appears to be sampled from the same superclass distribution $P_{Y_{\mathrm{s}},X}$. The following lemma leverages the observation to specify the contamination matrix M_{Sub} characterizing Sub-Conf learning.

Lemma 16. *Denote the base distributions as*

$$B := \begin{pmatrix} P_{Y=1,X} \\ \vdots \\ P_{Y=K,X} \end{pmatrix} = P \tag{79}$$

and the data-generating distributions as

$$\bar{P} := \begin{pmatrix} P_{Y \in Y_{\mathrm{s}},X} \\ \vdots \\ P_{Y \in Y_{\mathrm{s}},X} \end{pmatrix}.$$

Inserting the confidence weights into the identity matrix, we define the contamination matrix

$$M_{\mathrm{Sub}} := \begin{pmatrix} \frac{P_{Y \in Y_{\mathrm{s}}|X}}{P_{Y=1|X}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{P_{Y \in Y_{\mathrm{s}}|X}}{P_{Y=K|X}} \end{pmatrix}. \tag{80}$$

Then, $\bar{P} = M_{\mathrm{Sub}} B$*, and* M_{Sub} *characterizes the data-generating process of Sub-Conf learning (30).*

Proof. For each $j \in [K]$, $\left(M_{\mathrm{Sub}} B\right)_j = P_{Y \in Y_{\mathrm{s}},X}$ follows from (78). Thus, $\bar{P} = M_{\mathrm{Sub}} B$. It further implies all observed instances are labeled with the same superclass Y_{s}, meaning we can drop the observed labels, and the observed examples $\{x_i\}_{i=1}^n$ is equivalent to a set of i.i.d. samples from $P_{X|Y \in Y_{\mathrm{s}}}$ (30). □

Comparing $\bar{P} = M_{\text{Sub}}B$ with the formulation framework $\bar{P} = M_{\text{corr}}B$ (35), we observe that in Sub-Conf learning, M_{corr} is realized as M_{Sub}:

$$M_{\text{corr}} \to M_{\text{Sub}}.$$

4.3.2 Single-Class Confidence (SC-Conf) Learning (Cao et al., 2021a)

We compare the formulation of SC-Conf (28) with Sub-Conf (30) and observe that SC-Conf is a special case of Sub-Conf when $\mathsf{Y}_{\text{s}} = \{y_{\text{s}}\}$ being a singleton. Thus, we straightforwardly obtain the matrix formulation of SC-Conf from Lemma 16 be replacing Y_{s} in (80) with y_{s}:

Lemma 17. *Let the base distributions B be defined by (79) and the data-generating distributions be defined by*

$$\bar{P} := \begin{pmatrix} \mathsf{P}_{Y=y_{\text{s}},X} \\ \vdots \\ \mathsf{P}_{Y=y_{\text{s}},X} \end{pmatrix}.$$

Define the contamination matrix

$$M_{\text{SC}} := \begin{pmatrix} \frac{\mathsf{P}_{Y=y_{\text{s}}|X}}{\mathsf{P}_{Y=1|X}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\mathsf{P}_{Y=y_{\text{s}}|X}}{\mathsf{P}_{Y=K|X}} \end{pmatrix} \tag{81}$$

by substituting Y_{s} in (80) with y_{s}. Then, $\bar{P} = M_{\text{SC}}B$ and M_{SC} characterizes the data-generating process of SC-Conf learning (28).

Since SC-Conf is a special case of Sub-Conf, we have the reduction path

$$M_{\text{corr}} \to M_{\text{Sub}} \to M_{\text{SC}}.$$

4.3.3 Positive-confidence (Pconf) Learning (Ishida et al., 2018)

Comparing (6) with (28), we see that Pconf is a special case of SC-Conf when $K = 2$ and $y_{\text{s}} = \text{p}$ since $r_{\text{n}}(X) = 1 - r_{\text{p}}(X)$. A further modification to (81) we obtain the contamination matrix M_{Pconf} characterizing Pconf learning.

Lemma 18. *Let $B := \begin{pmatrix} \mathsf{P}_{Y=\text{p},X} \\ \mathsf{P}_{Y=\text{n},X} \end{pmatrix} = P$ and $\bar{P} := \begin{pmatrix} \mathsf{P}_{Y=\text{p},X} \\ \mathsf{P}_{Y=\text{p},X} \end{pmatrix}$. Define*

$$M_{\text{Pconf}} := \begin{pmatrix} \frac{\mathsf{P}_{Y=\text{p}|X}}{\mathsf{P}_{Y=\text{p}|X}} & 0 \\ 0 & \frac{\mathsf{P}_{Y=\text{p}|X}}{\mathsf{P}_{Y=\text{n}|X}} \end{pmatrix}. \tag{82}$$

Then, $\bar{P} = M_{\text{Pconf}}B$, and M_{Pconf} characterizes the data-generating process of Pconf learning (6).

The entry replacement that converts (81) to (82) implies the reduction path

$$M_{\text{corr}} \to M_{\text{Sub}} \to M_{\text{SC}} \to M_{\text{Pconf}}.$$

4.3.4 Soft-Label Learning (Ishida et al., 2023)

The difference between the soft-label and the previous confidence-based settings (Sub-Conf, SC-Conf, and Pconf) is how x is sampled. The sample distributions condition on the label information in the previous settings, while that in soft-label is P_X. Replacing the confidence weight $\frac{\mathsf{P}_{Y=y_s|X}}{\mathsf{P}_{Y=j|X}}$ in (77) with $\frac{1}{\mathsf{P}_{Y=j|X}}$,

$$\frac{1}{\mathsf{P}_{Y=j|X}} \cdot \mathsf{P}_{Y=j,X} = \mathsf{P}_X$$

explains how to convert $\mathsf{P}_{Y=j,X}$ to P_X. Therefore, filling the j-th diagonal entry of the identity matrix with $\frac{1}{\mathsf{P}_{Y=j|X}}$, we obtain the contamination matrix M_{Soft} for soft-label learning:

Lemma 19. *Let the base distributions B be defined by (79). Denote the data-generating distribution as*

$$\bar{P} := \begin{pmatrix} \mathsf{P}_X \\ \vdots \\ \mathsf{P}_X \end{pmatrix}. \tag{83}$$

Define the contamination matrix

$$M_{\text{Soft}} := \begin{pmatrix} \frac{1}{\mathsf{P}_{Y=1|X}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\mathsf{P}_{Y=K|X}} \end{pmatrix}. \tag{84}$$

Then, $\bar{P} = M_{\text{Soft}} B$, and M_{Soft} characterizes the data-generating process in (33).

Unlike SC-Conf and Pconf, which are special cases of Sub-Conf with Y_s taking only one label, the generation process of a soft-label can be viewed as assigning $\mathsf{Y}_s := [K]$. Considering the entire label space results in $\mathsf{P}_{Y\ [K]|X} = 1$; it coincides with the meaning of P_X that samples x regardless of the labels. Although technically the soft-label setting is not a special case of Sub-Conf (recalling the $\mathsf{Y}_s \quad [K]$ assumption from Section 4.3.1), M_{Soft} (84) is reduced from M_{Sub} (80) by realizing $\mathsf{P}_{Y\ \mathsf{Y}_s|X}$ as $\mathsf{P}_{Y\ [K]|X} = 1$. Therefore, we obtain the following reduction path

$$M_{\text{corr}} \to M_{\text{Sub}} \to M_{\text{Soft}}.$$

5 Risk Rewrite via Decontamination

We have demonstrated the capability of the proposed formulation component (35) in the last section. This section shows how the proposed framework provides a unified methodology for solving the risk rewrite problem. Specifically, given each contamination matrix described in Section 4, we show how to construct the corrected losses (38) to perform the risk rewrite via (39). We then recover each rewrite to the corresponding form reported in the literature to justify its feasibility. Because this paper focuses on a unified methodology for rewriting the classification risk instead of the designs of practical training objectives, we assume the required parameters are given or can be estimated accurately from the observed data.

Similar to the previous section, we only provide proofs that appear for the first time to avoid repeating similar proofs. For auxiliary lemmas and results whose proofs are similar to the previous ones, we refer to the omitted proofs in Appendix D. In particular, the omitted proofs in Section 5.1 can be found in Appendix D.1, those in Section 5.2 can be found in Appendix D.2, and those in Section 5.3 can be found in Appendix D.3.

5.1 MCD Scenarios

We apply the framework to conduct the risk rewrites for WSLs formulated in Section 4.1 and summarized in Table 7. A general approach is to show that the inversion method discussed in Proposition 1 provides the decontamination matrix $M_{\text{corr}}^{\dagger}$ required in (38).

5.1.1 Unlabeled-Unlabeled (UU) Learning

We justify the proposed framework for UU learning via the following steps.

Step 1: Corrected Loss Design and Risk Rewrite.
The three milestones in the proposed framework are (1) finding the contamination matrix M_{corr} that characterizes the data-generating process (35) of a weakly supervised scenario, (2) finding the decontamination matrix $M_{\mathrm{corr}}^{\dagger}$ that compensates for the contamination effect (37), which is then used in (3) the construction of corrected losses (38) for the risk rewrite (39).

Section 4.1.1 has reached the first milestone (35) as (48) of the form $\bar{P} = M_{\mathrm{UU}} B$ finds

$$M_{\mathrm{UU}} = \begin{pmatrix} 1 - \gamma_1 & \gamma_1 \\ \gamma_2 & 1 - \gamma_2 \end{pmatrix}$$

that connects the data-generating distributions $\bar{P} = \begin{pmatrix} P_{\mathrm{U}_1} \\ P_{\mathrm{U}_2} \end{pmatrix}$ and the base distributions $B = \begin{pmatrix} P_{X|Y=\mathrm{p}} \\ P_{X|Y=\mathrm{n}} \end{pmatrix}$. Note that B is not the risk-defining distributions $P = \begin{pmatrix} P_{Y=\mathrm{p},X} \\ P_{Y=\mathrm{n},X} \end{pmatrix}$, we need an additional step before reaching the second milestone. To further link $\bar{P}$ with P, we still need a M_{trsf} that satisfies $B = M_{\mathrm{trsf}} P$. Introducing the prior matrix $\Pi = \begin{pmatrix} \pi_{\mathrm{p}} & 0 \\ 0 & \pi_{\mathrm{n}} \end{pmatrix}$, we see that choosing $M_{\mathrm{trsf}} := \Pi^{-1}$ fulfills the need:

$$M_{\mathrm{trsf}} P = \begin{pmatrix} \pi_{\mathrm{p}}^{-1} & 0 \\ 0 & \pi_{\mathrm{n}}^{-1} \end{pmatrix} \begin{pmatrix} P_{Y=\mathrm{p},X} \\ P_{Y=\mathrm{n},X} \end{pmatrix} = \begin{pmatrix} \frac{P_{Y=\mathrm{p},X}}{P_{Y=\mathrm{p}}} \\ \frac{P_{Y=\mathrm{n},X}}{P_{Y=\mathrm{n}}} \end{pmatrix} = \begin{pmatrix} P_{X|Y=\mathrm{p}} \\ P_{X|Y=\mathrm{n}} \end{pmatrix} = B.$$

Hence, we can instantiate $\bar{P} = M_{\mathrm{corr}} M_{\mathrm{trsf}} P$ (36) as

$$\begin{pmatrix} P_{\mathrm{U}_1} \\ P_{\mathrm{U}_2} \end{pmatrix} = M_{\mathrm{UU}} \Pi^{-1} \begin{pmatrix} P_{Y=\mathrm{p},X} \\ P_{Y=\mathrm{n},X} \end{pmatrix} \tag{85}$$

in UU learning.

Next, we use Proposition 1 to derive the decontamination matrix $M_{\mathrm{corr}}^{\dagger}$ to reach the second milestone (37).

Corollary 20. *Assume M_{UU} in (85) is invertible. Then, defining the decontamination matrix for UU learning as*

$$M_{\mathrm{UU}}^{\dagger} := \Pi M_{\mathrm{UU}}^{-1}$$

gives rise to $M_{\mathrm{UU}}^{\dagger} \bar{P} = P$.

Proof. Suggested by Proposition 1, the inverse matrix ΠM_{UU}^{-1} cancels out the contamination brought by $M_{\mathrm{UU}} \Pi^{-1}$ in (85). Assigning $M_{\mathrm{UU}}^{\dagger} = \Pi M_{\mathrm{UU}}^{-1}$ and repeating the proof of Proposition 1, we have

$$M_{\mathrm{UU}}^{\dagger} \bar{P} = \Pi M_{\mathrm{UU}}^{-1} \bar{P} = \Pi M_{\mathrm{UU}}^{-1} M_{\mathrm{UU}} \Pi^{-1} P = P$$

that completes the proof. □

Now we will move on to the third milestone. With $M_{\mathrm{UU}}^{\dagger}$ in hand, we devise the corrected losses $\bar{L}$ to achieve the risk rewrite for UU learning. We denote the corrected loss at the $\bar{k}$-th entry of $\bar{L}$ as $\bar{\ell}_{\bar{k}} := \ell_{\bar{Y}=\bar{k}}(g(X))$, where $\bar{k} \quad \bar{Y}$ is a class of the observed data[5]. In UU learning, $\bar{Y} = \{\mathrm{U}_1, \mathrm{U}_2\}$. The following theorem proves rewrite (9) in Section 2.2.3.

[5]The definition of the corrected loss $\bar{\ell}_{\bar{k}}$ is in contrast to the original loss $\ell_k := \ell_{Y=k}(g(X))$.

Theorem 21. *Let $\gamma_1, \gamma_2 > 0$ and $\gamma_1 + \gamma_2 = 1$. Then, $M_{\mathrm{UU}}^{\dagger}$ defined in Corollary 20 is feasible. Moreover, the vector of corrected losses suggested by (38)*

$$\begin{pmatrix} \bar{\ell}_{\mathrm{U}_1} & \bar{\ell}_{\mathrm{U}_2} \end{pmatrix} = \bar{L} \quad := L \quad M_{\mathrm{UU}}^{\dagger}$$

with

$$\begin{aligned} \bar{\ell}_{\mathrm{U}_1} &= \frac{(1-\gamma_2)\pi_{\mathrm{p}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{p}} + \frac{-\gamma_2\pi_{\mathrm{n}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{n}}, \\ \bar{\ell}_{\mathrm{U}_2} &= \frac{-\gamma_1\pi_{\mathrm{p}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{p}} + \frac{(1-\gamma_1)\pi_{\mathrm{n}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{n}} \end{aligned} \tag{86}$$

achieves the following risk rewrite:

$$R(g) = \mathsf{E}_{\mathrm{U}_1}\left[\bar{\ell}_{\mathrm{U}_1}\right] + \mathsf{E}_{\mathrm{U}_2}\left[\bar{\ell}_{\mathrm{U}_2}\right]. \tag{87}$$

Proof. Since $\gamma_1 + \gamma_2 = 1$,

$$M_{\mathrm{UU}}^{-1} = \begin{pmatrix} 1-\gamma_1 & \gamma_1 \\ \gamma_2 & 1-\gamma_2 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1-\gamma_2}{1-\gamma_1-\gamma_2} & \frac{-\gamma_1}{1-\gamma_1-\gamma_2} \\ \frac{-\gamma_2}{1-\gamma_1-\gamma_2} & \frac{1-\gamma_1}{1-\gamma_1-\gamma_2} \end{pmatrix}$$

exists. Thus, it is feasible for us to define $M_{\mathrm{UU}}^{\dagger} := \Pi M_{\mathrm{UU}}^{-1}$ according to Corollary 20. Following (38), we construct

$$\bar{L} \quad := L \quad M_{\mathrm{UU}}^{\dagger}$$

and obtain

$$\begin{aligned} \begin{pmatrix} \bar{\ell}_{\mathrm{U}_1} & \bar{\ell}_{\mathrm{U}_2} \end{pmatrix} &= L \quad \Pi M_{\mathrm{UU}}^{-1} \\ &= \begin{pmatrix} \ell_{\mathrm{p}} & \ell_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} \pi_{\mathrm{p}} & 0 \\ 0 & \pi_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} \frac{1-\gamma_2}{1-\gamma_1-\gamma_2} & \frac{-\gamma_1}{1-\gamma_1-\gamma_2} \\ \frac{-\gamma_2}{1-\gamma_1-\gamma_2} & \frac{1-\gamma_1}{1-\gamma_1-\gamma_2} \end{pmatrix} \\ &= \begin{pmatrix} \ell_{\mathrm{p}} & \ell_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} \frac{(1-\gamma_2)\pi_{\mathrm{p}}}{1-\gamma_1-\gamma_2} & \frac{-\gamma_1\pi_{\mathrm{p}}}{1-\gamma_1-\gamma_2} \\ \frac{-\gamma_2\pi_{\mathrm{n}}}{1-\gamma_1-\gamma_2} & \frac{(1-\gamma_1)\pi_{\mathrm{n}}}{1-\gamma_1-\gamma_2} \end{pmatrix} \end{aligned} \tag{88}$$

that gives (86).

Next, with the critical component $\bar{L}$ in hand, applying (39), we obtain

$$\begin{aligned} R(g) &= \int_X \bar{L} \quad \bar{P}\,\mathrm{d}x \\ &= \int_X \left(P_{\mathrm{U}_1}\,\bar{\ell}_{\mathrm{U}_1} + P_{\mathrm{U}_2}\,\bar{\ell}_{\mathrm{U}_2}\right)\mathrm{d}x \\ &= \mathsf{E}_{\mathrm{U}_1}\left[\bar{\ell}_{\mathrm{U}_1}\right] + \mathsf{E}_{\mathrm{U}_2}\left[\bar{\ell}_{\mathrm{U}_2}\right], \end{aligned} \tag{89}$$

where the first equality holds since according to Corollary 20,

$$\bar{L} \quad \bar{P} = L \quad M_{\mathrm{UU}}^{\dagger}\bar{P} = L \quad P.$$

□

In (88), we do not need to specify the instance in ℓ_{p} and ℓ_{n} to be x^{u_1} or x^{u_2} since the equality holds for any instance x. We only need to distinguish x^{u_1} from x^{u_2} when the corrected losses multiply the data distributions. In particular, the detailed form of rewrite (89) using (8) is

$$\begin{aligned} R(g) &= \mathsf{E}_{\mathrm{U}_1}\left[\bar{\ell}_{\mathrm{U}_1}\right] + \mathsf{E}_{\mathrm{U}_2}\left[\bar{\ell}_{\mathrm{U}_2}\right] \\ &= \mathsf{E}_{x^{\mathrm{u}_1}\ P_{\mathrm{U}_1}}\left[\frac{(1-\gamma_2)\pi_{\mathrm{p}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{p}}(X^{\mathrm{u}_1}) + \frac{-\gamma_2\pi_{\mathrm{n}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{n}}(X^{\mathrm{u}_1})\right] \\ &\quad + \mathsf{E}_{x^{\mathrm{u}_2}\ P_{\mathrm{U}_2}}\left[\frac{-\gamma_1\pi_{\mathrm{p}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{p}}(X^{\mathrm{u}_2}) + \frac{(1-\gamma_1)\pi_{\mathrm{n}}}{1-\gamma_1-\gamma_2}\ell_{\mathrm{n}}(X^{\mathrm{u}_2})\right]. \end{aligned}$$

The freedom from specifying x in (88) eliminates the notational burden of distinguishing $\ell_Y(X^{\mathrm{u}_1})$ from $\ell_Y(X^{\mathrm{u}_2})$, allowing us to exploit the advantage of matrix multiplication while constructing the corrected losses. The freedom also enables separated treatments for the data distributions (e.g., formulating $\bar{P} = M_{\mathrm{UU}}\Pi^{-1}P$) and the corrected losses (e.g., devising $\bar{L} = L\ M_{\mathrm{UU}}^{\dagger}$).

Step 2: Recovering the previous result(s).
Lastly, we verify the feasibility of our rewrite by showing that our rewrite corresponds to an existing result. By parameter substitution, we replace γ_1 with $1-\theta$, γ_2 with θ, π_{n} with $1-\pi_{\mathrm{p}}$, ℓ_{p} with $\ell(g(X))$, and ℓ_{n} with $\ell(-g(X))$. Then, (86) becomes

$$\begin{aligned} \frac{(1-\theta)\pi_{\mathrm{p}}}{\theta-\theta}\ell(g(X)) + \frac{-\theta\,(1-\pi_{\mathrm{p}})}{\theta-\theta}\ell(-g(X)) &= \bar{\ell}_{+}(g(X)), \\ \frac{\theta(1-\pi_{\mathrm{p}})}{\theta-\theta}\ell(-g(X)) + \frac{-(1-\theta)\pi_{\mathrm{p}}}{\theta-\theta}\ell(g(X)) &= \bar{\ell}_{-}(-g(X)), \end{aligned}$$

recovering the corrected loss functions (8) and the constants reported in Theorem 4 of Lu et al. (2019).

5.1.2 Positive-Unlabeled (PU) Learning

Recall that all WSLs discussed in Section 4.1 share the same base distributions B (46). Further, as shown in Table 7, the contamination matrix of every WSL scenario beneath UU learning except M_{Sconf} is a child of M_{UU} on the reduction graph. It means $\bar{P} = M_{\mathrm{UU}}\Pi^{-1}P$ (85) is a general form for every child scenario in Table 7 (with different realizations of γ_1 and γ_2). Hence, we can reuse Theorem 21 to conduct the risk rewrite for every child scenario on the reduction graph. PU learning is the first of such examples.

Step 1: Corrected Loss Design and Risk Rewrite.
By the following corollary, we prove the rewrite (5) in Section 2.2.1.

Corollary 22. *For PU learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{\mathrm{P}}\left[\bar{\ell}_{\mathrm{P}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right], \tag{90}$$

where

$$\begin{aligned} \bar{\ell}_{\mathrm{P}} &= \pi_{\mathrm{p}}\ell_{\mathrm{p}} - \pi_{\mathrm{p}}\ell_{\mathrm{n}}, \\ \bar{\ell}_{\mathrm{U}} &= \ell_{\mathrm{n}}. \end{aligned} \tag{91}$$

Step 2: Recovering the previous result(s).
Since P_{P} is $P_{X|Y=\mathrm{p}}$ and P_{U} is P_X, we swap the notations to obtain

$$\begin{aligned} R(g) &= \mathsf{E}_{\mathrm{P}}\left[\bar{\ell}_{\mathrm{P}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right] \\ &= \mathsf{E}_{\mathrm{P}}\left[\pi_{\mathrm{p}}\ell_{\mathrm{p}} - \pi_{\mathrm{p}}\ell_{\mathrm{n}}\right] + \mathsf{E}_{\mathrm{U}}\left[\ell_{\mathrm{n}}\right] \\ &= \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[\ell_{\mathrm{p}}\right] - \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[\ell_{\mathrm{n}}\right] + \mathsf{E}_{X}\left[\ell_{\mathrm{n}}\right] \end{aligned}$$

from (90), which corresponds to the risk estimators (2) in Kiryo et al. (2017) and (3) in du Plessis et al. (2015).

Moreover, with an additional symmetric assumption of $\ell_{\mathrm{p}} + \ell_{\mathrm{n}} = 1$, one further obtains

$$\begin{aligned} R(g) &= \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[\ell_{\mathrm{p}}\right] - \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[\ell_{\mathrm{n}}\right] + \mathsf{E}_X\left[\ell_{\mathrm{n}}\right] \\ &= \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[\ell_{\mathrm{p}}\right] - \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[1 - \ell_{\mathrm{p}}\right] + \mathsf{E}_X\left[\ell_{\mathrm{n}}\right] \\ &= \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[\ell_{\mathrm{p}}\right] - \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[1\right] + \pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[\ell_{\mathrm{p}}\right] + \mathsf{E}_X\left[\ell_{\mathrm{n}}\right] \\ &= 2\pi_{\mathrm{p}}\mathsf{E}_{X|Y=\mathrm{p}}\left[\ell_{\mathrm{p}}\right] - \pi_{\mathrm{p}} + \mathsf{E}_X\left[\ell_{\mathrm{n}}\right]. \end{aligned}$$

This expression recovers several risk rewrites such as (4) of Kiryo et al. (2017), (3) of Niu et al. (2016), (2) of du Plessis et al. (2015)[6], and (3) of du Plessis et al. (2014).

5.1.3 Similar-Unlabeled (SU) Learning

According to Table 7, M_{SU} is a child of M_{UU} on the reduction graph. Thus, we can follow the same steps illustrated in Section 5.1.2 to justify the proposed framework.

Step 1: Corrected Loss Design and Risk Rewrite.
The following corollary combines (86) and (87) to conduct the risk rewrite.

Corollary 23. *Assume $\pi_{\mathrm{p}} = 1/2$. For SU learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{\tilde{\mathrm{S}}}\left[\bar{\ell}_{\tilde{\mathrm{S}}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right],$$

where

$$\begin{aligned} \bar{\ell}_{\tilde{\mathrm{S}}} &= \frac{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{p}} - \frac{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{n}}, \\ \bar{\ell}_{\mathrm{U}} &= -\frac{\pi_{\mathrm{n}}}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{p}} + \frac{\pi_{\mathrm{p}}}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{n}}. \end{aligned} \tag{92}$$

Step 2: Recovering the previous result(s).
To recover Theorem 1 of Bao et al. (2018), we first need to restore $\mathsf{E}_{\mathrm{S}}\left[\cdot\right]$ from $\mathsf{E}_{\tilde{\mathrm{S}}}\left[\cdot\right]$ in Corollary 23. The following lemma provides a means for us to do so.

Lemma 24. *Given B (46) and following the SU learning notations, we have*

$$M_{\mathrm{SU}}B = \begin{pmatrix} P_{\tilde{\mathrm{S}}} \\ P_{\mathrm{U}} \end{pmatrix} = \bar{P},$$

where

$$M_{\mathrm{SU}} := \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2 \int_{x \in \mathcal{X}} P_{x|Y=\mathrm{p}}\mathrm{d}x}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2 \int_{x \in \mathcal{X}} P_{x|Y=\mathrm{n}}\mathrm{d}x}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2} \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix}.$$

Proof. Since $\int_{x \in \mathcal{X}} P_{x|Y=\mathrm{p}}\mathrm{d}x = 1$ and $\int_{x \in \mathcal{X}} P_{x|Y=\mathrm{n}}\mathrm{d}x = 1$, we have $M_{\mathrm{SU}} = M_{\mathrm{SU}}$, and hence $M_{\mathrm{SU}}B = M_{\mathrm{SU}}B = \begin{pmatrix} P_{\tilde{\mathrm{S}}} \\ P_{\mathrm{U}} \end{pmatrix}$. The last equality follows from Lemma 5. □

[6] As the 0-1 loss is symmetric.

Lemma 24 allows us to slightly revise the derivation (89) as follows:

$$\begin{aligned} R(g) &= \int_x \int_{x'} \bar{L} \;\; \bar{P}\,\mathrm{d}x = \int_x \int_{x'} \bar{L} \;\; M_{\mathrm{SU}} B\,\mathrm{d}x \\ &= \int_x \int_{x'} \begin{pmatrix} \bar{\ell}_{\tilde{\mathrm{S}}} & \bar{\ell}_{\mathrm{U}} \end{pmatrix} \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2 \int_x \int_{x'} P_{x\,,x'\,|Y=\mathrm{p}}\mathrm{d}x}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2 \int_x \int_{x'} P_{x\,,x'\,|Y=\mathrm{n}}\mathrm{d}x}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} P_{x|Y=\mathrm{p}} \\ P_{x|Y=\mathrm{n}} \end{pmatrix} \mathrm{d}x \\ &\overset{(a)}{=} \int_x \int_{x'} \int_x \int_{x'} P_{\mathrm{S}} \bar{\ell}_{\tilde{\mathrm{S}}}\,\mathrm{d}x\,\mathrm{d}x' + \int_x \int_{x'} P_{\mathrm{U}}\,\bar{\ell}_{\mathrm{U}}\,\mathrm{d}x \\ &= \mathsf{E}_{\mathrm{S}}\left[\bar{\ell}_{\tilde{\mathrm{S}}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right], \end{aligned}$$

where equality (a) follows from the SU formulation (10).

Then, denoting

$$L(X) := \frac{1}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) - \frac{1}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X), \tag{93}$$

$$L_-(X) := -\frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) + \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X) \tag{94}$$

and continuing with (92), we obtain

$$\begin{aligned} \mathsf{E}_{\mathrm{S}}\left[\bar{\ell}_{\tilde{\mathrm{S}}}\right] &= \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\mathsf{E}_{\mathrm{S}}\left[\frac{1}{2\pi_{\mathrm{p}} - 1}\left(\ell_{\mathrm{p}} - \ell_{\mathrm{n}}\right)\right] \\ &= \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\mathsf{E}_{\mathrm{S}}\left[L(X)\right] \\ &\overset{(b)}{=} \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\mathsf{E}_{\mathrm{S}}\left[\frac{L(X) + L(X')}{2}\right] \end{aligned}$$

and

$$\begin{aligned} \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right] &= \mathsf{E}_{\mathrm{U}}\left[-\frac{\pi_{\mathrm{n}}}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{p}} + \frac{\pi_{\mathrm{p}}}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{n}}\right] \\ &= \mathsf{E}_{\mathrm{U}}\left[L_-(f(X))\right] \end{aligned} \tag{95}$$

that prove rewrite (11) in Section 2.2.4 and recover Theorem 1 of Bao et al. (2018) by matching notations[7]. The following lemma justifies equality (b).

Lemma 25. *Let* (x, x') P_{S} *defined by (10). Then,* $\mathsf{E}_{\mathrm{S}}\left[\frac{L(X)}{2}\right] = \mathsf{E}_{\mathrm{S}}\left[\frac{L(X')}{2}\right]$.

The derivation demonstrates the flexibility of the proposed framework in which a slight modification of M_{SU} recovers the pairwise distribution P_{S} required for $\mathsf{E}_{\mathrm{S}}[\cdot]$. Moreover, the technique developed here significantly reduces the proof length in Appendix B of Bao et al. (2018). Later in Section 5.1.5, we apply the same trick to recover Theorem 1 of Shimada et al. (2021) for SDU learning.

We remark that the result recovered in this paper is merely Theorem 1 of Bao et al. (2018) but not the last expression in (5) of Bao et al. (2018), which later was implemented as the objective (10) for optimization. It is because, pointed out by Negishi (2023), the additional assumption $P_{\mathrm{S}}(x, x') = P_{\tilde{\mathrm{S}}}(x)P_{\tilde{\mathrm{S}}}(x')$ required for achieving (5) of Bao et al. (2018) is impractical. We note that the remedy proposed by Negishi (2023) can be analyzed by the proposed framework, but we omit it due to the amount of overlap with the analyses in Sections 5.1.1 and 5.1.3.

[7]The matching to the notations of Bao et al. (2018) is as follows: π_{p} is π_+, π_{n} is π_-, $\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2$ is π_{S}, P_{S} is p_{S}, P_{U} is p, ℓ_{p} is $\ell(f(X), +1)$, ℓ_{n} is $\ell(f(X), -1)$, $L(X)$ by definition is $\frac{1}{2\pi_+ - 1}\left(\ell(f(X), +1) - \ell(f(X), -1)\right)$, and $L_-(f(X))$ by definition is $-\frac{\pi_-}{2\pi_+ - 1}\ell(f(X), +1) + \frac{\pi_+}{2\pi_+ - 1}\ell(f(X), -1)$.

5.1.4 Pairwise Comparison (Pcomp) Learning

We follow the steps illustrated in Section 5.1.2 to justify the proposed framework since, by Table 7, M_{Pcomp} is reduced from M_{UU}.

Step 1: Corrected Loss Design and Risk Rewrite.
The following corollary combines (86) and (87) to achieve rewrite (17) in Section 2.2.7.

Corollary 26. *For Pcomp learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{\mathrm{Sup}}\left[\bar{\ell}_{\mathrm{Sup}}\right] + \mathsf{E}_{\mathrm{Inf}}\left[\bar{\ell}_{\mathrm{Inf}}\right],$$

where

$$\begin{aligned} \bar{\ell}_{\mathrm{Sup}} &= \ell_{\mathrm{p}} - \pi_{\mathrm{p}}\ell_{\mathrm{n}}, \\ \bar{\ell}_{\mathrm{Inf}} &= -\pi_{\mathrm{n}}\ell_{\mathrm{p}} + \ell_{\mathrm{n}}. \end{aligned} \tag{96}$$

Step 2: Recovering the previous result(s).
It is straightforward to recover Theorem 3 of Feng et al. (2021) by matching notations[8]. Since x is a variable and can be substituted by x , we express Corollary 26 as

$$R(g) = \mathsf{E}_{x \sim P_{\mathrm{Sup}}}\left[\ell_{\mathrm{p}}(x) - \pi_{\mathrm{p}}\ell_{\mathrm{n}}(x)\right] + \mathsf{E}_{x \sim P_{\mathrm{Inf}}}\left[\ell_{\mathrm{n}}(x\,) - \pi_{\mathrm{n}}\ell_{\mathrm{p}}(x\,)\right], \tag{97}$$

recovering (5) of Feng et al. (2021).

5.1.5 Similar-dissimilar-unlabeled (SDU) Learning

We justify the applicability of the proposed framework for DU and SD separately. Firstly, we start with DU learning, which is similar to SU learning in the sense that pairwise information is provided. From Lemmas 5 and 7, we see that the pairwise distributions are treated similarly. Thus, following the same steps in Section 5.1.3, we conduct the risk rewrite for DU learning.

Step 1: Corrected Loss Design and Risk Rewrite for DU Learning.
The following corollary is a variant of Corollary 23.

Corollary 27. *Assume $\pi_{\mathrm{p}} = 1/2$. For DU learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{\tilde{\mathrm{D}}}\left[\bar{\ell}_{\tilde{\mathrm{D}}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right],$$

where

$$\begin{aligned} \bar{\ell}_{\tilde{\mathrm{D}}} &= 2\pi_{\mathrm{p}}\pi_{\mathrm{n}}\left(\frac{1}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}}\ell_{\mathrm{p}} - \frac{1}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}}\ell_{\mathrm{n}}\right), \\ \bar{\ell}_{\mathrm{U}} &= -\frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}}\ell_{\mathrm{p}} + \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}}\ell_{\mathrm{n}}. \end{aligned} \tag{98}$$

Step 2: Recovering the previous result(s) for DU Learning.
We reuse the trick in Lemma 24 for restoring the pairwise distribution P_{S} to restore P_{D} needed here, allowing us to recover the rewrite (15) in Theorem 1 of Shimada et al. (2021) and the first result in Theorem 7.3 of Sugiyama et al. (2022). The derivation resembles that of SU learning. We start with the next lemma, adapted from Lemma 24.

Lemma 28. *Given B (46) and following the DU learning notations, we have*

$$M_{\mathrm{DU}}B = \begin{pmatrix} P_{\tilde{\mathrm{D}}} \\ P_{\mathrm{U}} \end{pmatrix} = \bar{P},$$

[8]The matching is as follows: P_{Sup} is $\tilde{p}_{+}(x)$, P_{Inf} is $\tilde{p}_{-}(x)$, ℓ_{p} is $\ell(f(x), +1)$, and ℓ_{n} is $\ell(f(x), -1)$.

where

$$M_{\mathrm{DU}} := \begin{pmatrix} \frac{\int_x \mathsf{X} P_{x\ |Y=\mathrm{n}} \mathrm{d}x}{2} & \frac{\int_x \mathsf{X} P_{x\ |Y=\mathrm{p}} \mathrm{d}x}{2} \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix}.$$

We apply Lemma 28 to slightly revise the derivation of (89) as follows:

$$\begin{aligned} R(g) &= \int_x \mathsf{X} \bar{L}\ \bar{P}\, \mathrm{d}x = \int_x \mathsf{X} \bar{L}\ M_{\mathrm{DU}} B\, \mathrm{d}x \\ &= \int_x \mathsf{X} \begin{pmatrix} \bar{\ell}_{\tilde{\mathrm{D}}} & \bar{\ell}_{\mathrm{U}} \end{pmatrix} \begin{pmatrix} \frac{\int_x \mathsf{X} P_{x\ |Y=\mathrm{n}} \mathrm{d}x}{2} & \frac{\int_x \mathsf{X} P_{x\ |Y=\mathrm{p}} \mathrm{d}x}{2} \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} P_{x|Y=\mathrm{p}} \\ P_{x|Y=\mathrm{n}} \end{pmatrix} \mathrm{d}x \\ &= \int_x \mathsf{X} \int_x \mathsf{X} P_{\mathrm{D}}\, \bar{\ell}_{\tilde{\mathrm{D}}}\, \mathrm{d}x\, \mathrm{d}x + \int_x \mathsf{X} P_{\mathrm{U}}\, \bar{\ell}_{\mathrm{U}}\, \mathrm{d}x \\ &= \mathsf{E}_{\mathrm{D}}\left[\bar{\ell}_{\tilde{\mathrm{D}}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right], \end{aligned}$$

where the second to last equality follows from the DU formulation (12). Denoting

$$L_+(X) := \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{p}}(X) - \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{n}}(X), \tag{99}$$

recalling $L(X)$ from (93), and continuing with (98), we have

$$\begin{aligned} \mathsf{E}_{\mathrm{D}}\left[\bar{\ell}_{\tilde{\mathrm{D}}}\right] &= 2\pi_{\mathrm{p}}\pi_{\mathrm{n}} \mathsf{E}_{\mathrm{D}}\left[\frac{1}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}} \ell_{\mathrm{p}} - \frac{1}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}} \ell_{\mathrm{n}}\right] \\ &= 2\pi_{\mathrm{p}}\pi_{\mathrm{n}} \mathsf{E}_{\mathrm{D}}\left[-L(X)\right] \\ &\overset{\text{(a)}}{=} 2\pi_{\mathrm{p}}\pi_{\mathrm{n}} \mathsf{E}_{\mathrm{D}}\left[-\frac{L(X) + L(X')}{2}\right] \end{aligned}$$

and

$$\begin{aligned} \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right] &= \mathsf{E}_{\mathrm{U}}\left[-\frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}} \ell_{\mathrm{p}} + \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}} \ell_{\mathrm{n}}\right] \\ &= \mathsf{E}_{\mathrm{U}}\left[L_+(X)\right] \end{aligned} \tag{100}$$

that prove rewrite (13) in Section 2.2.5. By matching notations, we recover (15) in Theorem 1 of Shimada et al. (2021) [9] . Equality (a) follows from the next lemma.

Lemma 29. *Let* (x, x') P_{D} *defined in (12). Then,* $\mathsf{E}_{\mathrm{D}}\left[\frac{L(X)}{2}\right] = \mathsf{E}_{\mathrm{D}}\left[\frac{L(X')}{2}\right]$.

Secondly, we consider the rewrite of SD learning. To do so, we apply the knowledge acquired from SU and DU learning (Corollaries 23 and 27).

Step 1: Corrected Loss Design and Risk Rewrite for SD Learning.
We provide another variant of Corollary 23 to conduct the risk rewrite.

Corollary 30. *Assume* $\pi_{\mathrm{p}} = 1/2$. *For SD learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{\tilde{\mathrm{S}}}\left[\bar{\ell}_{\tilde{\mathrm{S}}}\right] + \mathsf{E}_{\tilde{\mathrm{D}}}\left[\bar{\ell}_{\tilde{\mathrm{D}}}\right],$$

where

$$\begin{aligned} \bar{\ell}_{\tilde{\mathrm{S}}} &= \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\left(\frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{p}} - \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{n}}\right), \\ \bar{\ell}_{\tilde{\mathrm{D}}} &= 2\pi_{\mathrm{p}}\pi_{\mathrm{n}}\left(-\frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{p}} + \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{n}}\right). \end{aligned} \tag{101}$$

[9] The matching to the notations of Shimada et al. (2021) is as follows: π_{p} is π_+, π_{n} is π_-, $\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2$ is π_{S}, $2\pi_{\mathrm{p}}\pi_{\mathrm{n}}$ is π_{D}, P_{S} is $p_{\mathrm{S}}(x, x')$, P_{D} is $p_{\mathrm{D}}(x, x')$, P_{U} is $p_{\mathrm{U}}(x)$, ℓ_{p} is $\ell(f(X), +1)$, ℓ_{n} is $\ell(f(X), -1)$, $L(X)$ is $\tilde{L}(f(X))$, $L_+(X)$ is $L(f(X), +1)$, and $L_-(X)$ is $L(f(X), -1)$.

Step 2: Recovering the previous result(s) for SD Learning.
We apply the same strategy as in Lemma 28 to obtain the needed $\boldsymbol{P}_{\mathrm{S}}$ and $\boldsymbol{P}_{\mathrm{D}}$. We begin with the next lemma, adapted from Lemma 24, to recover (16) in Theorem 1 of Shimada et al. (2021) and the second result in Theorem 7.3 of Sugiyama et al. (2022).

Lemma 31. *Given B (46) and following the SD learning notations, we have*

$$M_{\mathrm{SD}} B = \begin{pmatrix} \boldsymbol{P}_{\tilde{\mathrm{S}}} \\ \boldsymbol{P}_{\tilde{\mathrm{D}}} \end{pmatrix} = \bar{P},$$

where

$$M_{\mathrm{SD}} := \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2 \int_{x \in \mathcal{X}} \boldsymbol{P}_{x\,|Y=\mathrm{p}} \mathrm{d}x}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2 \int_{x \in \mathcal{X}} \boldsymbol{P}_{x\,|Y=\mathrm{n}} \mathrm{d}x}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2} \\ \frac{\int_{x \in \mathcal{X}} \boldsymbol{P}_{x\,|Y=\mathrm{n}} \mathrm{d}x}{2} & \frac{\int_{x \in \mathcal{X}} \boldsymbol{P}_{x\,|Y=\mathrm{p}} \mathrm{d}x}{2} \end{pmatrix}.$$

We apply Lemma 31 to slightly revise the derivation of (89) as follows:

$$\begin{aligned} R(g) &= \int_{x \in \mathcal{X}} \bar{L} \;\; \bar{P} \, \mathrm{d}x = \int_{x \in \mathcal{X}} \bar{L} \;\; M_{\mathrm{SD}} B \, \mathrm{d}x \\ &= \int_{x \in \mathcal{X}} \begin{pmatrix} \bar{\ell}_{\tilde{\mathrm{S}}} & \bar{\ell}_{\tilde{\mathrm{D}}} \end{pmatrix} \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2 \int_{x \in \mathcal{X}} \boldsymbol{P}_{x\,|Y=\mathrm{p}} \mathrm{d}x}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2 \int_{x \in \mathcal{X}} \boldsymbol{P}_{x\,|Y=\mathrm{n}} \mathrm{d}x}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2} \\ \frac{\int_{x \in \mathcal{X}} \boldsymbol{P}_{x\,|Y=\mathrm{n}} \mathrm{d}x}{2} & \frac{\int_{x \in \mathcal{X}} \boldsymbol{P}_{x\,|Y=\mathrm{p}} \mathrm{d}x}{2} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{x|Y=\mathrm{p}} \\ \boldsymbol{P}_{x|Y=\mathrm{n}} \end{pmatrix} \mathrm{d}x \\ &= \int_{x \in \mathcal{X}} \int_{x \in \mathcal{X}} \boldsymbol{P}_{\mathrm{S}} \, \bar{\ell}_{\tilde{\mathrm{S}}} \, \mathrm{d}x \, \mathrm{d}x + \int_{x \in \mathcal{X}} \int_{x \in \mathcal{X}} \boldsymbol{P}_{\mathrm{D}} \, \bar{\ell}_{\tilde{\mathrm{D}}} \, \mathrm{d}x \, \mathrm{d}x \\ &= \mathsf{E}_{\mathrm{S}} \left[\bar{\ell}_{\tilde{\mathrm{S}}} \right] + \mathsf{E}_{\mathrm{D}} \left[\bar{\ell}_{\tilde{\mathrm{D}}} \right], \end{aligned}$$

where the second to last equality follows from the SD formulation (14). Recalling $\boldsymbol{L}_+(X)$ (99) and $\boldsymbol{L}_-(X)$ (94) and continuing with (101),

$$\begin{aligned} \mathsf{E}_{\mathrm{S}} \left[\bar{\ell}_{\tilde{\mathrm{S}}} \right] &= \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2 \right) \mathsf{E}_{\mathrm{S}} \left[\frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{p}} - \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{n}} \right] \\ &= \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2 \right) \mathsf{E}_{\mathrm{S}} \left[\boldsymbol{L}_+(X) \right] \\ &\overset{\text{(b)}}{=} \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2 \right) \mathsf{E}_{\mathrm{S}} \left[\frac{\boldsymbol{L}_+(X) + \boldsymbol{L}_+(X')}{2} \right] \end{aligned}$$

and

$$\begin{aligned} \mathsf{E}_{\mathrm{D}} \left[\bar{\ell}_{\tilde{\mathrm{D}}} \right] &= 2\pi_{\mathrm{p}}\pi_{\mathrm{n}} \mathsf{E}_{\mathrm{D}} \left[-\frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{p}} + \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{n}} \right] \\ &= 2\pi_{\mathrm{p}}\pi_{\mathrm{n}} \mathsf{E}_{\mathrm{D}} \left[\boldsymbol{L}_-(X) \right] \\ &\overset{\text{(c)}}{=} 2\pi_{\mathrm{p}}\pi_{\mathrm{n}} \mathsf{E}_{\mathrm{D}} \left[\frac{\boldsymbol{L}_-(X) + \boldsymbol{L}_-(X')}{2} \right] \end{aligned}$$

prove rewrite (15) in Section 2.2.6. We also recover (16) in Theorem 1 of Shimada et al. (2021) via matching notations. The required matches can be found in the paragraph before Lemma 29. The equality (b) holds by applying Lemma 25 with $\boldsymbol{L}(X)$ replaced by $\boldsymbol{L}_+(X)$, and (c) follows from Lemma 29 with $\boldsymbol{L}(X)$ replaced by $\boldsymbol{L}_-(X)$.

An intriguing observation worth mentioning is that the losses $\boldsymbol{L}_+(X)$ and $\boldsymbol{L}_-(X)$ applied to decontaminate the unlabeled data in SU and DU learning ((95) and (100)) are now used to decontaminate the similar and the dissimilar data in SD learning, respectively. One can also quickly draw the same conclusion from Table 4. Knowing the reason behind this observation would help to transfer one corrected loss developed in one scenario to another weakly supervised scenario.

5.1.6 Similarity-Confidence (Sconf) Learning

Since M_{Sconf} (55) is not a child of M_{UU} (49) on the reduction graph, a direct application of Theorem 21 is infeasible. Nevertheless, we demonstrate how our framework is applied to rewrite the classification risk for Sconf learning. We make a small adjustment to the framework that instead of showing $\bar{L}\ \ \bar{P} = L\ \ M^\dagger \bar{P} = L\ \ P$, we show that for loss vector $\bar{L}$ with a certain property,

$$\int_x \ \ _{\mathsf{X}} \bar{L}\ \ \bar{P}\,\mathrm{d}x\ = \bar{L}\ \ \tilde{M}_{\mathrm{Sconf}} P. \tag{102}$$

The idea behind this approach is to accommodate x sampled from P_X (18). Suppose, informally, we have the equation above. Then, the right-hand side of (102) will produce $L\ \ P$ if we can compute a decontamination matrix $\tilde{M}^\dagger_{\mathrm{Sconf}}$ satisfying $\tilde{M}^\dagger_{\mathrm{Sconf}} \tilde{M}_{\mathrm{Sconf}} = I$ and assign $\bar{L}\ \ := L\ \ \tilde{M}^\dagger_{\mathrm{Sconf}}$. Lastly, integrating over x on both sides, we obtain the key equation

$$\int_x \ \ _{\mathsf{X}} \int_x \ \ _{\mathsf{X}} \bar{L}\ \ \bar{P}\,\mathrm{d}x\,\mathrm{d}x = \int_x \ \ _{\mathsf{X}} L\ \ P\,\mathrm{d}x$$

for risk rewrite.

Step 1: Corrected Loss Design and Risk Rewrite.
Let us follow the notations in Section 4.1.6. We begin with two technical lemmas and leave their proofs to Appendix D.1. The first technical lemma shows how to achieve (102).

Lemma 32. *Assume the formulation $\bar{P} = M_{\mathrm{Sconf}} B$ (58) is given. Suppose a vector of corrected losses $\bar{L}$ of the form $\begin{pmatrix} \tilde{\ell}_1(x) & \tilde{\ell}_2(x) \end{pmatrix}$ is independent of x. Then, we have*

$$\int_x \ \ _{\mathsf{X}} \bar{L}\ \ \bar{P}\,\mathrm{d}x\ = \bar{L}\ \ \tilde{M}_{\mathrm{Sconf}} P, \tag{103}$$

where

$$\tilde{M}_{\mathrm{Sconf}} = \begin{pmatrix} \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{r - \pi_{\mathrm{n}}} \mathrm{d}x & \int_x \frac{\pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{p}}}{r - \pi_{\mathrm{n}}} \mathrm{d}x \\ \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}}}{\pi_{\mathrm{p}} - r} \mathrm{d}x & \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{\pi_{\mathrm{p}} - r} \mathrm{d}x \end{pmatrix}.$$

The second technical lemma computes the decontamination matrix.

Lemma 33. *Let*

$$\tilde{M}^\dagger_{\mathrm{Sconf}} := \begin{pmatrix} \frac{r - \pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} & 0 \\ 0 & \frac{\pi_{\mathrm{p}} - r}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \end{pmatrix}.$$

Then,

$$\tilde{M}^\dagger_{\mathrm{Sconf}} \tilde{M}_{\mathrm{Sconf}} = I.$$

Next, we follow the sketch above to instantiate the corrected losses as

$$\bar{L}\ \ := L\ \ \tilde{M}^\dagger_{\mathrm{Sconf}} = \begin{pmatrix} \frac{r - \pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{p}}(X) & \frac{\pi_{\mathrm{p}} - r}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{n}}(X) \end{pmatrix}.$$

Putting $\tilde{M}^\dagger_{\mathrm{Sconf}}$, $\bar{L}$, and (103) together, we have the following rewrite.

Theorem 34. *Assume $\pi_{\mathrm{p}} = 1/2$. The classification risk of Sconf learning can be expressed by*

$$R(g) = \mathsf{E}_{X,X}\ \left[\frac{r - \pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{p}}(X) + \frac{\pi_{\mathrm{p}} - r}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}} \ell_{\mathrm{n}}(X) \right]. \tag{104}$$

Proof. Integrating both sides of (103) over x and applying Lemma 33, we obtain

$$\begin{aligned}\int_x \int_{x'} \bar{L} \ \bar{P} \, \mathrm{d}x \, \mathrm{d}x' &= \int_x \int_{x'} \bar{L} \ \tilde{M}_{\mathrm{Sconf}} P \, \mathrm{d}x \\ &= \int_x \int_{x'} L \ \tilde{M}_{\mathrm{Sconf}}^{-1} \tilde{M}_{\mathrm{Sconf}} P \, \mathrm{d}x = R(g).\end{aligned}$$

On the other hand, substituting $\bar{L}$ with $\begin{pmatrix} \frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) & \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X) \end{pmatrix}$ and $\bar{P}$ with $\begin{pmatrix} \boldsymbol{P}_X \boldsymbol{P}_X \\ \boldsymbol{P}_X \boldsymbol{P}_X \end{pmatrix}$,

$$\begin{aligned}\int_x \int_{x'} \bar{L} \ \bar{P} \, \mathrm{d}x \, \mathrm{d}x' &= \int_x \int_{x'} \boldsymbol{P}_x \boldsymbol{P}_{x'} \left(\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{p}}(x) + \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{n}}(x) \right) \mathrm{d}x \, \mathrm{d}x' \\ &= \mathsf{E}_{X,X'} \left[\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) + \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X) \right]\end{aligned}$$

completes the proof of the theorem. □

Step 2: Recovering the previous result(s).
From the above derivation, we have achieved the first half of the rewrite in (19). Notice that (56) can be rephrased as

$$\left(\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}} \right) \boldsymbol{P}_X \boldsymbol{P}_X = \left(\pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|\mathrm{p}} - \pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|\mathrm{n}} \right) \boldsymbol{P}_{X'|\mathrm{p}} + \left(\pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|\mathrm{n}} - \pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|\mathrm{p}} \right) \boldsymbol{P}_{X'|\mathrm{n}}$$

and that (57) can be rephrased as

$$\left(\frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{n}}} \right) \boldsymbol{P}_X \boldsymbol{P}_X = \left(\pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|\mathrm{n}} - \pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|\mathrm{p}} \right) \boldsymbol{P}_{X'|\mathrm{p}} + \left(\pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|\mathrm{p}} - \pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|\mathrm{n}} \right) \boldsymbol{P}_{X'|\mathrm{n}}.$$

Thus, when $\pi_{\mathrm{p}} = 1/2$, we can repeat the proof steps in Lemma 9 to rephrase (58) as

$$\begin{pmatrix} \boldsymbol{P}_X \boldsymbol{P}_X \\ \boldsymbol{P}_X \boldsymbol{P}_X \end{pmatrix} = \begin{pmatrix} \frac{\pi_{\mathrm{p}}\left(\pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|\mathrm{p}} - \pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|\mathrm{n}}\right)}{r-\pi_{\mathrm{n}}} & \frac{\pi_{\mathrm{p}}\left(\pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|\mathrm{n}} - \pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|\mathrm{p}}\right)}{r-\pi_{\mathrm{n}}} \\ \frac{\pi_{\mathrm{n}}\left(\pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|\mathrm{n}} - \pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|\mathrm{p}}\right)}{\pi_{\mathrm{p}}-r} & \frac{\pi_{\mathrm{n}}\left(\pi_{\mathrm{p}}^2 \boldsymbol{P}_{X|\mathrm{p}} - \pi_{\mathrm{n}}^2 \boldsymbol{P}_{X|\mathrm{n}}\right)}{\pi_{\mathrm{p}}-r} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{X'|\mathrm{p}} \\ \boldsymbol{P}_{X'|\mathrm{n}} \end{pmatrix}.$$

Comparing the equation above with $\bar{P} = M_{\mathrm{Sconf}} B$, we see that it is still feasible to formulate $\bar{P}$ with X and X' in M_{Sconf} and B of (58) swapped. Then, repeating the same argument in **Step 1** with x and x' swapped, we obtain

$$R(g) = \mathsf{E}_{X',X} \left[\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X') + \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X') \right]. \tag{105}$$

Therefore, the following combines (104) and (105) to obtain

$$\begin{aligned} R(g) &= \frac{1}{2}(R(g) + R(g)) \\ &= \frac{1}{2}\mathsf{E}_{X,X'} \left[\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X) + \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X) \right] + \frac{1}{2}\mathsf{E}_{X,X'} \left[\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{p}}(X') + \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}}\ell_{\mathrm{n}}(X') \right] \\ &= \mathsf{E}_{X,X'} \left[\frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} \frac{\ell_{\mathrm{p}}(X) + \ell_{\mathrm{p}}(X')}{2} + \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} \frac{\ell_{\mathrm{n}}(X) + \ell_{\mathrm{n}}(X')}{2} \right]\end{aligned}$$

that recovers rewrite (19) in Section 2.2.8. By matching notations, we recover Theorem 3 of Cao et al. (2021b) [10].

[10] The matching to the notations of Cao et al. (2021b) is as follows: π_{p} is π_+, π_{n} is π_-, r is s, $\ell_{\mathrm{p}}(X)$ is $\ell(g(X), +1)$, and $\ell_{\mathrm{n}}(X)$ is $\ell(g(X), -1)$.

5.2 CCN Scenarios

The proposed framework is now applied to conduct the risk rewrites for WSLs discussed in Section 4.2 and summarized in Table 8. Counterintuitively, we demonstrate that finding an inverse matrix (e.g., Proposition 1) is not the only way to solve the risk rewrite problem. Introduced in Proposition 2, the new technique exploited in this subsection, marginal chain, calculates the decontamination matrix for (37) via applying the conditional probability formula twice during a chain of matrix multiplications.

5.2.1 Generalized CCN

We justify the proposed framework for generalized CCN learning via the following steps. Derived equations will be applied to solve the risk rewrite problem for WSLs discussed in Section 4.2.

Step 1: Corrected Loss Design.
Let us follow the notations in Section 4.2.1. Same as what we have illustrated in the beginning of Section 5.1.1, the proposed framework achieves three milestones to rewrite the risk. We apply Lemma 10 to achieve the first milestone, $\bar{P} = M_{\text{gCCN}} P$. This is done by noting that for generalized CCN, $\bar{P} = M_{\text{gCCN}} B$ and $B = P$ are given by Lemma 10 (i.e., do not need to handle M_{trsf} discussed in Section 3.2 since M_{trsf} is the identity matrix when $B = P$).

The second milestone is to find $M_{\text{gCCN}}^{\dagger}$ to achieve $M_{\text{gCCN}}^{\dagger} \bar{P} = P$. Since M_{gCCN} (64) is identical to M (40), a direct application of Proposition 2 gives the decontamination matrix

$$M_{\text{gCCN}}^{\dagger} := \begin{pmatrix} \mathsf{P}_{Y=1|S=s_1,X} & \mathsf{P}_{Y=1|S=s_2,X} & \cdots & \mathsf{P}_{Y=1|S=s_{|\mathsf{S}|},X} \\ \mathsf{P}_{Y=2|S=s_1,X} & \mathsf{P}_{Y=2|S=s_2,X} & \cdots & \mathsf{P}_{Y=2|S=s_{|\mathsf{S}|},X} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{P}_{Y=K|S=s_1,X} & \mathsf{P}_{Y=K|S=s_2,X} & \cdots & \mathsf{P}_{Y=K|S=s_{|\mathsf{S}|},X} \end{pmatrix} \tag{106}$$

that satisfies the $M_{\text{gCCN}}^{\dagger} \bar{P} = P$ requirement.

The final milestone is achieved by instantiating the corrected loss (38) as $\bar{L} := L \ M_{\text{gCCN}}^{\dagger}$. We denote the k-th entry of L is $\ell_{Y=k}$ with $k \in [K]$ and the j-th entry of $\bar{L}$ is $\bar{\ell}_{S=s_j}$ with $j \in [|\mathsf{S}|]$.

Despite Proposition 2's simplicity, the construction of $M_{\text{gCCN}}^{\dagger}$ is somewhat surprising. $M_{\text{gCCN}}^{\dagger}$, to our best knowledge, contributes to a first loss correction result relaxing the invertibility constraint. Unlike $M_{\text{UU}}^{\dagger}$ (Corollary 20), which needs to compute an inverse matrix, one can construct $M_{\text{gCCN}}^{\dagger}$ by calculating each entry $\mathsf{P}_{Y|S,X}$ in (106), to which, we point out a systematic way in Section 5.2.2.

Step 2: Classification Risk Rewrite.
With $\bar{L}$ in hand, the following theorem provides an intermediate form of risk rewrite.

Theorem 35. *Let $\bar{P}$ and P are given by (62) and (63), respectively. Denote $\bar{L} := L M_{\text{gCCN}}^{\dagger}$. Then, $\bar{L} \ \bar{P} = L \ P$ and*

$$R(g) = \int_{\mathcal{X}} L \ P \mathrm{d}x = \int_{\mathcal{X}} \bar{L} \ \bar{P} \mathrm{d}x. \tag{107}$$

Proof. Since $M_{\text{gCCN}}^{\dagger}$ is given by Proposition 2, $M_{\text{gCCN}}^{\dagger} \bar{P} = P$. Thus, following the framework (39), we have $\bar{L} \ \bar{P} = L \ M_{\text{gCCN}}^{\dagger} \bar{P} = L \ P$ implying (107). □

Theorem 35 will be applied to derive the respective rewrites for WSLs discussed in Section 4.2 in the rest of this subsection. In particular, we explain how to realize $M_{\text{gCCN}}^{\dagger}$ (106) for a given CCN scenario. Then, the risk rewrite (107) automatically carries over for the scenario considered, and the respective $\bar{L}$ specifies the corrected losses in the rewrite.

5.2.2 Proper Partial-Label (PPL) Learning

$M_{\mathrm{gCCN}}^{\dagger}$ provides an abstraction for us to construct the corrected losses $\bar{L}$. Next, we focus on deriving the actual form of $\boldsymbol{P}_{Y|S,X}$ in $M_{\mathrm{gCCN}}^{\dagger}$ (106) to explicitly express $\bar{\ell}_S$ for PPL.

Step 1: Corrected Loss Design and Risk Rewrite.
Let us follow the notations in Section 4.2.2. The following lemma specifies the form of $\boldsymbol{P}_{Y|S,X}$ to instantiate $M_{\mathrm{gCCN}}^{\dagger}$.

Lemma 36. *$M_{\mathrm{PPL}}^{\dagger}$ corresponds to realizing $M_{\mathrm{gCCN}}^{\dagger}$ (106) with*

$$\boldsymbol{P}_{Y=i|S=s_j,X} := \frac{\boldsymbol{P}_{Y=i|X}\mathsf{I}\left[Y=i \quad s_j\right]}{\sum_{a \quad s_j}\boldsymbol{P}_{Y=a|X}}. \tag{108}$$

Proof. Recall that the decontamination matrix of M_{gCCN} (64) is $M_{\mathrm{gCCN}}^{\dagger}$ (106) and M_{PPL} is a reduction of M_{gCCN} via $\boldsymbol{P}_{S|Y,X} = C(S,X)\mathsf{I}\left[Y \quad S\right]$ (65). Thus, to find out the (i,j) entry of $M_{\mathrm{PPL}}^{\dagger}$, we need to find out the form of $\boldsymbol{P}_{Y=i|S=s_j,X}$ subject to (65).

Applying Theorem 1 of Wu et al. (2023) directly gives

$$\boldsymbol{P}_{Y=i|S=s_j,X} = \frac{\boldsymbol{P}_{Y=i|X}\mathsf{I}\left[Y=i \quad s_j\right]}{\sum_{a \quad s_j}\boldsymbol{P}_{Y=a|X}},$$

which completes the proof. For completeness, we provide a derivation as follows.

Note that $\boldsymbol{P}_{S|Y,X} = C(S,X)\mathsf{I}\left[Y \quad S\right]$ (65) implies

$$\begin{aligned}\sum_{b \quad Y\setminus S}\boldsymbol{P}_{S,Y=b|X} &= \sum_{b \quad Y\setminus S}\boldsymbol{P}_{S|Y=b,X}\boldsymbol{P}_{Y=b|X}\\ &= \sum_{b \quad Y\setminus S}C(S,X)\mathsf{I}\left[b \quad S\right]\boldsymbol{P}_{Y=b|X}\\ &= 0.\end{aligned}$$

Therefore, $\boldsymbol{P}_{S|X} = \sum_{a \quad S}\boldsymbol{P}_{S,Y=a|X} + \sum_{b \quad Y\setminus S}\boldsymbol{P}_{S,Y=b|X} = \sum_{a \quad S}\boldsymbol{P}_{S,Y=a|X}$. Utilizing this fact, we obtain

$$\begin{aligned}\boldsymbol{P}_{Y|S,X} = \frac{\boldsymbol{P}_{S,Y|X}}{\boldsymbol{P}_{S|X}} &= \frac{\boldsymbol{P}_{S|Y,X}\boldsymbol{P}_{Y|X}}{\sum_{a \quad S}\boldsymbol{P}_{S|Y=a,X}\boldsymbol{P}_{Y=a|X}}\\ &= \frac{C(S,X)\mathsf{I}\left[Y \quad S\right]\boldsymbol{P}_{Y|X}}{\sum_{a \quad S}C(S,X)\mathsf{I}\left[Y=a \quad S\right]\boldsymbol{P}_{Y=a|X}}\\ &= \frac{\boldsymbol{P}_{Y|X}\mathsf{I}\left[Y \quad S\right]}{\sum_{a \quad S}\boldsymbol{P}_{Y=a|X}}\end{aligned}$$

that finishes the proof for Theorem 1 of Wu et al. (2023). □

We have shown that $M_{\mathrm{PPL}}^{\dagger}$ is derived from $M_{\mathrm{gCCN}}^{\dagger}$. Thus, we can follow Theorem 35 to construct the corrected losses using (108) and obtain the risk rewrite (27) for PPL in Section 2.2.12.

Corollary 37. *Given $M_{\mathrm{PPL}}^{\dagger}$ defined by (108), we denote the corrected losses $\bar{L} := L \; M_{\mathrm{PPL}}^{\dagger}$. Then, for PPL learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{S,X}\left[\bar{\ell}_S\right],$$

where

$$\bar{\ell}_S = \sum_{i \quad S}\frac{\boldsymbol{P}_{Y=i|X}}{\sum_{a \quad S}\boldsymbol{P}_{Y=a|X}}\ell_{Y=i}. \tag{109}$$

Proof. Given (108), the j-th entry of $\bar{L}$ is of the form

$$\begin{aligned}\bar{\ell}_{S=s_j} = \left(L \quad M_{\mathrm{PPL}}^{\dagger}\right)_j &= \sum_{i=1}^{K} \frac{P_{Y=i|X}\,\mathsf{I}\left[Y=i \quad s_j\right]}{\sum_a \quad s_j P_{Y=a|X}} \ell_{Y=i} \\ &= \sum_{i \quad s_j} \frac{P_{Y=i|X}}{\sum_a \quad s_j P_{Y=a|X}} \ell_{Y=i}.\end{aligned}$$

Then, since $M_{\mathrm{PPL}}^{\dagger}$ is a realization of $M_{\mathrm{gCCN}}^{\dagger}$ according to Lemma 36, we continue (107) to express the risk as

$$R(g) = \int_{x \quad X} \bar{L} \quad \bar{P}\mathrm{d}x = \int_{x \quad X} \sum_{j=1}^{|S|} P_{S=s_j,x} \bar{\ell}_{S=s_j} \mathrm{d}x = \mathsf{E}_{S,X}\left[\bar{\ell}_S\right].$$

□

Step 2: Recovering the previous result(s).
We finish this part by pointing out Corollary 37 recovers Theorem 3 of Wu et al. (2023).

5.2.3 Provably Consistent Partial-Label (PCPL) Learning

It is fairly straightforward to apply the proposed framework to rewrite the classification risk. However, it is more involved in recovering the existing result.

Step 1: Corrected Loss Design and Risk Rewrite.
The argument for obtaining the risk rewrite for PCPL is similar to that of PPL. From Section 4.2.3 we know that PCPL is a special case of PPL that only differs in the choice of $C(S, X)$. Since $C(S, X)$ is independent of (108), $M_{\mathrm{PCPL}}^{\dagger}$ and $M_{\mathrm{PPL}}^{\dagger}$ are identical. Hence, following the notations in Section 4.2.3 and repeating the proof of Corollary 37, we obtain the risk rewrite for PCPL:

Corollary 38. *The decontamination matrix $M_{\mathrm{PCPL}}^{\dagger}$ for PCPL equals $M_{\mathrm{PPL}}^{\dagger}$. If we define the corrected losses as $\bar{L} := L \quad M_{\mathrm{PCPL}}^{\dagger}$, the classification risk for PCPL learning can be rewritten as*

$$R(g) = \mathsf{E}_{S,X}\left[\bar{\ell}_S\right],$$

where

$$\bar{\ell}_S = \sum_{i \quad S} \frac{P_{Y=i|X}}{\sum_a \quad S P_{Y=a|X}} \ell_{Y=i}. \tag{110}$$

Step 2: Recovering the previous result(s).
In order to recover (8) of Feng et al. (2020b), we need to reorganize the sum in (110) by leveraging a unique property of a pair of partial-labels $(s, s\,)$ that complement each other. The following technical lemma states the required property, with proof deferred to Appendix D.2.

Lemma 39. *Let $(s, s\,)$ be a pair of partial-labels satisfying $s = Y \setminus s$. Then,*

$$P_{S=s,X}\bar{\ell}_{S=s} + P_{S=s\,,X}\bar{\ell}_{S=s} = P_{S=s,X} \sum_{i=1}^{K} \frac{P_{Y=i|X}\ell_{Y=i}}{\sum_a \quad s P_{Y=a|X}}.$$

Denote $s_j := Y \setminus s_j$ for every $s_j \quad S$. Then, Lemma 39 implies

$$\begin{aligned}\sum_{j=1}^{|S|} 2P_{S=s_j,X}\bar{\ell}_{S=s_j} &= \sum_{j=1}^{|S|} \left(P_{S=s_j,X}\bar{\ell}_{S=s_j} + P_{S=s_j,X}\bar{\ell}_{S=s_j}\right) \\ &= \sum_{j=1}^{|S|} P_{S=s_j,X} \sum_{i=1}^{K} \frac{P_{Y=i|X}\ell_{Y=i}}{\sum_a \quad s_j P_{Y=a|X}}.\end{aligned}$$

Hence, continuing from Corollary 38,

$$\begin{aligned}
\mathsf{E}_{S,X}\left[\bar{\ell}_S\right] &= \int_x \sum_{j=1}^{|\mathcal{S}|} P_{S=s_j,x}\bar{\ell}_{S=s_j}\mathrm{d}x \\
&= \frac{1}{2}\int_x \sum_{j=1}^{|\mathcal{S}|} P_{S=s_j,x}\sum_{i=1}^{K}\frac{P_{Y=i|x}\ell_{Y=i}}{\sum_a P_{Y=a|x}}\mathrm{d}x \\
&= \frac{1}{2}\mathsf{E}_{S,X}\left[\sum_{i=1}^{K}\frac{P_{Y=i|X}}{\sum_a P_{Y=a|X}}\ell_{Y=i}\right]
\end{aligned}$$

shows that the rewrite from the framework recovers (25) in Section 2.2.11. By matching notations, we also recover (8) of Feng et al. (2020b)[11].

5.2.4 Multi-Complementary-Label (MCL) Learning

Step 1: Corrected Loss Design and Risk Rewrite.
Let us follow the notations in Section 4.2.4. As discussed in Section 4.2.4, MCL is a special case of PPL. Thus, we can modify Lemma 36 based on the notations in Section 4.2.4 to construct the decontamination matrix $M_{\mathrm{MCL}}^{\dagger}$ for MCL. Then, following the same steps for proving Corollary 37, we instantiate $\bar{L}$ to conduct the risk rewrite for MCL:

Corollary 40. *The (i,j) entry of the decontamination matrix $M_{\mathrm{MCL}}^{\dagger}$ is of the form*

$$P_{Y=i|\bar{S}=\bar{s}_j,X} = \frac{P_{Y=i|X}\,\mathsf{I}\left[Y=i \;/\; \bar{s}_j\right]}{\sum_{a\,/\,\bar{s}_j} P_{Y=a|X}}. \tag{111}$$

Define the corrected losses $\bar{L} := L\ M_{\mathrm{MCL}}^{\dagger}$. Then, for MCL learning, the classification risk can be rewritten as

$$R(g) = \mathsf{E}_{\bar{S},X}\left[\bar{\ell}_{\bar{S}}\right],$$

where

$$\bar{\ell}_{\bar{S}} = \sum_{i\,/\,\bar{S}} \frac{P_{Y=i|X}}{\sum_{a\,/\,\bar{S}} P_{Y=a|X}}\ell_{Y=i}. \tag{112}$$

Step 2: Recovering the previous result(s).
Although legitimate, the risk rewrite (112) following the marginal chain approach appears different from Theorem 3 of Feng et al. (2020a), to which we resort to the inversion approach (Proposition 1) that finds another decontamination matrix, termed M_{MCL}^{-1}, to recover. As a preparation step, we denote N_d as the number of multi-complementary-labels with size d and group rows of M_{MCL} (71) by the size of labels as follows.

$$M_{\mathrm{MCL}} = \begin{pmatrix} P_{|\bar{S}|=1}M_1 \\ P_{|\bar{S}|=2}M_2 \\ \vdots \\ P_{|\bar{S}|=K-1}M_{K-1} \end{pmatrix}, \tag{113}$$

[11] The matching to the notations of Feng et al. (2020b) is as follows: $P_{S,X}$ is $\tilde{p}(x,Y)$, $P_{Y=i|X}$ is $p(y=i|x)$, and $\ell_{Y=i}$ is $L(f(x),i)$.

where for $d \in [K-1]$, each block is of the form[12]

$$M_d = \frac{1}{\binom{K-1}{d}} \begin{pmatrix} \mathsf{I}[Y=1 \mid \bar{s}_{d,1}] & \mathsf{I}[Y=2 \mid \bar{s}_{d,1}] & \cdots & \mathsf{I}[Y=K \mid \bar{s}_{d,1}] \\ \mathsf{I}[Y=1 \mid \bar{s}_{d,2}] & \mathsf{I}[Y=2 \mid \bar{s}_{d,2}] & \cdots & \mathsf{I}[Y=K \mid \bar{s}_{d,2}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{I}[Y=1 \mid \bar{s}_{d,N_d}] & \mathsf{I}[Y=2 \mid \bar{s}_{d,N_d}] & \cdots & \mathsf{I}[Y=K \mid \bar{s}_{d,N_d}] \end{pmatrix}. \tag{114}$$

To maintain the equality $\bar{P} = M_{\mathrm{MCL}} P$ established in Lemma 13, we also rearrange $\bar{P}$ (69) as

$$\begin{pmatrix} \mathsf{P}_{\bar{S}=\bar{s}_{1,1},X} & \cdots \mathsf{P}_{\bar{S}=\bar{s}_{1,N_1},X} & \cdots \mathsf{P}_{\bar{S}=\bar{s}_{K-1,1},X} & \cdots \mathsf{P}_{\bar{S}=\bar{s}_{K-1,N_{K-1}},X} \end{pmatrix}. \tag{115}$$

As a sanity check, we see that for any $d \in [K-1]$ and $j \in [N_d]$,

$$\begin{aligned} \left(\mathsf{P}_{|\bar{S}|=d}\, M_d\, P\right)_j &= \mathsf{P}_{|\bar{S}|=d} \cdot \frac{1}{\binom{K-1}{d}} \sum_Y \mathsf{I}[Y \mid \bar{s}_{d,j}]\, \mathsf{P}_{Y,X} \\ &= \sum_{d=1}^{K-1} \mathsf{P}_{|\bar{s}_{d,j}|=d} \cdot \frac{1}{\binom{K-1}{d}} \sum_{Y \in \bar{s}_{d,j}} \mathsf{P}_{Y,X}\, \mathsf{I}[|\bar{s}_{d,j}| = d] \\ &= \mathsf{P}_{\bar{S}=\bar{s}_{d,j},X}. \end{aligned} \tag{116}$$

The next lemma is crucial for us to devise the decontamination matrix M_{MCL}^{-1} via the inversion approach. We defer its proof to the later part of this sub-subsection.

Lemma 41. *Let $i^\star \in Y$ be fixed. Then, for every $d \in [K-1]$,*

$$\mathsf{P}_{Y=i^\star,X} = \sum_{j=1}^{N_d} \left(1 - \frac{K-1}{d} \mathsf{I}\left[Y = i^\star \mid \bar{S} = \bar{s}_{d,j}\right]\right) \mathsf{P}_{\bar{S}=\bar{s}_{d,j},X \mid\mid |\bar{S}|=d}.$$

Moreover, the inverse matrix M_d^{-1} of M_d (114) is of the form

$$\begin{pmatrix} 1 - \frac{K-1}{d}\mathsf{I}[Y=1 \mid \bar{s}_{d,1}] & 1 - \frac{K-1}{d}\mathsf{I}[Y=1 \mid \bar{s}_{d,2}] & \cdots & 1 - \frac{K-1}{d}\mathsf{I}[Y=1 \mid \bar{s}_{d,N_d}] \\ 1 - \frac{K-1}{d}\mathsf{I}[Y=2 \mid \bar{s}_{d,1}] & 1 - \frac{K-1}{d}\mathsf{I}[Y=2 \mid \bar{s}_{d,2}] & \cdots & 1 - \frac{K-1}{d}\mathsf{I}[Y=2 \mid \bar{s}_{d,N_d}] \\ \vdots & \vdots & \ddots & \vdots \\ 1 - \frac{K-1}{d}\mathsf{I}[Y=K \mid \bar{s}_{d,1}] & 1 - \frac{K-1}{d}\mathsf{I}[Y=K \mid \bar{s}_{d,2}] & \cdots & 1 - \frac{K-1}{d}\mathsf{I}[Y=K \mid \bar{s}_{d,N_d}] \end{pmatrix}. \tag{117}$$

Applying the lemma, we construct

$$M_{\mathrm{MCL}}^{-1} := \begin{pmatrix} M_1^{-1} & M_2^{-1} & \cdots & M_{K-1}^{-1} \end{pmatrix} \tag{118}$$

and obtain $M_{\mathrm{MCL}}^{-1} \bar{P} = P$ since $\bar{P} = M_{\mathrm{MCL}} P$ (116) and

$$M_{\mathrm{MCL}}^{-1} M_{\mathrm{MCL}} = \sum_{d=1}^{K-1} M_d^{-1} \mathsf{P}_{|\bar{S}|=d} M_d = \sum_{d=1}^{K-1} \mathsf{P}_{|\bar{S}|=d} M_d^{-1} M_d = \sum_{d=1}^{K-1} \mathsf{P}_{|\bar{S}|=d} I = I.$$

We remark that M_{MCL}^{-1} plays the same role as $M_{\mathrm{MCL}}^{\dagger}$ realized by (111), as they both are decontamination matrices (designed to convert $\bar{P}$ back to P and used to construct the corrected losses $\bar{L}$). Distinct symbols

[12] Comparing to (71) where we use one index to denote a total of $|\mathcal{S}|$ partial-labels, M_d uses a pair of indices d and j to denote the j-th partial-label with size d. It is easy to verify that $\sum_{d=1}^{K-1} N_d = \sum_{d=1}^{K-1} \binom{K-1}{d} = 2^K - 2 = |\mathcal{S}|$.

are merely used to reflect the difference that $M_{\mathrm{MCL}}^{\dagger}$ results from the marginal chain method while M_{MCL}^{-1} comes from the inversion approach. Then, applying the framework (38), $\bar{L} \;:= L \;\; M_{\mathrm{MCL}}^{-1}$ leads to

$$\bar{L} \;\; \bar{P} = L \;\; M_{\mathrm{MCL}}^{-1} \bar{P} = L \;\; P.$$

With the corrected losses $\bar{L}$ in hand, the following theorem provides the risk rewrite (23) for MCL via the inversion approach and recovers Theorem 3 of Feng et al. (2020a)[13].

Theorem 42. *For MCL learning, the classification risk can be expressed as follows.*

$$R(g) = \mathsf{E}_{\bar{S},X}\left[\bar{\ell}_{\bar{S}}\right] = \sum_{d=1}^{K-1} \mathsf{P}_{|\bar{S}|=d} \mathsf{E}_{\bar{S},X||\bar{S}|=d}\left[\bar{\ell}_{\bar{S}}\right],$$

where

$$\bar{\ell}_{\bar{S}} = \sum_{i \,/\, \bar{S}} \ell_{Y=i} - \frac{K - 1 - |\bar{S}|}{|\bar{S}|} \sum_{\bar{s} \;\; \bar{S}} \ell_{Y=\bar{s}}.$$

Proof. We first establish

$$R(g) = \int_{\mathcal{X}} \bar{L} \;\; \bar{P} \mathrm{d}x = \mathsf{E}_{\bar{S},X}\left[\bar{\ell}_{\bar{S}}\right]$$

since $\bar{L} \;\; \bar{P} = L \;\; P$, where $\bar{P}$ is specified in (115) and $\bar{L} \;\; = L \;\; M_{\mathrm{MCL}}^{-1}$ with the $\bar{S}$-th entry being $\bar{\ell}_{\bar{S}}$. Also, recall that $\mathsf{P}_{\bar{S},X} = \sum_{d=1}^{K-1} \mathsf{P}_{|\bar{S}|=d} \mathsf{P}_{\bar{S},X||\bar{S}|=d}$ in Section 4.2.4. Thus, decomposing the probability by the size of $\bar{S}$, we have

$$\mathsf{E}_{\bar{S},X}\left[\bar{\ell}_{\bar{S}}\right] = \sum_{d=1}^{K-1} \mathsf{P}_{|\bar{S}|=d} \mathsf{E}_{\bar{S},X||\bar{S}|=d}\left[\bar{\ell}_{\bar{S}}\right].$$

Lastly, M_{MCL}^{-1} (118) and M_d^{-1} (117) imply, when $\bar{S} = \bar{s}_{d,j}$,

$$\bar{\ell}_{\bar{S}=\bar{s}_{d,j}} = \left(L \;\; M_d^{-1}\right)_j = \sum_{i=1}^{K} \ell_{Y=i} \left(1 - \frac{K-1}{d} \mathsf{I}\left[Y = i \quad \bar{s}_{d,j}\right]\right) = \sum_{i=1}^{K} \ell_{Y=i} - \frac{K-1}{d} \sum_{i \;\; \bar{s}_{d,j}} \ell_{Y=i}.$$

A simple reorganization and substituting d with $|\bar{S}|$ shows

$$\bar{\ell}_{\bar{S}} = \sum_{i \,/\, \bar{S}} \ell_{Y=i} + \sum_{i \;\; \bar{S}} \ell_{Y=i} - \frac{K-1}{|\bar{S}|} \sum_{i \;\; \bar{S}} \ell_{Y=i} = \sum_{i \,/\, \bar{S}} \ell_{Y=i} - \frac{K - 1 - |\bar{S}|}{|\bar{S}|} \sum_{i \;\; \bar{S}} \ell_{Y=i}.$$

□

Now we return to the postponed proof.

Proof. **of Lemma 41.** We start with identifying M_d^{-1}. Denote $\{\bar{s}_{d,1}, \ldots, \bar{s}_{d,N_d}\}$, the set of multi-complementary-labels of size d, as $\bar{\mathsf{S}}_d$. Let us focus on the sized-d data-generating distribution

$$\bar{P}_d = \begin{pmatrix} \mathsf{P}_{\bar{S}=\bar{s}_{d,1},X||\bar{S}|=d} \\ \vdots \\ \mathsf{P}_{\bar{S}=\bar{s}_{d,N_d},X||\bar{S}|=d} \end{pmatrix}.$$

[13]The matching to the notations of Feng et al. (2020a) is as follows: $\mathsf{P}_{\bar{S},X||\bar{S}|=d}$ is $\bar{p}(x, \bar{Y}|s = d)$, $\mathsf{P}_{|\bar{S}|=d}$ is $p(s = d)$, and $\bar{\ell}_{\bar{S}}$ is $\bar{L}_d(f(x), \bar{Y})$.

Note that $\bar{P}_d$ corresponds to extracting the entries from (115) that generate sized-d data and then dividing them by $P_{|\bar{S}|=d}$. Thus, $\bar{P} = M_{\mathrm{MCL}} P$ in Lemma 13 implies $\bar{P}_d = M_d P$ and its j-th entry is expressed as

$$P_{\bar{S}=\bar{s}_{d,j},X||\bar{S}|=d} = \frac{1}{\binom{K-1}{d}} \sum_{i=1}^{K} \mathbb{I}\left[Y = i \;/\; \bar{s}_{d,j}\right] P_{Y=i,X}. \tag{119}$$

The equality hints to us that if one manages to collect certain multi-complementary-labels $\bar{s}$ to form an equation resembling $\sum_{\bar{s}} P_{\bar{s},X||\bar{s}|=d} = c_3 \cdot P_{Y=i,X}$ for some constant c_3, then a reciprocal operation $\frac{1}{c_3}$ recovers $P_{Y=i,X}$ we need (recall we want to find M_d^{-1} achieving $M_d^{-1}\bar{P}_d = P$). To achieve such a goal, we fix on class $i^\star$ and collect elements in $\bar{\mathcal{S}}_d$ that do not contain $i^\star$ to form $E_d^{i^\star} := \left\{\bar{s}_{d,j} | \bar{s}_{d,j} \quad \bar{\mathcal{S}}_d, i^\star \;/\; \bar{s}_{d,j}\right\}$ to connect $P_{\bar{S},X||\bar{S}|=d}$ with $P_{Y=i^\star,X}$ as follows. Summing (119) over all elements in $E_d^{i^\star}$, we obtain

$$\begin{aligned}
\sum_{\bar{s} \;\; E_d^{i^\star}} P_{\bar{S}=\bar{s},X||\bar{S}|=d} &= \sum_{\bar{s} \;\; E_d^{i^\star}} \frac{1}{\binom{K-1}{d}} \sum_{i=1}^{K} \mathbb{I}\left[Y = i \;/\; \bar{s}\right] P_{Y=i,X} \\
&= \frac{1}{\binom{K-1}{d}} \left[\binom{K-2}{d} \sum_{\substack{i=1 \\ i=i^\star}}^{K} P_{Y=i,X} + \binom{K-1}{d} P_{Y=i^\star,X} \right].
\end{aligned}$$

The last equality holds since there are $\binom{K-2}{d}$ multi-complementary-labels $\bar{s} \quad \bar{\mathcal{S}}_d$ such that $i = i^\star$ and neither of them is in $\bar{s}$ (i.e., $i \;/\; \bar{s}$ and $i^\star \;/\; \bar{s}$), and there are $\binom{K-1}{d}$ multi-complementary-labels $\bar{s} \quad \bar{\mathcal{S}}_d$ such that $i = i^\star$ and i is not in $\bar{s}$. Then, we regroup the sums by pulling $\binom{K-2}{d} P_{Y=i^\star,X}$ out of $\binom{K-1}{d} P_{Y=i^\star,X}$ to combine with $\binom{K-2}{d} \sum_{\substack{i=1 \\ i=i^\star}}^{K} P_{Y=i,X}$. It leads to

$$\begin{aligned}
\sum_{\bar{s} \;\; E_d^{i^\star}} P_{\bar{S}=\bar{s},X||\bar{S}|=d} &= \frac{1}{\binom{K-1}{d}} \left[\binom{K-2}{d} \sum_{i=1}^{K} P_{Y=i,X} + \binom{K-2}{d-1} P_{Y=i^\star,X} \right] \\
&= \frac{K-1-d}{K-1} P_X + \frac{d}{K-1} P_{Y=i^\star,X}.
\end{aligned} \tag{120}$$

Denoting $\bar{\mathcal{S}}_d \backslash E_d^{i^\star} = \{\bar{s}_{d,j} | \bar{s}_{d,j} \quad \bar{\mathcal{S}}_d, i^\star \quad \bar{s}_{d,j}\}$ as $I_d^{i^\star}$ and rearranging terms in the above equation according to the reciprocal idea illustrated above, we have

$$\begin{aligned}
P_{Y=i^\star,X} &= \frac{K-1}{d} \left(\sum_{\bar{s} \;\; E_d^{i^\star}} P_{\bar{S}=\bar{s},X||\bar{S}|=d} - \frac{K-1-d}{K-1} P_X \right) \qquad (121) \\
&\overset{(a)}{=} \frac{K-1}{d} \left(P_X - \sum_{\bar{s} \;\; I_d^{i^\star}} P_{\bar{S}=\bar{s},X||\bar{S}|=d} - \frac{K-1-d}{K-1} P_X \right) \\
&= P_X - \frac{K-1}{d} \sum_{\bar{s} \;\; I_d^{i^\star}} P_{\bar{S}=\bar{s},X||\bar{S}|=d}.
\end{aligned}$$

Equality (a) holds since $|\bar{S}|$ and X are independent (Feng et al., 2020a), which implies

$$P_X = P_{X||\bar{S}|=d} = \sum_{\bar{s} \;\; \bar{\mathcal{S}}_d} P_{\bar{S}=\bar{s},X||\bar{S}|=d} = \sum_{\bar{s} \;\; E_d^{i^\star}} P_{\bar{S}=\bar{s},X||\bar{S}|=d} + \sum_{\bar{s} \;\; I_d^{i^\star}} P_{\bar{S}=\bar{s},X||\bar{S}|=d}.$$

Continuing the derivation, we have

$$\begin{aligned}
P_{Y=i^\star,X} &= \sum_{j=1}^{N_d} P_{\bar{S}=\bar{s}_{d,j},X||\bar{S}|=d} - \sum_{j=1}^{N_d} \frac{K-1}{d} \mathbb{I}\left[Y = i^\star \quad \bar{s}_{d,j}\right] P_{\bar{S}=\bar{s}_{d,j},X||\bar{S}|=d} \\
&= \sum_{j=1}^{N_d} \left(1 - \frac{K-1}{d} \mathbb{I}\left[Y = i^\star \quad \bar{s}_{d,j}\right] \right) P_{\bar{S}=\bar{s}_{d,j},X||\bar{S}|=d},
\end{aligned} \tag{122}$$

proving the first part of the lemma.

The derivation of turning (120) to (121) is a reciprocal action. Thus, if we view $1 - \frac{K-1}{d}\mathbb{I}[Y = i^\star \quad \bar{s}_{d,j}]$ as the $(i^\star, j)$ entry of some matrix M , (122) can be interpreted as $P_{i^\star} = \left(M\ \bar{P}_d\right)_{i^\star}$, suggesting $M\ M_d = I$ since $\bar{P}_d = M_d P$. We formalize this intuition in the next lemma.

Lemma 43. *Let M be of the form (117), and recall M_d is defined by (114). Then, $M\ M_d = I$, meaning $M = M_d^{-1}$.*

The above lemma finishes the proof of Lemma 41. □

Proof. **of Lemma 43.** Let d be fixed. Denoted by $A_{i,k}$, the (i,k) entry of $M\ M_d$, is the inner product of i-th row of M (117) and the k-th column of M_d (114)

$$A_{i,k} = \sum_{j=1}^{N_d} \left(1 - \frac{K-1}{d}\mathbb{I}[Y = i \quad \bar{s}_{d,j}]\right)\left(\frac{1}{\binom{K-1}{d}}\mathbb{I}[Y = k \ / \ \bar{s}_{d,j}]\right) = \sum_{j=1}^{N_d} c_{i,k}.$$

In the following, we will show that the calculation results in the identity matrix

$$A_{i,k} = \begin{cases} 1, & \text{if } i = k, \\ 0, & \text{if } i = k, \end{cases}$$

to complete the proof.

When $i = k$, we have 4 possible cases: (i) Both i and k are in $\bar{s}_{d,j}$, (ii) Both of them are not in $\bar{s}_{d,j}$, (iii) $i \quad \bar{s}_{d,j}$ and $k \ / \ \bar{s}_{d,j}$, and (iv) $i \ / \ \bar{s}_{d,j}$ and $k \quad \bar{s}_{d,j}$. For cases (i) and (iv), the coefficients $c_{i,k}$ are 0 since $\mathbb{I}[k \ / \ \bar{s}_{d,j}] = 0$ if $k \quad \bar{s}_{d,j}$. For case (ii), the coefficient $c_{i,k}$ is $\frac{1}{\binom{K-1}{d}}$. The number of such $\bar{s}_{d,j}$ is $\binom{K-2}{d}$ since we are counting the ways of forming a set of size d from $K-2$ elements. For case (iii), the coefficient $c_{i,k}$ is $\left(1 - \frac{K-1}{d}\right)\frac{1}{\binom{K-1}{d}}$. The number of such $\bar{s}_{d,j}$ is $\binom{K-2}{d-1}$ since we are counting the ways of forming a set of size $d-1$ from $k-2$ elements. Thus, if $i = k$,

$$\begin{aligned} A_{i,k} &= \frac{1}{\binom{K-1}{d}}\binom{K-2}{d} + \left(1 - \frac{K-1}{d}\right)\frac{1}{\binom{K-1}{d}}\binom{K-2}{d-1} \\ &= \frac{\binom{K-2}{d}}{\binom{K-1}{d}} + \frac{\binom{K-2}{d-1}}{\binom{K-1}{d}} - \frac{\frac{K-1}{d}\binom{K-2}{d-1}}{\binom{K-1}{d}} = 0 \end{aligned}$$

since

$$\binom{K-2}{d} + \binom{K-2}{d-1} = \binom{K-1}{d} = \frac{K-1}{d}\binom{K-2}{d-1}.$$

When $i = k$, we have 2 possible cases: (i) Both i and k are in $\bar{s}_{d,j}$, (ii) Both are not in $\bar{s}_{d,j}$. For case (i), the coefficient $c_{i,k}$ is 0. For case (ii), the coefficient $c_{i,k}$ is $\frac{1}{\binom{K-1}{d}}$, and the number of such $\bar{s}_{d,j}$ is $\binom{K-1}{d}$, as we want to form a set of size d from $K-1$ candidates. Therefore, if $i = k$,

$$A_{i,k} = \frac{1}{\binom{K-1}{d}}\binom{K-1}{d} = 1.$$

□

We want to elaborate more on the role of Theorem 1 of Wu et al. (2023) in the analyses discussed in Section 5.2. Firstly, as shown in the proof of Lemma 41, it aids the execution of the inversion approach (Proposition 1). The properness $C(S,X)\mathbb{I}[Y \quad S]$ (65) can be instantiated to define the entries of M_d (114),

which in turn establishes the key equation (120) enabling us to identify the entries of M_d^{-1} (122). Composing M_d^{-1}, we obtain M_{MCL}^{-1}, a crucial element for applying our framework (38).

Secondly, Theorem 1 of Wu et al. (2023) contributes to the marginal chain approach (Proposition 2) as well. The key equations (108) and (111) realised from Theorem 1 of Wu et al. (2023) provide the entries of $M_{\mathrm{PPL}}^{\dagger}$ (Lemma 36, Section 5.2.2), $M_{\mathrm{PCPL}}^{\dagger}$ (Section 5.2.3), and $M_{\mathrm{MCL}}^{\dagger}$ (Section 5.2.4) when applying (38). Therefore, the combined advantage of our framework and Theorem 1 of Wu et al. (2023) provides CCN scenarios unified analyses whose key steps can also be rationally interpreted. Moreover, as will be shown later, we compare the marginal chain and the inversion approaches via a CL example in Section 5.2.5. A CL example is the simplest way to convey the differences between the two methods without burying the essence in complicated derivations.

5.2.5 Complementary-Label (CL) Learning

Step 1: Corrected Loss Design and Risk Rewrite.
Note that the parameters chosen for the construction of M_{CL} (76) in Section 4.2.5 reduces M_{MCL} (113) to be M_1 of (114). That is, assigning $\mathsf{P}_{|\bar{S}|=d} = 1$ for $d = 1$, $\mathsf{P}_{|\bar{S}|=d} = 0$ for $d > 1$, and $\bar{s}_{1,j} = \{j\}$ for all $j \quad [K]$ in (113), we have

$$M_{\mathrm{MCL}} \to M_1 = \frac{1}{K-1}\begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix} = M_{\mathrm{CL}}.$$

Hence, the proof steps of Theorem 42 carry over to CL learning. With a simple rearranging on

$$\begin{aligned} \bar{\ell}_{\bar{S}} &= \sum_{i/\bar{S}} \ell_{Y=i} - \frac{K-1-|\bar{S}|}{|\bar{S}|} \sum_{\bar{s}\ \bar{S}} \ell_{Y=\bar{s}} \\ &= \sum_{i=1}^{K} \ell_{Y=i} - \frac{K-1}{|\bar{S}|} \sum_{\bar{s}\ \bar{S}} \ell_{Y=\bar{s}} \end{aligned}$$

and assigning $|\bar{S}| = 1$, we arrive at (21):

Corollary 44. *For CL learning, the classification risk can be expressed as*

$$R(g) = \mathsf{E}_{\bar{S},X}\left[\bar{\ell}_{\bar{S}}\right] = \mathsf{E}_{\bar{S},X}\left[\sum_{i=1}^{K} \ell_{Y=i} - (K-1)\ell_{\bar{S}}\right].$$

Step 2: Recovering the previous result(s).
The rewrite above recovers Theorem 1 of Ishida et al. (2019) if we substitute $\bar{S}$ with $\bar{Y}$ and $\ell_{\bar{S}}$ with $\ell(\bar{Y}, g(X))$. Moreover, if we choose $d = 1$ and $\bar{s}_{1,j} = \{j\}$ for all $j \quad [K]$, the decontamination matrix provided by (117) becomes

$$M_1^{-1} = \begin{pmatrix} -(K-2) & 1 & \cdots & 1 \\ 1 & -(K-2) & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & -(K-2) \end{pmatrix}, \tag{123}$$

which translates the corrected losses $\bar{L} \quad = L \quad M_1^{-1}$ as

$$L \quad \left(-(K-2)\mathbf{I}_K + \mathbf{11}^{\top}\right),$$

recovering (9) of Ishida et al. (2019).

Comparing inversion with marginal chain via an example.
We use a simple CL example to demonstrate the differences between the inversion (Proposition 1) and the marginal chain (Proposition 2) approaches and explain how the intuition of decontamination is implemented. Here, we focus on comparing how a decontamination matrix $M_{\text{corr}}^{\dagger}$ achieves $M_{\text{corr}}^{\dagger}\bar{P} = P$ (37) since when the equality is established, the downstream construction of the corrected losses and the risk rewrite follow the framework. For this example, let us choose $K = 4$ and simplify $\mathsf{P}_{Y=k,X}$ as p_k. Applying (76), the contamination process defining the data-generating distributions is expressed as

$$\bar{P} = M_{\text{CL}}P = \frac{1}{3}\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} \frac{p_2+p_3+p_4}{3} \\ \frac{p_1+p_3+p_4}{3} \\ \frac{p_1+p_2+p_4}{3} \\ \frac{p_1+p_2+p_3}{3} \end{pmatrix}.$$

Equation (123), simplified from (117), provides the decontamination matrix from the inversion approach:

$$M_{\text{CL}}^{-1} = \begin{pmatrix} -2 & 1 & 1 & 1 \\ 1 & -2 & 1 & 1 \\ 1 & 1 & -2 & 1 \\ 1 & 1 & 1 & -2 \end{pmatrix}.$$

Then, the inversion approach (Proposition 1) achieves the decontamination (37) by showing

$$\begin{aligned} M_{\text{CL}}^{-1}\bar{P} &= \frac{1}{3}\begin{pmatrix} -2 & 1 & 1 & 1 \\ 1 & -2 & 1 & 1 \\ 1 & 1 & -2 & 1 \\ 1 & 1 & 1 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} \\ &= \frac{1}{3}\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = P. \end{aligned} \tag{124}$$

On the other hand, equation (111) produces the decontamination matrix from the marginal chain approach:

$$M_{\text{CL}}^{\dagger} = \begin{pmatrix} \frac{0\cdot p_1}{p_2+p_3+p_4} & \frac{p_1}{p_1+p_3+p_4} & \frac{p_1}{p_1+p_2+p_4} & \frac{p_1}{p_1+p_2+p_3} \\ \frac{p_2}{p_2+p_3+p_4} & \frac{0\cdot p_2}{p_1+p_3+p_4} & \frac{p_2}{p_1+p_2+p_4} & \frac{p_2}{p_1+p_2+p_3} \\ \frac{p_3}{p_2+p_3+p_4} & \frac{p_3}{p_1+p_3+p_4} & \frac{0\cdot p_3}{p_1+p_2+p_4} & \frac{p_3}{p_1+p_2+p_3} \\ \frac{p_4}{p_2+p_3+p_4} & \frac{p_4}{p_1+p_3+p_4} & \frac{p_4}{p_1+p_2+p_4} & \frac{0\cdot p_4}{p_1+p_2+p_3} \end{pmatrix}.$$

Then, the marginal chain approach (Proposition 2) achieves the decontamination (37) by showing

$$M_{\mathrm{CL}}^{\dagger}\bar{P} = \begin{pmatrix} \frac{0\cdot p_1}{p_2+p_3+p_4} & \frac{p_1}{p_1+p_3+p_4} & \frac{p_1}{p_1+p_2+p_4} & \frac{p_1}{p_1+p_2+p_3} \\ \frac{p_2}{p_2+p_3+p_4} & \frac{0\cdot p_2}{p_1+p_3+p_4} & \frac{p_2}{p_1+p_2+p_4} & \frac{p_2}{p_1+p_2+p_3} \\ \frac{p_3}{p_2+p_3+p_4} & \frac{p_3}{p_1+p_3+p_4} & \frac{0\cdot p_3}{p_1+p_2+p_4} & \frac{p_3}{p_1+p_2+p_3} \\ \frac{p_4}{p_2+p_3+p_4} & \frac{p_4}{p_1+p_3+p_4} & \frac{p_4}{p_1+p_2+p_4} & \frac{0\cdot p_4}{p_1+p_2+p_3} \end{pmatrix} \begin{pmatrix} \frac{p_2+p_3+p_4}{3} \\ \frac{p_1+p_3+p_4}{3} \\ \frac{p_1+p_2+p_4}{3} \\ \frac{p_1+p_2+p_3}{3} \end{pmatrix} \tag{125}$$

$$= \begin{pmatrix} \frac{p_1+p_1+p_1}{3} \\ \frac{p_2+p_2+p_2}{3} \\ \frac{p_3+p_3+p_3}{3} \\ \frac{p_4+p_4+p_4}{3} \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = P.$$

Comparing (124) and (125), we see that the intuition of decontamination is realized differently. The inversion approach (124) directly cancels out the effect of M_{corr} without relying on any property of P. In contrast, the marginal chain method (125) leverages the fact that P is a probability vector and carries out a procedure similar to importance reweighting to resolve the contamination. Both methods have respective merits, and we hope the comparison will inspire new thoughts leveraging certain properties of P for the corrected loss design and the study of decontamination.

5.3 Confidence-based Scenarios

The proposed framework is now applied to conduct the risk rewrites for WSLs discussed in Section 4.3 and summarized in Table 9.

5.3.1 Subset Confidence (Sub-Conf) Learning

Step 1: Corrected Loss Design and Risk Rewrite.
Let us follow the notations in Section 4.3.1. Recall that Lemma 16 has reached the first milestone (36) by showing $\bar{P} = M_{\mathrm{Sub}}P$. To reach the second milestone (37), we apply Proposition 1 to construct the decontamination matrix $M_{\mathrm{Sub}}^{\dagger}$ to cancel out the contamination caused by M_{Sub} (80) as follows.

Lemma 45. *Assume $\mathsf{P}_{Y\ \ Y_{\mathrm{s}}|X} > 0$ for all possible outcomes of X. Choosing*

$$M_{\mathrm{Sub}}^{\dagger} := M_{\mathrm{Sub}}^{-1} = \begin{pmatrix} \frac{\mathsf{P}_{Y=1|X}}{\mathsf{P}_{Y\ \ Y_{\mathrm{s}}|X}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\mathsf{P}_{Y=K|X}}{\mathsf{P}_{Y\ \ Y_{\mathrm{s}}|X}} \end{pmatrix}, \tag{126}$$

we have $M_{\mathrm{Sub}}^{\dagger}\bar{P} = P$, where $\bar{P}$ and $M_{\mathrm{Sub}}P$ are given by Lemma 16.

Proof. The assumption $\mathsf{P}_{Y\ \ Y_{\mathrm{s}}|X} > 0$ implies M_{Sub} is invertible. As suggested by Proposition 1, we define $M_{\mathrm{Sub}}^{\dagger} := M_{\mathrm{Sub}}^{-1}$. Then,

$$M_{\mathrm{Sub}}^{\dagger}M_{\mathrm{Sub}} = \begin{pmatrix} \frac{\mathsf{P}_{Y=1|X}}{\mathsf{P}_{Y\ \ Y_{\mathrm{s}}|X}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\mathsf{P}_{Y=K|X}}{\mathsf{P}_{Y\ \ Y_{\mathrm{s}}|X}} \end{pmatrix} \begin{pmatrix} \frac{\mathsf{P}_{Y\ \ Y_{\mathrm{s}}|X}}{\mathsf{P}_{Y=1|X}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\mathsf{P}_{Y\ \ Y_{\mathrm{s}}|X}}{\mathsf{P}_{Y=K|X}} \end{pmatrix} = I$$

implies $M_{\mathrm{Sub}}^{\dagger}\bar{P} = M_{\mathrm{Sub}}^{-1}M_{\mathrm{Sub}}P = P$ that proves the lemma. □

With $M_{\text{Sub}}^{\dagger}$ in hand, the next theorem defines the corrected losses $\bar{L}$ and achieves the risk rewrite (31).

Theorem 46. *For Sub-Conf learning, the classification risk can be written as*

$$R(g) = \pi_{Y_{\mathrm{s}}} \mathsf{E}_{X|Y \quad Y_{\mathrm{s}}} \left[\sum_{i=1}^{K} \frac{r_i(X)}{r_{Y_{\mathrm{s}}}(X)} \ell_i \right].$$

Proof. Given Lemma 45, we can define $\bar{L} \quad := L \quad M_{\text{Sub}}^{\dagger}$ so that

$$\bar{L}_i \quad = \left(L \quad M_{\text{Sub}}^{\dagger} \right)_i = \frac{P_{Y=i|X}}{P_{Y \quad Y_{\mathrm{s}}|X}} \ell_i$$

for each $i \quad [K]$ and

$$\bar{L} \quad \bar{P} = L \quad M_{\text{Sub}}^{\dagger} \bar{P} = L \quad P.$$

Therefore, we can apply (39) to obtain (31) as follows.

$$\begin{aligned} R(g) \quad &= \quad \int_{x \quad X} L \quad P \mathrm{d}x = \int_{x \quad X} \bar{L} \quad \bar{P} \mathrm{d}x = \int_{x \quad X} \sum_{i=1}^{K} \frac{P_{Y=i|X}}{P_{Y \quad Y_{\mathrm{s}}|X}} \ell_i \cdot P_{Y \quad Y_{\mathrm{s}}} P_{X|Y \quad Y_{\mathrm{s}}} \mathrm{d}x \\ &= \quad P_{Y \quad Y_{\mathrm{s}}} \mathsf{E}_{X|Y \quad Y_{\mathrm{s}}} \left[\sum_{i=1}^{K} \frac{P_{Y=i|X}}{P_{Y \quad Y_{\mathrm{s}}|X}} \ell_i \right] \\ &= \quad \pi_{Y_{\mathrm{s}}} \mathsf{E}_{X|Y \quad Y_{\mathrm{s}}} \left[\sum_{i=1}^{K} \frac{r_i(X)}{r_{Y_{\mathrm{s}}}(X)} \ell_i \right]. \end{aligned}$$

The last equality follows the notations in Section 2.2.14. □

Step 2: Recovering the previous result(s).
Notation matching gives

$$R(g) = \pi_{Y_{\mathrm{s}}} \mathsf{E}_{p(x|y \quad Y_{\mathrm{s}})} \left[\sum_{y=1}^{K} \frac{r^{y}(x)}{r^{Y_{\mathrm{s}}}(x)} \ell(g(x), y) \right],$$

recovering Theorem 6 of Cao et al. (2021a)[14].

5.3.2 Single-Class Confidence (SC-Conf) Learning

Step 1: Corrected Loss Design and Risk Rewrite.
The SC-Conf derivation resembles that in Section 5.3.1 since M_{SC} is a child of M_{Sub} on the reduction graph. Thus, following the notations in Section 4.3.2, assuming $P_{Y=y_{\mathrm{s}}|X} > 0$ for all possible outcomes of X, and replacing the set Y_{s} in $M_{\text{Sub}}^{\dagger}$ (126) with a singleton y_{s}, we have

$$M_{\text{SC}}^{\dagger} := \begin{pmatrix} \frac{P_{Y=1|X}}{P_{Y=y_{\mathrm{s}}|X}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{P_{Y=K|X}}{P_{Y=y_{\mathrm{s}}|X}} \end{pmatrix}$$

satisfying $M_{\text{SC}}^{\dagger} \bar{P} = P$. We also obtain $\bar{L} \quad = L \quad M_{\text{SC}}^{\dagger}$ and $\bar{L} \quad \bar{P} = L \quad P$ by inheriting the proof of Lemma 45. Then, a modification to Theorem 46 by replacing $\bar{L}_i \quad = \frac{P_{Y=i|X}}{P_{Y \quad Y_{\mathrm{s}}|X}} \ell_i$ with

$$\bar{L}_i \quad = \left(L \quad M_{\text{SC}}^{\dagger} \right)_i = \frac{P_{Y=i|X}}{P_{Y=y_{\mathrm{s}}|X}} \ell_i = \frac{r_i(X)}{r_{y_{\mathrm{s}}}(X)} \ell_i$$

proves the risk rewrite (29) for SC-Conf learning:

[14]The matching is as follows: $P_{X|Y \quad Y_{\mathrm{s}}}$ is $p(x|y \quad Y_{\mathrm{s}})$, $r_i(X)$ is $r^{i}(X)$, $r_{Y_{\mathrm{s}}}(X)$ is $r^{Y_{\mathrm{s}}}(X)$, and ℓ_i is $\ell(g(X), i)$.

Corollary 47. *For SC-Conf learning, the classification risk can be written as*

$$R(g) = \pi_{y_\mathrm{s}} \mathsf{E}_{X|Y=y_\mathrm{s}} \left[\sum_{i=1}^{K} \frac{r_i(X)}{r_{y_\mathrm{s}}(X)} \ell_i \right].$$

Step 2: Recovering the previous result(s).
By matching notations, we obtain

$$R(g) = \pi_{y_\mathrm{s}} \mathsf{E}_{p(x|y_\mathrm{s})} \left[\sum_{y=1}^{K} \frac{r^y(x)}{r^{y_\mathrm{s}}(x)} \ell(g(x), y) \right],$$

recovering Theorem 1 of Cao et al. (2021a)[15].

5.3.3 Positive-confidence (Pconf) Learning

Step 1: Corrected Loss Design and Risk Rewrite.
Let us follow the notations in Section 4.3.3. Recall that M_{Pconf} is a child of M_{SC} on the reduction graph with $K = 2$ and $y_\mathrm{S} = \mathrm{p}$. Thus, assuming $\mathsf{P}_{Y=\mathrm{p}|X} > 0$ for all possible outcomes of X and replacing K and y_s in Section 5.3.2 accordingly, we obtain the decontamination matrix

$$M_{\mathrm{Pconf}}^{\dagger} := \begin{pmatrix} \frac{\mathsf{P}_{Y=\mathrm{p}|X}}{\mathsf{P}_{Y=\mathrm{p}|X}} & 0 \\ 0 & \frac{\mathsf{P}_{Y=\mathrm{n}|X}}{\mathsf{P}_{Y=\mathrm{p}|X}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1-r(X)}{r(X)} \end{pmatrix}$$

and the rewrite (7) reviewed in Section 2.2.2.

Corollary 48. *For Pconf learning, the classification risk can be written as*

$$R(g) = \pi_\mathrm{p} \mathsf{E}_\mathrm{P} \left[\ell_\mathrm{p} + \frac{1 - r(X)}{r(X)} \ell_\mathrm{n} \right].$$

Step 2: Recovering the previous result(s).
By matching notations, we obtain

$$R(g) = \pi_+ \mathsf{E}_+ \left[\ell(g(x)) + \frac{1 - r(x)}{r(x)} \ell(-g(x)) \right],$$

recovering Theorem 1 of Ishida et al. (2018)[16].

5.3.4 Soft-Label Learning

Step 1: Corrected Loss Design and Risk Rewrite.
We follow the notations in Section 4.3.4. As discussed in Section 4.3.4, M_{Soft} is a special case of M_{Sub} when $\mathsf{Y}_\mathrm{s} := [K]$. Thus, reducing $M_{\mathrm{Sub}}^{\dagger}$ (126) by assigning $\mathsf{P}_{Y\ \mathsf{Y}_\mathrm{s}|X} = \mathsf{P}_{Y\ [K]|X} = 1$, we obtain the the decontamination matrix

$$M_{\mathrm{Soft}}^{\dagger} := \begin{pmatrix} \mathsf{P}_{Y=1|X} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathsf{P}_{Y=K|X} \end{pmatrix}.$$

Then, we follow the same argument in Theorem 46 to achieve the rewrite (34) by the next corollary.

Corollary 49. *For soft-label learning, the classification risk can be written as*

$$R(g) = \mathsf{E}_X \left[\sum_{i=1}^{K} \mathsf{P}_{Y=i|X} \ell_i \right] = \mathsf{E}_X \left[\sum_{i=1}^{K} r_i(X) \ell_i \right].$$

[15]The matching is as follows: $\mathsf{P}_{X|Y=y_\mathrm{s}}$ is $p(x|y_\mathrm{s})$, $r_i(X)$ is $r^i(X)$, $r_{y_\mathrm{s}}(X)$ is $r^{y_\mathrm{s}}(X)$, and ℓ_i is $\ell(g(X), i)$.

[16]The matching is as follows: π_p is π_+, $\mathsf{P}_{X|Y=\mathrm{p}}$ is $p(x|y=+1)$, $\mathsf{P}_{Y=\mathrm{p}|X}$ is $r(x)$, $\mathsf{P}_{Y=\mathrm{n}|X}$ is $1 - r(x)$, ℓ_p is $\ell(g(x))$, and ℓ_n is $\ell(-g(x))$.

Step 2: Recovering the previous result(s).
Ishida et al. (2023) did not focus on the classification risk rewrite problem. We can modify Corollary 49 to provide a risk rewrite for binary soft-label learning mentioned by Ishida et al. (2023). Taking $K = 2$, we have

$$R(g) = \mathsf{E}_X \left[P_{Y=\mathrm{p}|X} \ell_\mathrm{p} + P_{Y=\mathrm{n}|X} \ell_\mathrm{n} \right] = \mathsf{E}_X \left[r(X) \ell_\mathrm{p} + (1 - r(X)) \ell_\mathrm{n} \right].$$

6 New Risk Rewrites

The proposed framework can be applied to derive risk rewrites for new scenarios. In this section, we show how the proposed framework accommodates a different performance metric and adapts to a noisy environment.

6.1 The Balanced Error Rate

In the previous sections, we chose the most common metric, the classification risk, to elaborate our framework. Here, we demonstrate that the framework adapts to another common but different performance metric. The balanced error rate (BER) is defined as

$$R_\mathrm{BER}(g) := \frac{\mathsf{E}_{X \sim P_{X|Y=\mathrm{p}}} \left[\ell_\mathrm{p}(g(X)) \right] + \mathsf{E}_{X \sim P_{X|Y=\mathrm{n}}} \left[\ell_\mathrm{n}(g(X)) \right]}{2} \tag{127}$$

(Scott & Zhang, 2020). Next, we show how to apply the framework to obtain the risk rewrite of UU learning under BER.

Recall that in Section 5.1.1, for classification risk, $M_\mathrm{trsf} = \begin{pmatrix} 1/\pi_\mathrm{p} & 0 \\ 0 & 1/\pi_\mathrm{n} \end{pmatrix}$ was chosen to link the base distributions $B = \begin{pmatrix} P_{X|Y=\mathrm{p}} \\ P_{X|Y=\mathrm{n}} \end{pmatrix}$ and the risk-defining distributions $P = \begin{pmatrix} P_{Y=\mathrm{p},X} \\ P_{Y=\mathrm{n},X} \end{pmatrix}$:

$$\begin{pmatrix} P_{X|Y=\mathrm{p}} \\ P_{X|Y=\mathrm{n}} \end{pmatrix} = \begin{pmatrix} 1/\pi_\mathrm{p} & 0 \\ 0 & 1/\pi_\mathrm{n} \end{pmatrix} \begin{pmatrix} P_{Y=\mathrm{p},X} \\ P_{Y=\mathrm{n},X} \end{pmatrix}.$$

By definition (127), the risk-defining distributions become $\begin{pmatrix} P_{X|Y=\mathrm{p}}/2 \\ P_{X|Y=\mathrm{n}}/2 \end{pmatrix}$. Since the base distributions B remain unchanged, we must redefine $M_\mathrm{trsf} := \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, so that

$$\begin{pmatrix} P_{X|Y=\mathrm{p}} \\ P_{X|Y=\mathrm{n}} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} P_{X|Y=\mathrm{p}}/2 \\ P_{X|Y=\mathrm{n}}/2 \end{pmatrix}$$

satisfies $B = M_\mathrm{trsf} P$ for the BER setting.

With the newly defined $M_\mathrm{trsf} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ in hand, we can follow Step 1 in Section 5.1.1 to obtain the data-generating process

$$\begin{pmatrix} P_{\mathrm{U}_1} \\ P_{\mathrm{U}_2} \end{pmatrix} = M_\mathrm{UU} \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} P_{X|Y=\mathrm{p}}/2 \\ P_{X|Y=\mathrm{n}}/2 \end{pmatrix}.$$

Then, by assigning $M^\dagger_{\mathrm{UU-BER}} := \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} M_\mathrm{UU}^{-1}$ and following the similar derivation in the proof of Theorem 21, we achieve the risk rewrite for UU learning under BER:

$$R_\mathrm{BER}(g) = \mathsf{E}_{\mathrm{U}_1} \left[\bar{\ell}_{\mathrm{U}_1} \right] + \mathsf{E}_{\mathrm{U}_2} \left[\bar{\ell}_{\mathrm{U}_2} \right],$$

where

$$\begin{aligned}\bar{\ell}_{\mathrm{U}_1} &= \frac{1-\gamma_2}{2(1-\gamma_1-\gamma_2)}\ell_{\mathrm{p}} + \frac{-\gamma_2}{2(1-\gamma_1-\gamma_2)}\ell_{\mathrm{n}},\\ \bar{\ell}_{\mathrm{U}_2} &= \frac{-\gamma_1}{2(1-\gamma_1-\gamma_2)}\ell_{\mathrm{p}} + \frac{1-\gamma_1}{2(1-\gamma_1-\gamma_2)}\ell_{\mathrm{n}}.\end{aligned}$$

6.2 Learning with Label Noise

In this section, we show that the framework can easily handle the risk rewrite under label noise. We will take PU learning as an example. We use $\boldsymbol{P}$ to denote the clean distributions and $\boldsymbol{Q}$ for the noisy ones. The scenario is formulated as follows. According to Lemma 4, the data-generating process of the noisy PU learning is of the form

$$\begin{pmatrix}\boldsymbol{Q}_{\mathrm{P}}\\ \boldsymbol{Q}_{\mathrm{U}}\end{pmatrix} = \begin{pmatrix}1 & 0\\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}}\end{pmatrix}\begin{pmatrix}\boldsymbol{Q}_{X|Y=\mathrm{p}}\\ \boldsymbol{Q}_{X|Y=\mathrm{n}}\end{pmatrix},$$

where $\pi_{\mathrm{p}} = \boldsymbol{Q}_{Y=\mathrm{p}}$ and $\pi_{\mathrm{n}} = \boldsymbol{Q}_{Y=\mathrm{n}}$. We choose the MCD setting (Section 2.2.17) to formulate the label noise:

$$\begin{pmatrix}\boldsymbol{Q}_{X|Y=\mathrm{p}}\\ \boldsymbol{Q}_{X|Y=\mathrm{n}}\end{pmatrix} = \begin{pmatrix}1-\alpha_{\mathrm{p}} & \alpha_{\mathrm{p}}\\ \alpha_{\mathrm{n}} & 1-\alpha_{\mathrm{n}}\end{pmatrix}\begin{pmatrix}\boldsymbol{P}_{X|Y=\mathrm{p}}\\ \boldsymbol{P}_{X|Y=\mathrm{n}}\end{pmatrix},$$

where $\alpha_{\mathrm{p}} \geq 0$ and $\alpha_{\mathrm{n}} \geq 0$ are parameters describing the degree of noise perturbation. As in the previous section, we also choose BER as the performance metric.

By cascading matrices, the framework effortlessly links the data-generating distributions and the risk-defining distributions:

$$\begin{pmatrix}\boldsymbol{Q}_{\mathrm{P}}\\ \boldsymbol{Q}_{\mathrm{U}}\end{pmatrix} = \begin{pmatrix}1 & 0\\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}}\end{pmatrix}\begin{pmatrix}1-\alpha_{\mathrm{p}} & \alpha_{\mathrm{p}}\\ \alpha_{\mathrm{n}} & 1-\alpha_{\mathrm{n}}\end{pmatrix}\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}\begin{pmatrix}\boldsymbol{P}_{X|Y=\mathrm{p}}/2\\ \boldsymbol{P}_{X|Y=\mathrm{n}}/2\end{pmatrix}.$$

Knowing the contamination process, we apply the inversion method (Proposition 1) to construct the decontamination matrix

$$\begin{aligned}M^{\dagger}_{\mathrm{PU\text{-}BER}} &:= \begin{pmatrix}1/2 & 0\\ 0 & 1/2\end{pmatrix}\frac{1}{1-\alpha_{\mathrm{p}}-\alpha_{\mathrm{n}}}\begin{pmatrix}1-\alpha_{\mathrm{n}} & -\alpha_{\mathrm{p}}\\ -\alpha_{\mathrm{n}} & 1-\alpha_{\mathrm{p}}\end{pmatrix}\begin{pmatrix}1 & 0\\ -\pi_{\mathrm{p}}/\pi_{\mathrm{n}} & 1/\pi_{\mathrm{n}}\end{pmatrix}\\ &= \begin{pmatrix}\frac{(1-\alpha_{\mathrm{n}})\pi_{\mathrm{n}}+\alpha_{\mathrm{p}}\pi_{\mathrm{p}}}{2(1-\alpha_{\mathrm{p}}-\alpha_{\mathrm{n}})\pi_{\mathrm{n}}} & \frac{-\alpha_{\mathrm{p}}}{2(1-\alpha_{\mathrm{p}}-\alpha_{\mathrm{n}})\pi_{\mathrm{n}}}\\ \frac{-\alpha_{\mathrm{n}}\pi_{\mathrm{n}}-(1-\alpha_{\mathrm{p}})\pi_{\mathrm{p}}}{2(1-\alpha_{\mathrm{p}}-\alpha_{\mathrm{n}})\pi_{\mathrm{n}}} & \frac{1-\alpha_{\mathrm{p}}}{2(1-\alpha_{\mathrm{p}}-\alpha_{\mathrm{n}})\pi_{\mathrm{n}}}\end{pmatrix} := \begin{pmatrix}c_1 & c_2\\ c_3 & c_4\end{pmatrix}.\end{aligned}$$

Then, applying equations (38) and (39), we can rewrite the BER for PU learning as

$$R_{\mathrm{BER}}(g) = \mathsf{E}_{X\ \boldsymbol{Q}_{\mathrm{P}}}\left[\bar{\ell}_{\mathrm{P}}\right] + \mathsf{E}_{X\ \boldsymbol{Q}_{\mathrm{U}}}\left[\bar{\ell}_{\mathrm{U}}\right],$$

where

$$\begin{aligned}\bar{\ell}_{\mathrm{P}} &= c_1\ell_{\mathrm{p}} + c_3\ell_{\mathrm{n}},\\ \bar{\ell}_{\mathrm{U}} &= c_2\ell_{\mathrm{p}} + c_4\ell_{\mathrm{n}}.\end{aligned}$$

In the two subsections above, we have shown that the proposed framework can address risk rewrite under a different performance metric and a complex system. In the outlook part of the next section, we will further discuss the potential of the framework.

7 Conclusion and Outlook

We set out with the questions wishing to determine if there is a common way to interpret the formation of weak supervision and search for a generic treatment to solve WSL, to understand the essence of WSL. In response, we proposed a framework that unifies the formulations and analyses of a set of WSL scenarios to provide a common ground to connect, compare, and understand various weakly-supervised signals. The formulation component of the proposed framework, viewing WSL from a contamination perspective, associates a WSL data-generating process with a base distribution vector multiplied by a contamination matrix. By instantiating the contamination matrices of WSLs, we revealed a comprehensive reduction graph, Figure 1, connecting existing WSLs. Each vertex contains a contamination matrix and the section index of the WSL scenario which the matrix characterizes. Each edge represents the reduction relation of two WSLs. We can see three major branches from the abstract M_{corr}, corresponding to Tables 7, 8, and 9 we discussed in Section 4. The analysis component of the proposed framework, tackling the problem from a decontamination viewpoint, working with the technical building blocks Theorems 1 and 2 constitute a generic treatment to solve the risk rewrite problem. Section 5 discussed in depth how the analysis component conducts risk rewrite and recovers existing results for WSLs.

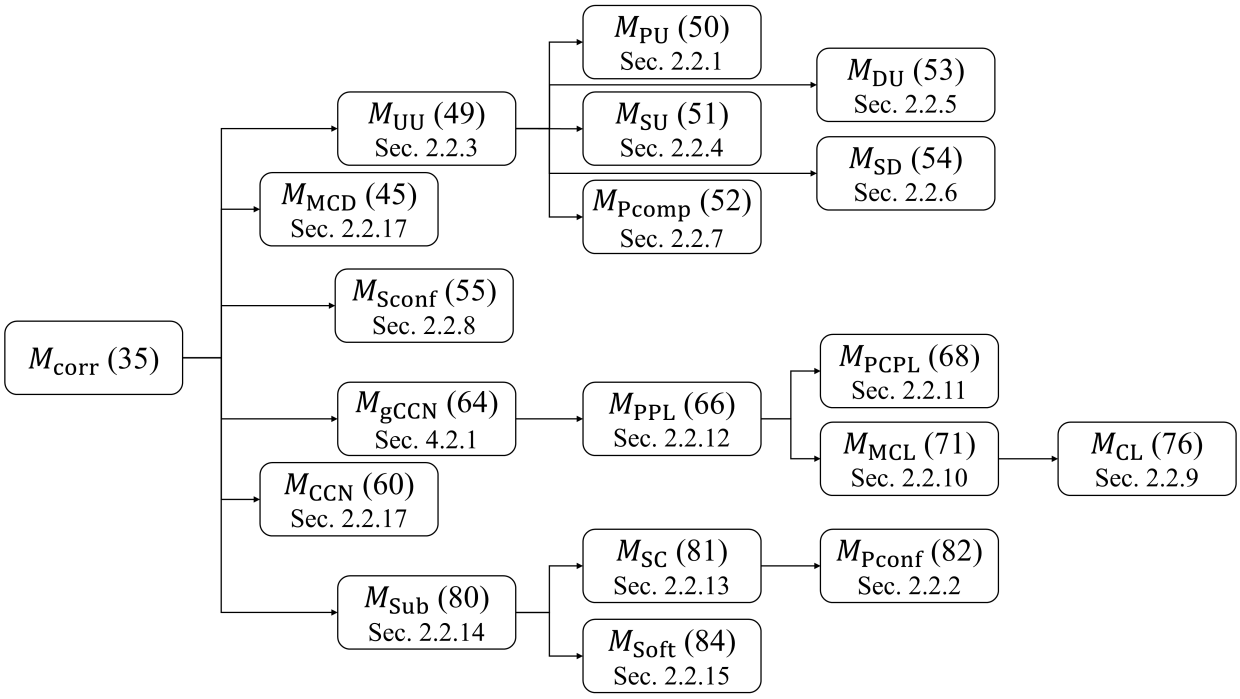

Figure 1: Depicting the reduction map from Tables 7, 8, and 9.

The application of the proposed framework results in a set of theorems. We summarize them in Table 10. The Formulation column consists of the results of the formulation component (35). The Decontamination and the Corrected losses columns correspond to the results of the analysis component ((37), (38), and (39)). The Recovery column justifies the framework by recovering results from the literature. Crucial results are marked red. Since the analyses of different scenarios are subsumed under a single framework, we now have a basis for transferring a technique developed for one scenario to another. In addition, these alternative proofs provide different ways of dissecting the risk, which in turn could aid in the development of a training objective by examining multiple risk decomposition approaches.

Table 10: Theorem Structure.

Model	Formulation (Find M s.t. $\bar{P} = MB$.)	Decontamination (Find $M^\dagger$ s.t. $P = M^\dagger \bar{P}$.)	Corrected losses (Rewrite via $\bar{L} = L \quad M^\dagger$ and $\bar{P}$.)	Recovery
Abstract model	(35)	(37) Proposition 1 and Proposition 2	(38) and (39)	
MCD				
UU	Lemma 3	Corollary 20	Theorem 21	(Notation swap.)
PU	Lemma 4	(Immediate reduction.)	Corollary 22	(Notation swap.)
SU	Lemma 5	(Immediate reduction.)	Corollary 23	Lemmas 24 and 25
Pcomp	Lemma 6	(Immediate reduction.)	Corollary 26	(Notation swap.)
DU	Lemma 7	(Immediate reduction.)	Corollary 27	Lemmas 28 and 29
SD	Lemma 8	(Immediate reduction.)	Corollary 30	Lemma 31
Sconf	Lemma 9	Lemmas 32 and 33	Theorem 34	(Notation swap.)
CCN				
gCCN	Lemma 10	(106) and Proposition 2	Theorem 35	(Notation swap.)
PPL	Lemma 11	Lemma 36	Corollary 37	(Notation swap.)
PCPL	Lemma 12	Corollary 38	Corollary 38	Lemma 39
MCL	Lemma 13	Corollary 40	Corollary 40	Theorem 42, Lemmas 41 and 43
CL	Lemma 15	(Immediate reduction.)	Corollary 44	(Notation swap.)
Sub-Conf	Lemma 16	Lemma 45	Theorem 46	(Notation swap.)
SC-Conf	Lemma 17	(Immediate reduction.)	Corollary 47	(Notation swap.)
Pconf	Lemma 18	(Immediate reduction.)	Corollary 48	(Notation swap.)
Soft	Lemma 19	(Immediate reduction.)	Corollary 49	(N/A.)

The proposed framework is abstract and flexible; hence, we would like to discuss its potential from the following aspects. Firstly, the performance measure focused on in this paper is the classification risk. With proper choices of P and L, our framework can be extended to other performance metrics, such as the balanced error rate, one-versus-rest risk, and cost-sensitive measures (Rifkin & Klautau, 2004; Zhang, 2004; Brodersen et al., 2010; du Plessis et al., 2014; Menon et al., 2015; Blanchard et al., 2016; Natarajan et al., 2017; Scott & Zhang, 2020). We have demonstrated the applicability of the proposed framework for the balanced error rate in Section 6. Secondly, we can explore the formulation capability by exploiting the power of matrix operations. Cascading matrices allow us to formulate complex scenarios, such as data containing preference relations collected in a noisy environment. Matrix addition allows us to categorize different contamination mechanisms into cases to capture the structural properties of a problem. A complicated scenario could undergo a sophisticated formulation procedure, but once we have the resulting contamination matrix, the problem boils down to calculating the corresponding decontamination matrix. Thirdly, the MCD scenarios discussed in this paper (Sections 4.1 and 5.1) belong to binary classification. A way of extending an MCD formulation to multiclass classification is to extend M_{MCD} (45) from a 2×2 matrix to a $K \times K$ one, in which $K^2 - K$ mixture rates are used to characterize the extended M_{gMCD}: the (i, j) entry is $\gamma_{i,j}$ if $i = j$ and is $1 - \sum_{j=i} \gamma_{i,j}$ for the i-th entry on the diagonal. Fourthly, the label-flipping probabilities $\boldsymbol{P}_{\bar{Y}|Y}$ in Natarajan et al. (2017) and Feng et al. (2020b) assume that the contaminated label $\bar{Y}$ is independent of X condition on the ture label Y. The formulation matrices, M_{CCN} (60) and M_{gCCN} (64), in contrast, take X into consideration. This formulation enables us to tackle the instance-dependent problem (Berthon et al., 2021) and the effect of sampling strategies such as SAR and SCAR (Elkan & Noto, 2008; Coudray et al., 2023) in the future. Fifthly, we hope that the analysis technique developed in Section 5.2, which combines marginal chain and properness, opens up a new possibility to search for invertibility-free methods for the risk rewrite problem. We also project its potential in research regarding the broader sense of contamination

and decontamination. Sixthly, the properness of Wu et al. (2023) provides an efficient technique to compute $\mathsf{P}_{Y|S,X}$ needed in M_{gCCN} (106). It would be intriguing to know if there are any other alternatives. Finally but not least, the proposed framework operating under matrix multiplication belongs to a broader question of under what circumstances does a function $f^\dagger$ exist with $P = f^\dagger(\bar{P})$ if $\bar{P} = f(P)$.

Acknowledgments

The authors were supported by the Institute for AI and Beyond, UTokyo. The first author would like to thank Professor Takashi Ishida (UTokyo) for valuable insights and discussions in extending the coverage of the framework, and the colleagues Xin-Qiang Cai, Masahiro Negishi, Wei Wang, and Yivan Zhang (in alphabetical order) for comments in improving the manuscript.

References

Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pp. 461–470, Stockholmsmässan, Stockholm, Sweden, 2018.

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event*, pp. 825–836, 2021.

Gilles Blanchard and Clayton Scott. Decontamination of mutually contaminated models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014*, pp. 1–9, Reykjavik, Iceland, 2014.

Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10(2):2780–2824, 2016.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition, ICPR 2010*, pp. 3121–3124, Istanbul, Turkey, 2010.

Yuzhou Cao, Lei Feng, Senlin Shu, Yitian Xu, Bo An, Gang Niu, and Masashi Sugiyama. Multi-class classification from single-class data with confidences. *CoRR*, abs/2106.08864, 2021a. URL `https://arxiv.org/abs/2106.08864`.

Yuzhou Cao, Lei Feng, Yitian Xu, Bo An, Gang Niu, and Masashi Sugiyama. Learning from similarity-confidence data. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event*, pp. 1272–1282, 2021b.

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA*, pp. 961–970, 2019.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *Proceedings of 38th International Conference on Machine Learning, ICML 2021, Virtual Event*, pp. 1507–1517, 2021.

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event*, pp. 1929–1938, 2020.

Jesús Cid-Sueiro. Proper losses for learning from partial labels. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, pp. 1574–1582, Lake Tahoe, Nevada, USA, 2012.

Olivier Coudray, Christine Keribin, Pascal Massart, and Patrick Pamphile. Risk bounds for positive-unlabeled learning under the selected at random assumption. *Journal of Machine Learning Research*, 24:107:1–107:31, 2023. URL https://jmlr.org/papers/v24/22-067.html.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011. URL https://dl.acm.org/doi/10.5555/1953048.2021049.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Clustering unclustered data: Unsupervised binary labeling of two datasets having different class balances. In *Proceedings of Conference on Technologies and Applications of Artificial Intelligence, TAAI 2013*, pp. 1–6, Taipei, Taiwan, 2013.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pp. 703–711, Montréal, Québec, Canada, 2014.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pp. 1386–1394, Lille, France, 2015.

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, Las Vegas, Nevada, USA, 2008.

Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event*, pp. 3072–3081, 2020a.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual Event*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/7bd28f15a49d5e5848d6ec70e584e625-Abstract.html.

Lei Feng, Senlin Shu, Nan Lu, Bo Han, Miao Xu, Gang Niu, Bo An, and Masashi Sugiyama. Pointwise binary classification with pairwise confidence comparisons. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event*, pp. 3252–3262, 2021.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, Conference Track Proceedings*, Toulon, France, 2017. URL https://openreview.net/forum?id=H12GRgcxg.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pp. 8536–8546, Montréal, Québec, Canada, 2018.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 5639–5649, Long Beach, California, USA, 2017.

Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification from positive-confidence data. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pp. 5921–5932, Montréal, Québec, Canada, 2018.

Takashi Ishida, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 2971–2980, Long Beach, California, USA, 2019.

Takashi Ishida, Ikko Yamane, Nontawat Charoenphakdee, Gang Niu, and Masashi Sugiyama. Is the performance of my deep network too good to be true? A direct approach to estimating the bayes error in binary classification. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda*, 2023. URL https://openreview.net/forum?id=FZdJQgy05rz.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pp. 2309–2318, Stockholmsmässan, Stockholm, Sweden, 2018.

Yasuhiro Katsura and Masato Uchida. Bridging ordinary-label learning and complementary-label learning. In *Proceedings of The 12th Asian Conference on Machine Learning, ACML 2020*, pp. 161–176, Bangkok, Thailand, 2020.

Julian Katz-Samuels, Gilles Blanchard, and Clayton Scott. Decontamination of mutual contamination models. *Journal of Machine Learning Research*, 20:41:1–41:57, 2019. URL http://jmlr.org/papers/v20/17-576.html.

Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 1675–1685, Long Beach, California, USA, 2017.

Nan Lu, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA*, New Orleans, Louisiana, USA, 2019. URL https://openreview.net/forum?id=B1xWcj0qYm.

Nan Lu, Tianyi Zhang, Gang Niu, and Masashi Sugiyama. Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, Online Event*, pp. 1115–1125, 2020. URL http://proceedings.mlr.press/v108/lu20c.html.

Nan Lu, Shida Lei, Gang Niu, Issei Sato, and Masashi Sugiyama. Binary classification from multiple unlabeled datasets via surrogate set classification. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event*, pp. 7134–7144, 2021. URL http://proceedings.mlr.press/v139/lu21c.html.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event*, pp. 6500–6510, 2020.

Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah M. Erfani, Shu-Tao Xia, Sudanthi N. R. Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pp. 3361–3370, Stockholmsmässan, Stockholm, Sweden, 2018.

Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pp. 125–134, Lille, France, 2015.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pp. 1196–1204, Lake Tahoe, Nevada, USA, 2013.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18:155:1–155:33, 2017. URL http://jmlr.org/papers/v18/15-226.html.

Masahiro Negishi. Pairwise-constraint classification in weakly supervised machine learning: Risk-consistent approach and classifier-consistent approach. Senior thesis, 2023.

Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pp. 1199–1207, Barcelona, Spain, 2016.

Curtis G. Northcutt, Tailin Wu, and Isaac L. Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017*, Sydney, Australia, 2017. URL http://auai.org/uai2017/proceedings/papers/35.pdf.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 2233–2241, Honolulu, Hawaii, USA, 2017.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *3rd International Conference on Learning Representations, ICLR 2015, Workshop Track Proceedings*, San Diego, California, USA, 2015. URL http://arxiv.org/abs/1412.6596.

Ryan M. Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004. URL https://jmlr.org/papers/v5/rifkin04a.html.

Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Semi-supervised AUC optimization based on positive-unlabeled learning. *Machine Learning*, 107(4):767–794, 2018.

Emanuele Sansone, Francesco G. B. De Natale, and Zhi-Hua Zhou. Efficient training for positive unlabeled learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2584–2598, 2019.

Clayton Scott and Jianxin Zhang. Learning from multiple corrupted sources, with application to learning from label proportions. *CoRR*, abs/1910.04665, 2019. URL http://arxiv.org/abs/1910.04665.

Clayton Scott and Jianxin Zhang. Learning from label proportions: A mutual contamination framework. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual Event*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/fcde14913c766cf307c75059e0e89af5-Abstract.html.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *COLT 2013 - The 26th Annual Conference on Learning Theory*, pp. 489–511, Princeton University, New Jersey, USA, 2013.

Takuya Shimada, Han Bao, Issei Sato, and Masashi Sugiyama. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. *Neural Computation*, 33(5):1234–1268, 2021.

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, Tomoya Sakai, and Gang Niu. *Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach*. Adaptive Computation and Machine Learning series. MIT Press, 2022. ISBN 9780262047074.

Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. In *3rd International Conference on Learning Representations, ICLR 2015, Workshop Track Proceedings*, San Diego, California, USA, 2015. URL http://arxiv.org/abs/1406.2080.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 5552–5560, Salt Lake City, Utah, USA, 2018.

Ambuj Tewari and Peter L. Bartlett. Chapter 14 - learning theory. In *Academic Press Library in Signal Processing: Volume 1*, pp. 775–816. Elsevier, 2014. doi: https://doi.org/10.1016/B978-0-12-396502-8.00014-0.

Brendan van Rooyen and Robert C. Williamson. Learning in the presence of corruption. *CoRR*, abs/1504.00091, 2015. URL http://arxiv.org/abs/1504.00091.

Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18:228:1–228:50, 2017. URL http://jmlr.org/papers/v18/16-315.html.

Qian-Wei Wang, Yu-Feng Li, and Zhi-Hua Zhou. Partial label learning with unlabeled data. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pp. 3755–3761, Macao, China, 2019.

Zhenguo Wu, Jiaqi Lv, and Masashi Sugiyama. Learning with proper partial labels. *Neural Computation*, 35(1):58–81, 2023. URL https://doi.org/10.1162/neco_a_01554.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 7164–7173, Long Beach, California, USA, 2019.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, Proceedings, Part I*, pp. 69–85, 2018.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004. URL https://jmlr.org/papers/volume5/zhang04b/zhang04b.pdf.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

A Use Cases for This Paper

There are several ways to use this paper in the study of WSLs. We provide some common use cases as follows:

Use Case 1: If you want a quick overview of WSL. Reading Section 2 would provide a quick catch-up on the formulations and results of various WSL scenarios.

Use Case 2: If you want to get the high-level idea of this paper. One could start with Section 3, which provides the abstract form of the proposed framework. Then, Sections 4.1.1 and 5.1.1 give the practical application of the framework to UU learning. If you are more familiar with PU learning, then you are welcome to proceed to Sections 4.1.2 and 5.1.1 to see how our framework interprets PU learning. In addition, these reading steps will reveal the connection between UU and PU.

Use Case 3: If you want to know how to apply the framework to a specific scenario. One could read Sections 4.a.b and 5.a.b at the same time, and come back to the rest of Sections 4 and 5 only as needed. Note that "a.b" represents the index of a subsubsection.

Use Case 4: If you want to know the connections discovered and the analysis techniques developed in this paper. Section 4 provides detailed discussions of the formulations, and Figure 1 summarizes the relationship. Section 5 provides detailed explanations of how our framework is applied to rewrite the classification risk. The extensive analyses from Section 5 are summarized in Table 10.

B Notations

Table 11: Notations and Aliases.

Name of the notation	Expression	Aliases	Convention
Example	(y, x)		(x, y)
Binary classes	$\{\mathrm{p}, \mathrm{n}\}$		$\{+1, -1\}$
Multiple classes	$\{1, \cdots, K\}$	$[K]$	
Compound classes of $[K]$	$2^{[K]} \setminus \{\ , [K]\}$	$\mathcal{S}$	
A subset of classes	$\mathcal{Y}_s \quad [K]$		
Joint distribution	$\Pr(Y = y, X = x)$	$P_{Y=y,x}$, $P_{Y=y,X}$, or $P_{Y,X}$	$\Pr(x, y)$
Class prior	$\Pr(Y = y)$	π_y	
Marginal	$\Pr(X)$	P_X	
Class-conditional	$\Pr(X = x \mid Y = y)$	$P_{x\mid y}$, $P_{X\mid y}$, $P_{x\mid Y=y}$, or $P_{X\mid Y=y}$	
Class probability	$\Pr(Y = y \mid X = x)$	$P_{Y=y\mid x}$, $P_{Y=y\mid X}$, or $P_{Y\mid X}$	$\eta(x)$
Confidence	$\Pr(Y = y \mid X = x)$	$r_y(X)$, $r_y(x)$, or $r(X)$ if $y = \mathrm{p}$	$r^y(x)$ or $r(x)$
Sample size probability	$\Pr(\lvert S\rvert = d)$	$P_{\lvert S\rvert=d}$ or $q_{\lvert S\rvert}$	
Hypothesis and its space	$g \quad \mathcal{G}$		
Loss of g	$\ell_{Y=y}(g(x))$	ℓ_y, $\ell_y(X)$, or $\ell_Y(g(X))$	$\ell(g(X), Y)$
Classification risk	$\mathsf{E}_{Y,X}\left[\ell_Y(g(X))\right]$	$R(g)$	$\mathsf{E}_{X,Y}\left[\ell(g(X), Y)\right]$
The j-th entry of vector V	$(V)_j$	V_j	
Indicator function of E	$\mathsf{I}\left[E\right]$		
Complement of set s	$\mathcal{Y} \setminus s$	$\bar{s}$	
Identity matrix	I		
MCD parameters	γ_{p} and γ_{n}		
UU parameters	γ_1 and γ_2		$1 - \theta$ and θ
CCN parameters	$P_{\bar{Y}\mid Y,X}$	$P_{S\mid Y,X}$ or $P_{\bar{S}\mid Y,X}$	ρ_+ and ρ_-

C Omitted Proofs in Section 4

Omitted proofs in Section 4 are provided in this appendix. We first restate a claim in the main body of the paper, and then provide the corresponding proof.

C.1 Omitted Proofs in Section 4.1

Proof of Lemma 5

Lemma 5. *Let B (46) be the base distributions and*

$$\bar{P} := \begin{pmatrix} \boldsymbol{P}_{\tilde{\mathrm{S}}} \\ \boldsymbol{P}_{\mathrm{U}} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{SU}} := \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix}, \tag{51}$$

which satisfies $\bar{P} = M_{\mathrm{SU}}B$, characterizes the data-generating distributions $\bar{P}$.

Proof. The proof steps follow that of Lemma 4. By definitions,

$$M_{\mathrm{SU}}B = \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix}.$$

Since

$$\frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2}\boldsymbol{P}_{X|Y=\mathrm{p}} + \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2}\boldsymbol{P}_{X|Y=\mathrm{n}} = \frac{\pi_{\mathrm{p}}^2\boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|Y=\mathrm{p}}}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} = \boldsymbol{P}_{\tilde{\mathrm{S}}}$$

and

$$\pi_{\mathrm{p}}\boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}\boldsymbol{P}_{X|Y=\mathrm{n}} = \boldsymbol{P}_X = \boldsymbol{P}_{\mathrm{U}},$$

the first entry of the resulting vector equals $\boldsymbol{P}_{\tilde{\mathrm{S}}}$ and the second entry equals $\boldsymbol{P}_{\mathrm{U}}$, we achieve $\bar{P} = M_{\mathrm{SU}}B$. □

Proof of Lemma 6

Lemma 6. *Let B (46) be the base distributions and*

$$\bar{P} := \begin{pmatrix} \boldsymbol{P}_{\mathrm{Sup}} \\ \boldsymbol{P}_{\mathrm{Inf}} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{Pcomp}} := \begin{pmatrix} \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2} \\ \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}} & \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}} \end{pmatrix}, \tag{52}$$

which satisfies $\bar{P} = M_{\mathrm{Pcomp}}B$, characterizes the data-generating distributions $\bar{P}$.

Proof. By definitions,

$$M_{\mathrm{Pcomp}}B = \begin{pmatrix} \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2} \\ \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}} & \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix}.$$

Since

$$\frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2}\boldsymbol{P}_{X|Y=\mathrm{p}} + \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2}\boldsymbol{P}_{X|Y=\mathrm{n}} = \frac{\pi_{\mathrm{p}}\boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}^2\boldsymbol{P}_{X|Y=\mathrm{n}}}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2} = \boldsymbol{P}_{\mathrm{Sup}}$$

and

$$\frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}}\boldsymbol{P}_{X|Y=\mathrm{p}} + \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}}\boldsymbol{P}_{X|Y=\mathrm{n}} = \frac{\pi_{\mathrm{p}}^2\boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}\boldsymbol{P}_{X|Y=\mathrm{n}}}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}} = \boldsymbol{P}_{\mathrm{Inf}},$$

the first entry of the resulting vector equals $\boldsymbol{P}_{\mathrm{Sup}}$ and the second entry of the resulting vector equals $\boldsymbol{P}_{\mathrm{Inf}}$, which establishes $M_{\mathrm{Pcomp}}B = \bar{P}$. □

Proof of Lemma 7

Lemma 7. *Let B (46) be the base distributions and*

$$\bar{P} = \begin{pmatrix} \boldsymbol{P}_{\tilde{\mathrm{D}}} \\ \boldsymbol{P}_{\mathrm{U}} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{DU}} = \begin{pmatrix} 1/2 & 1/2 \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix}, \tag{53}$$

which satisfies $\bar{P} = M_{\mathrm{DU}}B$, characterizes the data-generating distributions $\bar{P}$.

Proof. Similar to the proofs of Lemmas 5 and 6, we begin with

$$M_{\mathrm{DU}}B = \begin{pmatrix} 1/2 & 1/2 \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix}.$$

Since $\left(\boldsymbol{P}_{X|Y=\mathrm{p}} + \boldsymbol{P}_{X|Y=\mathrm{n}}\right)/2 = \boldsymbol{P}_{\tilde{\mathrm{D}}}$ and $\pi_{\mathrm{p}}\boldsymbol{P}_{X|Y=\mathrm{p}} + \pi_{\mathrm{n}}\boldsymbol{P}_{X|Y=\mathrm{n}} = \boldsymbol{P}_X = \boldsymbol{P}_{\mathrm{U}}$, we have $M_{\mathrm{DU}}B = \bar{P}$. □

Proof of Lemma 8

Lemma 8. *Let B (46) be the base distributions and*

$$\bar{P} = \begin{pmatrix} \boldsymbol{P}_{\tilde{\mathrm{S}}} \\ \boldsymbol{P}_{\tilde{\mathrm{D}}} \end{pmatrix}.$$

Then, the contamination matrix

$$M_{\mathrm{SD}} = \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} \\ 1/2 & 1/2 \end{pmatrix} \tag{54}$$

which satisfies $\bar{P} = M_{\mathrm{SD}}B$, characterizes the data-generating distributions $\bar{P}$.

Proof. First, we begin with

$$M_{\mathrm{SD}}B = \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2} \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{X|Y=\mathrm{p}} \\ \boldsymbol{P}_{X|Y=\mathrm{n}} \end{pmatrix}.$$

Then, we have the lemma by reusing the calculations in the proofs of Lemmas 5 and 7. □

C.2 Omitted Proofs in Section 4.2

Proof of Lemma 11

Lemma 11. *Let the elements in S be* $\{s_1, s_2, \dots s_{|S|}\}$. *For each* $j \quad [|S|]$, *let the* j*-th entry of* $\bar{P}$ *be*

$$\bar{P}_j = P_{S=s_j,X} := C(S = s_j, X) \sum_{k \quad s_j} P_{Y=k,X},$$

which denotes the data-generating distribution of (s_j, X). *Assume the base distributions* B *and the contamination matrix* M_{PPL} *are given by (63) and (66), respectively. Then,* M_{PPL} *satisfies* $\bar{P} = M_{\mathrm{PPL}}B$ *and characterizes PPL learning (26).*

Proof. For each $j \quad [|S|]$,

$$\left(M_{\mathrm{PPL}}B\right)_j = \sum_{k=1}^{K} C(s_j, X) \mathsf{I}\left[Y = k \quad s_j\right] P_{Y=k,X} = C(s_j, X) \sum_{k \quad s_j} P_{Y=k,X} = \bar{P}_j.$$

Note that $C(s_j, X) \sum_{k \quad s_j} P_{Y=k,X}$ corresponds to (26) when the observed partial-label is s_j. □

Proof of Lemma 12

Lemma 12. *Let the elements in S be* $\{s_1, s_2, \dots s_{|S|}\}$. *For each* $j \quad [|S|]$, *let the* j*-th entry of* $\bar{P}$ *be*

$$\bar{P}_j = P_{S=s_j,X} := \frac{1}{2^{K-1} - 1} \sum_{k \quad s_j} P_{Y=k,X},$$

which denotes the data-generating distribution of (s_j, X). *Assume the base distributions* B *and the contamination matrix* M_{PCPL} *are given by (63) and (68), respectively. Then,* M_{PCPL} *satisfies* $\bar{P} = M_{\mathrm{PCPL}}B$ *and characterizes PCPL learning (24).*

Proof. For each $j \quad [|S|]$,

$$\left(M_{\mathrm{PCPL}}B\right)_j = \sum_{k=1}^{K} \frac{1}{2^{K-1} - 1} \mathsf{I}\left[Y = k \quad s_j\right] P_{Y=k,X} = \frac{1}{2^{K-1} - 1} \sum_{k \quad s_j} P_{Y=k,X} = \bar{P}_j.$$

Note that $\frac{1}{2^{K-1}-1} \sum_{k \quad s_j} P_{Y=k,X}$ corresponds to (24) when the observed partial-label is s_j. □

Proof of lemma 13

Lemma 13. *Suppose the base distributions* B, *the contamination matrix* M_{MCL}, *and the data-generating distributions* $\bar{P}$ *are given by (63), (71), and (70), respectively. Then,* M_{MCL} *satisfies* $\bar{P} = M_{\mathrm{MCL}}B$ *and characterizes MCL (22).*

Proof. For each $j \quad [N]$, we have

$$\begin{aligned}
\left(M_{\mathrm{MCL}}B\right)_j &= \sum_Y \frac{P_{|\bar{S}|=|\bar{s}_j|}}{\binom{K-1}{|\bar{s}_j|}} \mathsf{I}\left[Y \ / \ \bar{s}_j\right] P_{Y,X} \\
&= \sum_Y \frac{\sum_{d=1}^{K-1} \mathsf{I}\left[|\bar{s}_j| = d\right] P_{|\bar{s}_j|=d}}{\binom{K-1}{|\bar{s}_j|}} \mathsf{I}\left[Y \ / \ \bar{s}_j\right] P_{Y,X} \\
&= \sum_{d=1}^{K-1} \frac{P_{|\bar{s}_j|=d}}{\binom{K-1}{|\bar{s}_j|}} \sum_Y \mathsf{I}\left[Y \ / \ \bar{s}_j\right] P_{Y,X} \mathsf{I}\left[|\bar{s}_j| = d\right] \\
&= \sum_{d=1}^{K-1} \frac{P_{|\bar{s}_j|=d}}{\binom{K-1}{|\bar{s}_j|}} \sum_{Y/\bar{s}_j} P_{Y,X} \mathsf{I}\left[|\bar{s}_j| = d\right] \\
&= P_{\bar{S}=\bar{s}_j,X}.
\end{aligned}$$

□

D Omitted Proofs in Section 5

Omitted proofs in Section 5 are provided in this appendix. We first restate a claim in the main body of the paper, and then provide the corresponding proof.

D.1 Omitted Proofs in Section 5.1

Proof of Corollary 22

Corollary 22. *For PU learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{\mathrm{P}}\left[\bar{\ell}_{\mathrm{P}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right], \tag{90}$$

where

$$\begin{aligned}
\bar{\ell}_{\mathrm{P}} &= \pi_{\mathrm{p}}\ell_{\mathrm{p}} - \pi_{\mathrm{p}}\ell_{\mathrm{n}}, \\
\bar{\ell}_{\mathrm{U}} &= \ell_{\mathrm{n}}.
\end{aligned} \tag{91}$$

Proof. According to Table 7, M_{PU} is a child of M_{UU} on the reduction graph. Thus, replacing the subscripts $\{\mathrm{U}_1, \mathrm{U}_2\}$ of $\bar{P}$ and $\bar{L}$ with $\{\mathrm{P}, \mathrm{U}\}$ and assigning $\gamma_1 = 0$ and $\gamma_2 = \pi_{\mathrm{p}}$ as what we choose in Section 4.1.2, we follow the proof of Theorem 21 to conduct the risk rewrite: We first obtain the corrected losses (91) by plugging the assigned values into (86). Then, repeating the steps in (89), we achieve (90). □

Proof of Corollary 23

Corollary 23. *Assume $\pi_{\mathrm{p}} = 1/2$. For SU learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{\tilde{\mathrm{S}}}\left[\bar{\ell}_{\tilde{\mathrm{S}}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right],$$

where

$$\begin{aligned}
\bar{\ell}_{\tilde{\mathrm{S}}} &= \frac{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{p}} - \frac{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{n}}, \\
\bar{\ell}_{\mathrm{U}} &= -\frac{\pi_{\mathrm{n}}}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{p}} + \frac{\pi_{\mathrm{p}}}{2\pi_{\mathrm{p}} - 1}\ell_{\mathrm{n}}.
\end{aligned} \tag{92}$$

Proof. By Table 7, M_{SU} is a child of M_{UU}. Substituting the subscripts $\{\mathrm{U}_1, \mathrm{U}_2\}$ with subscripts $\{\tilde{\mathrm{S}}, \mathrm{U}\}$ and choosing $\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}$ and $\gamma_2 = \pi_{\mathrm{p}}$ as we did in Section 4.1.3, we obtain the corrected losses (92) by plugging the assigned values into (86). We note that $\pi_{\mathrm{p}} = 1/2$ ensures the choices of γ_1 and γ_2 above satisfy the $\gamma_1 + \gamma_2 = 1$ assumption discussed in Section 4.1.1. Then, we achieve the rewrite $R(g) = \mathsf{E}_{\tilde{\mathrm{S}}}\left[\bar{\ell}_{\tilde{\mathrm{S}}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right]$ by repeating the derivation for (89). □

Proof of Lemma 25

Lemma 25. *Let* (x, x') P_{S} *defined by (10). Then,* $\mathsf{E}_{\mathrm{S}}\left[\frac{L(X)}{2}\right] = \mathsf{E}_{\mathrm{S}}\left[\frac{L(X')}{2}\right]$.

Proof. For clarity, we simplify P_{S} as $c_1 P_{X|Y=\mathrm{p}} P_{X'|Y=\mathrm{p}} + c_2 P_{X|Y=\mathrm{n}} P_{X'|Y=\mathrm{n}}$, with $c_1 = \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2}$ and $c_2 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}^2}$. The lemma follows from

$$
\begin{aligned}
&\mathsf{E}_{\mathrm{S}}\left[L(X)\right] \\
&= \int_{x'} \int_{x} P_{\mathrm{S}} L(x)\, \mathrm{d}x\, \mathrm{d}x' \\
&= \int_{x'} \int_{x} \left(c_1 P_{x|Y=\mathrm{p}} P_{x'|Y=\mathrm{p}} + c_2 P_{x|Y=\mathrm{n}} P_{x'|Y=\mathrm{n}}\right) L(x)\, \mathrm{d}x\, \mathrm{d}x' \\
&= c_1 \int_{x} P_{x|Y=\mathrm{p}} L(x)\, \mathrm{d}x \int_{x'} P_{x'|Y=\mathrm{p}}\, \mathrm{d}x' + c_2 \int_{x} P_{x|Y=\mathrm{n}} L(x)\, \mathrm{d}x \int_{x'} P_{x'|Y=\mathrm{n}}\, \mathrm{d}x' \\
&= c_1 \int_{x} P_{x|Y=\mathrm{p}} L(x)\, \mathrm{d}x + c_2 \int_{x} P_{x|Y=\mathrm{n}} L(x)\, \mathrm{d}x,
\end{aligned}
$$

and similarly,

$$
\mathsf{E}_{\mathrm{S}}\left[L(X')\right] = c_1 \int_{x'} P_{x'|Y=\mathrm{p}} L(x')\, \mathrm{d}x' + c_2 \int_{x'} P_{x'|Y=\mathrm{n}} L(x')\, \mathrm{d}x'.
$$

□

Proof of Corollary 26

Corollary 26. *For Pcomp learning, the classification risk can be rewritten as*

$$
R(g) = \mathsf{E}_{\mathrm{Sup}}\left[\bar{\ell}_{\mathrm{Sup}}\right] + \mathsf{E}_{\mathrm{Inf}}\left[\bar{\ell}_{\mathrm{Inf}}\right],
$$

where

$$
\begin{aligned}
\bar{\ell}_{\mathrm{Sup}} &= \ell_{\mathrm{p}} - \pi_{\mathrm{p}} \ell_{\mathrm{n}}, \\
\bar{\ell}_{\mathrm{Inf}} &= -\pi_{\mathrm{n}} \ell_{\mathrm{p}} + \ell_{\mathrm{n}}.
\end{aligned} \tag{96}
$$

Proof. Since M_{UU} reduces to M_{Pcomp} with $\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}+\pi_{\mathrm{n}}^2}$ and $\gamma_2 = \frac{\pi_{\mathrm{p}}^2}{\pi_{\mathrm{p}}^2+\pi_{\mathrm{n}}}$ according to Table 7, we replace the subscripts $\{\mathrm{U}_1, \mathrm{U}_2\}$ with $\{\mathrm{Sup}, \mathrm{Inf}\}$ and instantiate (86) with the assigned values to obtain the corrected losses $\bar{\ell}_{\mathrm{Sup}}$ and $\bar{\ell}_{\mathrm{Inf}}$ (96). Then, repeating the same steps in (89), we have the corollary. □

Proof of Corollary 27

Corollary 27. *Assume* $\pi_{\mathrm{p}} = 1/2$. *For DU learning, the classification risk can be rewritten as*

$$
R(g) = \mathsf{E}_{\tilde{\mathrm{D}}}\left[\bar{\ell}_{\tilde{\mathrm{D}}}\right] + \mathsf{E}_{\mathrm{U}}\left[\bar{\ell}_{\mathrm{U}}\right],
$$

where

$$
\begin{aligned}
\bar{\ell}_{\tilde{\mathrm{D}}} &= 2\pi_{\mathrm{p}}\pi_{\mathrm{n}} \left(\frac{1}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}} \ell_{\mathrm{p}} - \frac{1}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}} \ell_{\mathrm{n}}\right), \\
\bar{\ell}_{\mathrm{U}} &= -\frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}} \ell_{\mathrm{p}} + \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{n}} - \pi_{\mathrm{p}}} \ell_{\mathrm{n}}.
\end{aligned} \tag{98}
$$

Proof. By Table 7, M_{DU} is reduced from M_{UU}. Thus, replacing $\{\mathrm{U}_1, \mathrm{U}_2\}$ with $\{\tilde{\mathrm{D}}, \mathrm{U}\}$, and assigning $\gamma_1 = 1/2$ and $\gamma_2 = \pi_{\mathrm{p}}$, we obtain the corrected losses (98) by plugging the assigned values into (86). Note that $\pi_{\mathrm{p}} = 1/2$ implies that the γ_1 and γ_2 assignments are feasible. Then, repeating the steps in (89), we have the corollary. □

Proof of Lemma 28

Lemma 28. *Given B (46) and following the DU learning notations, we have*

$$M_{\mathrm{DU}}B = \begin{pmatrix} P_{\tilde{\mathrm{D}}} \\ P_{\mathrm{U}} \end{pmatrix} = \bar{P},$$

where

$$M_{\mathrm{DU}} := \begin{pmatrix} \frac{\int_x \mathcal{X} P_{x\,|Y=\mathrm{n}}\mathrm{d}x}{2} & \frac{\int_x \mathcal{X} P_{x\,|Y=\mathrm{p}}\mathrm{d}x}{2} \\ \pi_{\mathrm{p}} & \pi_{\mathrm{n}} \end{pmatrix}.$$

Proof. Since $\int_x \mathcal{X} P_{x\,|Y=\mathrm{n}}\mathrm{d}x = 1$ and $\int_x \mathcal{X} P_{x\,|Y=\mathrm{p}}\mathrm{d}x = 1$, we have $M_{\mathrm{DU}} = M_{\mathrm{DU}}$ and hence $M_{\mathrm{DU}}B = M_{\mathrm{DU}}B = \begin{pmatrix} P_{\tilde{\mathrm{D}}} \\ P_{\mathrm{U}} \end{pmatrix}$. The last equality follows from Lemma 7. □

Proof of Lemma 29

Lemma 29. *Let (x, x') P_{D} defined in (12). Then, $\mathsf{E}_{\mathrm{D}}\left[\frac{L(X)}{2}\right] = \mathsf{E}_{\mathrm{D}}\left[\frac{L(X')}{2}\right]$.*

Proof. Recall $P_{\mathrm{D}} = \frac{1}{2}(P_{x|Y=\mathrm{p}}P_{x'\,|Y=\mathrm{n}} + P_{x|Y=\mathrm{n}}P_{x'\,|Y=\mathrm{p}})$. Following the similar argument in Lemma 25,

$$\begin{aligned} \mathsf{E}_{\mathrm{D}}\left[\frac{L(X)}{2}\right] &= \int_x \mathcal{X} \int_{x'} \mathcal{X} \left(P_{x|Y=\mathrm{p}}P_{x'\,|Y=\mathrm{n}} + P_{x|Y=\mathrm{n}}P_{x'\,|Y=\mathrm{p}}\right) \frac{L(x)}{4}\,\mathrm{d}x\,\mathrm{d}x' \\ &= \int_x \mathcal{X} \left(P_{x|Y=\mathrm{p}} + P_{x|Y=\mathrm{n}}\right) \frac{L(x)}{4}\,\mathrm{d}x \end{aligned}$$

and

$$\mathsf{E}_{\mathrm{D}}\left[\frac{L(X')}{2}\right] = \int_{x'} \mathcal{X} \left(P_{x'\,|Y=\mathrm{n}} + P_{x'\,|Y=\mathrm{p}}\right) \frac{L(x')}{4}\,\mathrm{d}x'$$

prove the lemma. □

Proof of Lemma 30

Corollary 30. *Assume $\pi_{\mathrm{p}} = 1/2$. For SD learning, the classification risk can be rewritten as*

$$R(g) = \mathsf{E}_{\tilde{\mathrm{S}}}\left[\bar{\ell}_{\tilde{\mathrm{S}}}\right] + \mathsf{E}_{\tilde{\mathrm{D}}}\left[\bar{\ell}_{\tilde{\mathrm{D}}}\right],$$

where

$$\begin{aligned} \bar{\ell}_{\tilde{\mathrm{S}}} &= \left(\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2\right)\left(\frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}} - \frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}\right), \\ \bar{\ell}_{\tilde{\mathrm{D}}} &= 2\pi_{\mathrm{p}}\pi_{\mathrm{n}}\left(-\frac{\pi_{\mathrm{n}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{p}} + \frac{\pi_{\mathrm{p}}}{\pi_{\mathrm{p}} - \pi_{\mathrm{n}}}\ell_{\mathrm{n}}\right). \end{aligned} \tag{101}$$

Proof. By Table 7, M_{SD} is reduced from M_{UU}. Thus, replacing $\{\mathrm{U}_1, \mathrm{U}_2\}$ with $\{\tilde{\mathrm{S}}, \tilde{\mathrm{D}}\}$, and assigning $\gamma_1 = \frac{\pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2}$ and $\gamma_2 = 1/2$, we obtain the corrected losses (101) by plugging the assigned values into (86). Note that $\pi_{\mathrm{p}} = 1/2$ implies that the γ_1 and γ_2 assignments are feasible. Then, repeating the steps in (89), we have the corollary. □

Proof of Lemma 31

Lemma 31. *Given B (46) and following the SD learning notations, we have*

$$M_{\mathrm{SD}} B = \begin{pmatrix} P_{\tilde{\mathrm{S}}} \\ P_{\tilde{\mathrm{D}}} \end{pmatrix} = \bar{P},$$

where

$$M_{\mathrm{SD}} := \begin{pmatrix} \frac{\pi_{\mathrm{p}}^2 \int_{x \in \mathcal{X}} P_{x\,|Y=\mathrm{p}} \mathrm{d}x}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2} & \frac{\pi_{\mathrm{n}}^2 \int_{x \in \mathcal{X}} P_{x\,|Y=\mathrm{n}} \mathrm{d}x}{\pi_{\mathrm{p}}^2 + \pi_{\mathrm{n}}^2} \\ \frac{\int_{x \in \mathcal{X}} P_{x\,|Y=\mathrm{n}} \mathrm{d}x}{2} & \frac{\int_{x \in \mathcal{X}} P_{x\,|Y=\mathrm{p}} \mathrm{d}x}{2} \end{pmatrix}.$$

Proof. Since $\int_{x \in \mathcal{X}} P_{x\,|Y=\mathrm{p}} \mathrm{d}x = 1$ and $\int_{x \in \mathcal{X}} P_{x\,|Y=\mathrm{n}} \mathrm{d}x = 1$, we have $M_{\mathrm{SD}} = M_{\mathrm{SD}}$ and hence $M_{\mathrm{SD}} B = M_{\mathrm{SD}} B = \begin{pmatrix} P_{\tilde{\mathrm{S}}} \\ P_{\tilde{\mathrm{D}}} \end{pmatrix}$. The last equality follows from Lemma 8. □

Proof of Lemma 32

Lemma 32. *Assume the formulation $\bar{P} = M_{\mathrm{Sconf}} B$ (58) is given. Suppose a vector of corrected losses $\bar{L}$ of the form $\begin{pmatrix} \tilde{\ell}_1(x) & \tilde{\ell}_2(x) \end{pmatrix}$ is independent of x. Then, we have*

$$\int_{x \in \mathcal{X}} \bar{L} \ \bar{P} \, \mathrm{d}x = \bar{L} \ \tilde{M}_{\mathrm{Sconf}} P, \tag{103}$$

where

$$\tilde{M}_{\mathrm{Sconf}} = \begin{pmatrix} \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{r - \pi_{\mathrm{n}}} \mathrm{d}x & \int_x \frac{\pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{p}}}{r - \pi_{\mathrm{n}}} \mathrm{d}x \\ \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}}}{\pi_{\mathrm{p}} - r} \mathrm{d}x & \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{\pi_{\mathrm{p}} - r} \mathrm{d}x \end{pmatrix}.$$

Proof. **of Lemma 32.** We replace $\bar{P}$ using $\bar{P} = M_{\mathrm{Sconf}} B$ (58). Since $\bar{L}$ is independent of x, we can move the integral over x into M_{Sconf} to obtain

$$\begin{aligned} \int_{x \in \mathcal{X}} \bar{L} \ \bar{P} \, \mathrm{d}x &= \int_x \bar{L} \begin{pmatrix} \frac{\pi_{\mathrm{p}}\left(\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}\right)}{r - \pi_{\mathrm{n}}} & \frac{\pi_{\mathrm{p}}\left(\pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{p}}\right)}{r - \pi_{\mathrm{n}}} \\ \frac{\pi_{\mathrm{n}}\left(\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}}\right)}{\pi_{\mathrm{p}} - r} & \frac{\pi_{\mathrm{n}}\left(\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}\right)}{\pi_{\mathrm{p}} - r} \end{pmatrix} \begin{pmatrix} P_{X|\mathrm{p}} \\ P_{X|\mathrm{n}} \end{pmatrix} \mathrm{d}x \\ &= \bar{L} \begin{pmatrix} \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{r - \pi_{\mathrm{n}}} \mathrm{d}x & \int_x \frac{\pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{p}}}{r - \pi_{\mathrm{n}}} \mathrm{d}x \\ \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}}}{\pi_{\mathrm{p}} - r} \mathrm{d}x & \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{\pi_{\mathrm{p}} - r} \mathrm{d}x \end{pmatrix} \begin{pmatrix} \pi_{\mathrm{p}} P_{X|\mathrm{p}} \\ \pi_{\mathrm{n}} P_{X|\mathrm{n}} \end{pmatrix}. \end{aligned}$$

Since $\pi_{\mathrm{p}} P_{X|\mathrm{p}} = P_{X,Y=\mathrm{p}}$ and $\pi_{\mathrm{n}} P_{X|\mathrm{n}} = P_{X,Y=\mathrm{n}}$, $\begin{pmatrix} \pi_{\mathrm{p}} P_{X|\mathrm{p}} \\ \pi_{\mathrm{n}} P_{X|\mathrm{n}} \end{pmatrix} = P$. Furthermore, comparing the equality in the above derivation with (103), we have

$$\tilde{M}_{\mathrm{Sconf}} = \begin{pmatrix} \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{r - \pi_{\mathrm{n}}} \mathrm{d}x & \int_x \frac{\pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{p}}}{r - \pi_{\mathrm{n}}} \mathrm{d}x \\ \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}}}{\pi_{\mathrm{p}} - r} \mathrm{d}x & \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{\pi_{\mathrm{p}} - r} \mathrm{d}x \end{pmatrix}$$

that completes the proof. □

Proof of Lemma 33

Lemma 33. *Let*

$$\tilde{M}^{\dagger}_{\mathrm{Sconf}} := \begin{pmatrix} \frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} & 0 \\ 0 & \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} \end{pmatrix}.$$

Then,

$$\tilde{M}^{\dagger}_{\mathrm{Sconf}} \tilde{M}_{\mathrm{Sconf}} = I.$$

Proof. We prove the lemma by examining each entry of $\tilde{M}^{\dagger}_{\mathrm{Sconf}} \tilde{M}_{\mathrm{Sconf}}$. The value of $(1,1)$ entry is

$$\begin{aligned} \frac{r-\pi_{\mathrm{n}}}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{r-\pi_{\mathrm{n}}} \mathrm{d}x &= \frac{1}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} \left(\pi_{\mathrm{p}}^2 \int_x P_{x\,|\mathrm{p}} \mathrm{d}x - \pi_{\mathrm{n}}^2 \int_x P_{x\,|\mathrm{n}} \mathrm{d}x \right) \\ &= \frac{\pi_{\mathrm{p}}^2 - \pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} = 1. \end{aligned}$$

The $(2,2)$ entry has value

$$\begin{aligned} \frac{\pi_{\mathrm{p}}-r}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} \int_x \frac{\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}}}{\pi_{\mathrm{p}}-r} \mathrm{d}x &= \frac{1}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} \left(\pi_{\mathrm{p}}^2 \int_x P_{x\,|\mathrm{p}} \mathrm{d}x - \pi_{\mathrm{n}}^2 \int_x P_{x\,|\mathrm{n}} \mathrm{d}x \right) \\ &= \frac{\pi_{\mathrm{p}}^2 - \pi_{\mathrm{n}}^2}{\pi_{\mathrm{p}}-\pi_{\mathrm{n}}} = 1. \end{aligned}$$

The $(1,2)$ entry is zero since $\int_x \left(\pi_{\mathrm{n}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{n}}^2 P_{x\,|\mathrm{p}}\right) \mathrm{d}x = 0$. Similarly, since $\int_x \left(\pi_{\mathrm{p}}^2 P_{x\,|\mathrm{n}} - \pi_{\mathrm{p}}^2 P_{x\,|\mathrm{p}}\right) \mathrm{d}x = 0$, the $(2,1)$ entry is also zero. □

D.2 Omitted Proofs in Section 5.2

Proof of Corollary 38

Corollary 38. *The decontamination matrix $M^{\dagger}_{\mathrm{PCPL}}$ for PCPL equals $M^{\dagger}_{\mathrm{PPL}}$. If we define the corrected losses as $\bar{L} := L \; M^{\dagger}_{\mathrm{PCPL}}$, the classification risk for PCPL learning can be rewritten as*

$$R(g) = \mathsf{E}_{S,X}\left[\bar{\ell}_S\right],$$

where

$$\bar{\ell}_S = \sum_{i \;\; S} \frac{P_{Y=i|X}}{\sum_{a \;\; S} P_{Y=a|X}} \ell_{Y=i}. \tag{110}$$

Proof. The proof follows the standard argument: First, find out $M^{\dagger}_{\mathrm{PCPL}}$, then construct the corrected losses to rewrite the risk.

Since M_{PCPL} is reduced from M_{PPL}, we can exploit Lemma 36. Note that the only difference $C(S,X)$ between the formulations of PCPL and PPL cancels itself out in the derivation of $P_{Y=i|S=s_j,X}$ (please refer to the proof of Lemma 36 for a detailed derivation), the (i,j) entry of $M^{\dagger}_{\mathrm{PCPL}}$ coincides with that of $M^{\dagger}_{\mathrm{PPL}}$ for all i and j, proving the first assertion.

Since $M^{\dagger}_{\mathrm{PPL}}$ and $M^{\dagger}_{\mathrm{PCPL}}$ are identical, $L \;\; M^{\dagger}_{\mathrm{PPL}} = L \;\; M^{\dagger}_{\mathrm{PCPL}}$ gives (110):

$$\begin{aligned} \bar{\ell}_{S=s_j} = \left(L \;\; M^{\dagger}_{\mathrm{PCPL}}\right)_j &= \sum_{i=1}^{K} \frac{P_{Y=i|X}\, \mathsf{I}\left[Y=i \quad s_j\right]}{\sum_{a \;\; s_j} P_{Y=a|X}} \ell_{Y=i} \\ &= \sum_{i \;\; s_j} \frac{P_{Y=i|X}}{\sum_{a \;\; s_j} P_{Y=a|X}} \ell_{Y=i}. \end{aligned}$$

Being identical to $M_{\mathrm{PPL}}^{\dagger}$ also means that $M_{\mathrm{PCPL}}^{\dagger}$ is derived from $M_{\mathrm{gCCN}}^{\dagger}$. Thus, we can continue (107) to rewrite the risk by repeating the proof of Corollary 37:

$$R(g) = \int_{x\ \ X} \bar{L}\ \ \bar{P}\mathrm{d}x = \int_{x\ \ X} \sum_{j=1}^{|S|} P_{S=s_j,x}\bar{\ell}_{S=s_j}\mathrm{d}x = \mathsf{E}_{S,X}\left[\bar{\ell}_S\right].$$

□

Proof of Lemma 39

Lemma 39. *Let (s, s') be a pair of partial-labels satisfying $s = Y \setminus s'$. Then,*

$$P_{S=s,X}\bar{\ell}_{S=s} + P_{S=s',X}\bar{\ell}_{S=s'} = P_{S=s,X}\sum_{i=1}^{K}\frac{P_{Y=i|X}\ell_{Y=i}}{\sum_{a\ \ s'} P_{Y=a|X}}.$$

Proof. Given M_{PCPL} (68), we apply $\bar{P} = M_{\mathrm{PCPL}}B$ to obtain

$$P_{S=s',X} = \frac{\sum_{k=1}^{K}\mathsf{I}\left[Y = k \quad s'\right]P_{Y=k,X}}{2^{K-1}-1}.$$

We also have

$$\bar{\ell}_{S=s'} = \frac{\sum_{i\ \ s'} P_{Y=i|X}\ell_{Y=i}}{\sum_{a\ \ s'} P_{Y=a|X}}$$

according to (110). Since

$$\frac{\sum_{k=1}^{K}\mathsf{I}\left[Y = k \quad s'\right]P_{Y=k,X}}{\sum_{a\ \ s'} P_{Y=a|X}} = P_X = \frac{\sum_{k=1}^{K}\mathsf{I}\left[Y = k \quad s\right]P_{Y=k,X}}{\sum_{a\ \ s'} P_{Y=a|X}},$$

$$\begin{aligned} P_{S=s',X}\bar{\ell}_{S=s'} &= \frac{\sum_{k=1}^{K}\mathsf{I}\left[Y = k \quad s'\right]P_{Y=k,X}}{2^{K-1}-1}\frac{\sum_{i\ \ s'} P_{Y=i|X}\ell_{Y=i}}{\sum_{a\ \ s'} P_{Y=a|X}} \\ &= \frac{\sum_{k=1}^{K}\mathsf{I}\left[Y = k \quad s\right]P_{Y=k,X}}{2^{K-1}-1}\frac{\sum_{i\ \ s'} P_{Y=i|X}\ell_{Y=i}}{\sum_{a\ \ s'} P_{Y=a|X}}. \end{aligned}$$

Thus,

$$\begin{aligned} P_{S=s,X}\bar{\ell}_{S=s} + P_{S=s',X}\bar{\ell}_{S=s'} &= \frac{\sum_{k=1}^{K}\mathsf{I}\left[Y = k \quad s\right]P_{Y=k,X}}{2^{K-1}-1}\frac{\sum_{i\ \ s} P_{Y=i|X}\ell_{Y=i}}{\sum_{a\ \ s'} P_{Y=a|X}} \\ &\quad + \frac{\sum_{k=1}^{K}\mathsf{I}\left[Y = k \quad s\right]P_{Y=k,X}}{2^{K-1}-1}\frac{\sum_{i\ \ s'} P_{Y=i|X}\ell_{Y=i}}{\sum_{a\ \ s'} P_{Y=a|X}} \\ &= P_{S=s,X}\frac{\sum_{i\ \ s} P_{Y=i|X}\ell_{Y=i} + \sum_{i\ \ s'} P_{Y=i|X}\ell_{Y=i}}{\sum_{a\ \ s'} P_{Y=a|X}} \\ &= P_{S=s,X}\sum_{i=1}^{K}\frac{P_{Y=i|X}\ell_{Y=i}}{\sum_{a\ \ s'} P_{Y=a|X}} \end{aligned}$$

proves the lemma. □

Proof of Corollary 40

Corollary 40. *The (i,j) entry of the decontamination matrix $M_{\mathrm{MCL}}^{\dagger}$ is of the form*

$$\mathsf{P}_{Y=i|\bar{S}=\bar{s}_j,X} = \frac{\mathsf{P}_{Y=i|X}\mathsf{I}\left[Y=i \;/\; \bar{s}_j\right]}{\sum_{a\,/\,\bar{s}_j}\mathsf{P}_{Y=a|X}}. \tag{111}$$

Define the corrected losses $\bar{L} := L\ M_{\mathrm{MCL}}^{\dagger}$. Then, for MCL learning, the classification risk can be rewritten as

$$R(g) = \mathsf{E}_{\bar{S},X}\left[\bar{\ell}_{\bar{S}}\right],$$

where

$$\bar{\ell}_{\bar{S}} = \sum_{i\,/\,\bar{S}} \frac{\mathsf{P}_{Y=i|X}}{\sum_{a\,/\,\bar{S}}\mathsf{P}_{Y=a|X}}\ell_{Y=i}. \tag{112}$$

Proof. The proof follows the standard strategy in Section 5.2: We will first find out $M_{\mathrm{MCL}}^{\dagger}$, and then construct the corrected losses $\bar{L}$ to rewrite the risk.

Based on the notion in (71), we denote the (j,i) entry of M_{MCL} as

$$\frac{\mathsf{P}_{|\bar{S}|=|\bar{s}_j|}}{\binom{K-1}{|\bar{s}_j|}}\mathsf{I}\left[Y=i \;/\; \bar{s}_j\right] = C(\bar{s}_j,X)\mathsf{I}\left[Y=i \;/\; \bar{s}_j\right] = \mathsf{P}_{\bar{S}=\bar{s}_j|Y=i,X}. \tag{128}$$

Expressing M_{MCL} via (128) allows us to apply the argument for (106) to show that the (i,j) entry of $M_{\mathrm{MCL}}^{\dagger}$ is of the form $\mathsf{P}_{Y=i|\bar{S}=\bar{s}_j,X}$. Specifically, assigning $(M)_{j,i}$ in (40) as $\mathsf{P}_{\bar{S}=\bar{s}_j|Y=i,X}$ and applying marginal chain (i.e., Proposition 2), we obtain

$$\begin{aligned}
\left(M_{\mathrm{MCL}}^{\dagger}M_{\mathrm{MCL}}P\right)_i &= \sum_{j=1}^{|\mathsf{S}|}\left(M_{\mathrm{MCL}}^{\dagger}\right)_{i,j}\sum_{k=1}^{K}\left(M_{\mathrm{MCL}}\right)_{j,k}P_k \\
&= \sum_{j=1}^{|\mathsf{S}|}\mathsf{P}_{Y=i|\bar{S}=\bar{s}_j,X}\sum_{k=1}^{K}\mathsf{P}_{\bar{S}=\bar{s}_j|Y=k,X}\mathsf{P}_{Y=k,X} \\
&= \mathsf{P}_{Y=i,X} = P_i.
\end{aligned}$$

Then, we follow the same argument in Lemma 36 to calculate $\mathsf{P}_{Y=i|\bar{S}=\bar{s}_j,X}$ subject to (128). Note that $\mathsf{P}_{\bar{S}|Y,X} = C(\bar{S},X)\mathsf{I}\left[Y \;/\; \bar{S}\right]$ in (128) implies

$$\begin{aligned}
\sum_{b\ \bar{S}}\mathsf{P}_{\bar{S},Y=b|X} &= \sum_{b\ \bar{S}}\mathsf{P}_{\bar{S}|Y=b,X}\mathsf{P}_{Y=b|X} \\
&= \sum_{b\ \bar{S}}C(\bar{S},X)\mathsf{I}\left[Y=b \;/\; \bar{S}\right]\mathsf{P}_{Y=b|X} \\
&= 0.
\end{aligned}$$

Thus, $\mathsf{P}_{\bar{S}|X} = \sum_{b\ \bar{S}}\mathsf{P}_{\bar{S},Y=b|X} + \sum_{a\,/\,\bar{S}}\mathsf{P}_{\bar{S},Y=a|X} = \sum_{a\,/\,\bar{S}}\mathsf{P}_{\bar{S},Y=a|X}$. The fact further implies

$$\begin{aligned}
\mathsf{P}_{Y|\bar{S},X} = \frac{\mathsf{P}_{\bar{S},Y|X}}{\mathsf{P}_{\bar{S}|X}} &= \frac{\mathsf{P}_{\bar{S}|Y,X}\mathsf{P}_{Y|X}}{\sum_{a\,/\,\bar{S}}\mathsf{P}_{\bar{S}|Y=a,X}\mathsf{P}_{Y=a|X}} \\
&= \frac{C(\bar{S},X)\mathsf{I}\left[Y \;/\; \bar{S}\right]\mathsf{P}_{Y|X}}{\sum_{a\,/\,\bar{S}}C(\bar{S},X)\mathsf{I}\left[Y=a \;/\; \bar{S}\right]\mathsf{P}_{Y=a|X}} \\
&= \frac{\mathsf{P}_{Y|X}\mathsf{I}\left[Y \;/\; \bar{S}\right]}{\sum_{a\,/\,\bar{S}}\mathsf{P}_{Y=a|X}}.
\end{aligned}$$

Therefore, for $Y = i$ and $\bar{S} = \bar{s}_j$, we achieve

$$\mathsf{P}_{Y=i|\bar{S}=\bar{s}_j,X} = \frac{\mathsf{P}_{Y=i|X}\,\mathsf{I}\left[Y = i \;/\; \bar{s}_j\right]}{\sum_{a\,/\,\bar{s}_j} \mathsf{P}_{Y=a|X}}$$

that proves (111).

With $M^{\dagger}_{\mathrm{MCL}}$ in hand, we repeat the same argument in Corollary 37 to obtain

$$\begin{aligned}\bar{\ell}_{\bar{S}=\bar{s}_j} = \begin{pmatrix} L & M^{\dagger}_{\mathrm{MCL}} \end{pmatrix}_j &= \sum_{i=1}^{K} \frac{\mathsf{P}_{Y=i|X}\,\mathsf{I}\left[Y = i \;/\; \bar{s}_j\right]}{\sum_{a\,/\,\bar{s}_j} \mathsf{P}_{Y=a|X}} \ell_{Y=i} \\ &= \sum_{i\,/\,\bar{s}_j} \frac{\mathsf{P}_{Y=i|X}}{\sum_{a\,/\,\bar{s}_j} \mathsf{P}_{Y=a|X}} \ell_{Y=i}\end{aligned}$$

and

$$\begin{aligned}R(g) &= \int_{x\ \mathsf{X}} L\ \ P\,\mathrm{d}x = \int_{x\ \mathsf{X}} L\ \ M^{\dagger}_{\mathrm{MCL}} M_{\mathrm{MCL}} P\,\mathrm{d}x \\ &= \int_{x\ \mathsf{X}} \bar{L}\ \ \bar{P}\,\mathrm{d}x = \int_{x\ \mathsf{X}} \sum_{j=1}^{|\mathsf{S}|} \mathsf{P}_{\bar{S}=\bar{s}_j,x} \bar{\ell}_{\bar{S}=\bar{s}_j}\,\mathrm{d}x = \mathsf{E}_{\bar{S},X}\left[\bar{\ell}_{\bar{S}}\right]\end{aligned}$$

to complete the risk rewrite of MCL. □

D.3 Omitted Proofs in Section 5.3

Proof of Corollary 47

Corollary 47. *For SC-Conf learning, the classification risk can be written as*

$$R(g) = \pi_{y_{\mathrm{s}}} \mathsf{E}_{X|Y=y_{\mathrm{s}}} \left[\sum_{i=1}^{K} \frac{r_i(X)}{r_{y_{\mathrm{s}}}(X)} \ell_i \right].$$

Proof. The corollary follows from notation substitution and the same argument for Theorem 46. Specifically, we replace M_{Sub} with M_{SC}, Y_{s} with y_{s}, and $\frac{\mathsf{P}_{Y=i|X}}{\mathsf{P}_{Y\ \mathsf{Y}_{\mathrm{s}}|X}} \ell_i$ with $\frac{\mathsf{P}_{Y=i|X}}{\mathsf{P}_{Y=y_{\mathrm{s}}|X}} \ell_i$. □

Proof of Corollary 48

Corollary 48. *For Pconf learning, the classification risk can be written as*

$$R(g) = \pi_{\mathrm{p}} \mathsf{E}_{\mathrm{P}} \left[\ell_{\mathrm{p}} + \frac{1 - r(X)}{r(X)} \ell_{\mathrm{n}} \right].$$

Proof. Since

$$M^{\dagger}_{\mathrm{Pconf}} M_{\mathrm{Pconf}} = \begin{pmatrix} \frac{\mathsf{P}_{Y=\mathrm{p}|X}}{\mathsf{P}_{Y=\mathrm{p}|X}} & 0 \\ 0 & \frac{\mathsf{P}_{Y=\mathrm{p}|X}}{\mathsf{P}_{Y=\mathrm{n}|X}} \end{pmatrix} \begin{pmatrix} \frac{\mathsf{P}_{Y=\mathrm{p}|X}}{\mathsf{P}_{Y=\mathrm{p}|X}} & 0 \\ 0 & \frac{\mathsf{P}_{Y=\mathrm{n}|X}}{\mathsf{P}_{Y=\mathrm{p}|X}} \end{pmatrix} = I,$$

we define $\bar{L}\ := L\ \ M^{\dagger}_{\mathrm{Pconf}}$ and apply (39) to rewrite the risk as follows

$$\begin{aligned}R(g) &= \int_{x\ \mathsf{X}} L\ \ P\mathrm{d}x = \int_{x\ \mathsf{X}} \bar{L}\ \ \bar{P}\mathrm{d}x = \int_{x\ \mathsf{X}} \begin{pmatrix} \ell_{\mathrm{p}} & \frac{1-r(X)}{r(X)}\ell_{\mathrm{n}} \end{pmatrix} \begin{pmatrix} \mathsf{P}_{Y=\mathrm{p},X} \\ \mathsf{P}_{Y=\mathrm{p},X} \end{pmatrix} \mathrm{d}x \\ &= \mathsf{P}_{Y=\mathrm{p}} \mathsf{E}_{X|Y=\mathrm{p}} \left[\ell_{\mathrm{p}} + \frac{1 - r(X)}{r(X)} \ell_{\mathrm{n}} \right] \\ &= \pi_{\mathrm{p}} \mathsf{E}_{\mathrm{P}} \left[\ell_{\mathrm{p}} + \frac{1 - r(X)}{r(X)} \ell_{\mathrm{n}} \right].\end{aligned}$$

□

Proof of Corollary 49

Corollary 49. *For soft-label learning, the classification risk can be written as*

$$R(g) = \mathsf{E}_X \left[\sum_{i=1}^{K} \mathsf{P}_{Y=i|X} \ell_i \right] = \mathsf{E}_X \left[\sum_{i=1}^{K} r_i(X) \ell_i \right].$$

Proof. Defining $\bar{L} := L \ M_{\text{Soft}}^{\dagger}$ and recalling $\bar{P}$ in (83), we apply (39) to obtain

$$\begin{aligned} R(g) &= \int_x \int_{\mathsf{X}} L \ P \mathrm{d}x = \int_x \int_{\mathsf{X}} \bar{L} \ \bar{P} \mathrm{d}x = \int_x \int_{\mathsf{X}} \sum_{i=1}^{K} \mathsf{P}_{Y=i|X} \ell_i \cdot \mathsf{P}_X \mathrm{d}x \\ &= \mathsf{E}_X \left[\sum_{i=1}^{K} \mathsf{P}_{Y=i|X} \ell_i \right]. \end{aligned}$$

□