
A Unified Framework for Generalization Error Analysis of Learning with Arbitrary Discrete Weak Features

Kosuke Sugiyama¹ Masato Uchida¹

Abstract

In many real-world applications, predictive tasks inevitably involve low-quality input features (Weak Features; WFs) which arise due to factors such as misobservations, missingness, or partial observations. While several methods have been proposed to estimate the true values of specific types of WFs and to solve a downstream task, a unified theoretical framework that comprehensively addresses these methods remains underdeveloped. In this paper, we propose a unified framework called Weak Features Learning (WFL), which accommodates arbitrary discrete WFs and a broad range of learning algorithms, and we demonstrate its validity. Furthermore, we introduce a class of algorithms that learn both the estimation model for WFs and the predictive model for a downstream task and perform a generalization error analysis under finite-sample conditions. Our results elucidate the interdependencies between the estimation errors of WFs and the prediction error of a downstream task, as well as the theoretical conditions necessary for the learning approach to achieve consistency. This work establishes a unified theoretical foundation, providing generalization error analysis and performance guarantees, even in scenarios where WFs manifest in diverse forms.

1. Introduction

The performance and explainability of machine learning models are highly dependent on the quality of the training data. However, in practical applications, obtaining high-quality data is often infeasible due to constraints such as

¹Major in Computer Science and Communications Engineering, Waseda University, Tokyo, Japan. Correspondence to: Kosuke Sugiyama <kosuke0322@asagi.waseda.jp>, Masato Uchida <m.uchida@waseda.jp>.

privacy concerns, high observation costs, and uncertainties in sensor measurements. Consequently, challenges often stem from low-quality labels (referred to as Weak Labels; WLs) that contain incorrect, incomplete, or ambiguous supervisory information, and low-quality features (referred to as Weak Features; WFs), which manifest as misobservations, missing values, or ambiguous observations. Extensive research has focused on WLs within the framework of weakly supervised learning (WSL), covering methods such as semi-supervised learning and learning with label noise, which have provided substantial theoretical guarantees (Chapelle et al., 2006; Elkan & Noto, 2008; Cour et al., 2011; Natarajan et al., 2013; Ishida et al., 2017). In contrast, research on WFs has proposed methods such as impute-then-regress (ItR), which imputes missing before regression (Josse et al., 2024; Bertsimas et al., 2021; Le Morvan et al., 2020b; 2021), and complementary features learning (CFL), which leverages complementary features (CFs) that differ from the true values (Sugiyama & Uchida, 2024). However, a unified framework that provides consistent theoretical guarantees across the various forms of WFs remains underexplored.

Motivated by these challenges, this study focuses on weak features learning (WFL), a generalized learning problem involving arbitrary WFs. A common approach involves sequentially or iteratively learning *feature estimation models* g to estimate the true values of WFs (referred to as exact values) and a *label prediction model* f to predict downstream task labels using the outputs of g . This strategy is considered rational for improving both the quality of WFs and the predictive performance of a downstream task (Yoon et al., 2018; Mattei & Frellsen, 2019; Le Morvan et al., 2020a; Ipsen et al., 2021; Zaffran et al., 2023; Ipsen et al., 2022; Sugiyama & Uchida, 2024). Indeed, prior approaches such as ItR and CFL are grounded in this learning strategy, utilizing various machine learning methods to construct g . These methods aim to improve the quality of WFs, thereby enhancing the generalization performance and explainability of f in a downstream task.

However, in the context of WFL, where various methods have been proposed, significant gaps remain in understanding how the learning of g and f impacts each other’s learning efficiency and under what conditions WFL can achieve

optimal g and f (i.e., consistency). For example, in ItR, while the conditions under which g and f become Bayes rules have been analyzed, the generalization error analysis under finite samples and any data distribution has not yet been sufficiently conducted (Josse et al., 2024; Bertsimas et al., 2021; Le Morvan et al., 2020b; 2021). Also, in CFL, which addresses CFs, theoretical analysis remains underdeveloped, and clear guidelines for handling diverse forms of WFs in a unified manner have yet to be established (Sugiyama & Uchida, 2024).

This study aims to investigate the mutual influence between estimating the exact values of WFs and learning a downstream task. Specifically, we focus on scenarios involving discrete WFs (hereafter referred to as discrete WFL) and provide a unified formalization. By performing finite-sample error analysis for a generalized class of learning algorithms, we systematically address these questions. The proposed theoretical framework extends beyond situations involving missing values or CFs. It also accommodates cases where WFs arise in the diverse forms discussed in the WFs literature, such as erroneous observations or scenarios where only a candidate set containing the exact value is observed (Natarajan et al., 2013; Cour et al., 2011; Feng et al., 2020; Xu et al., 2021). Consequently, the framework not only reinterprets existing approaches such as ItR and CFL but also offers new theoretical insights into previously unexplored WF settings. Furthermore, as the class of learning algorithms analyzed in this study encompasses various existing methods (Yoon et al., 2018; Mattei & Frellsen, 2019; Ipsen et al., 2021; Josse et al., 2024; Le Morvan et al., 2021; Ipsen et al., 2022; Sugiyama & Uchida, 2024), our results provide a unified theoretical evaluation framework for these methods. We have also developed a unified formulation and analysis for scenarios involving continuous WFs, and the results for continuous WFs are presented in a separate paper (Sugiyama & Uchida, 2025).

The main contributions of this study are as follows:

1. We propose a risk-based formulation to address arbitrary discrete WFs and demonstrate that the introduced objective function facilitates the learning of f , which captures the true input-output relationship. This validates the proposed formulation (Section 3.2).
2. Within the proposed formulation, we define the Learning Algorithm Class for discrete WFL (LAC-dWFL), which flexibly combines three steps: (i) learning g using WFs as weak supervision, (ii) learning f with a fixed g , and (iii) learning g with a fixed f . This framework accommodates both sequential and iterative learning approaches, offering a unified perspective on diverse methods (Section 3.3).
3. For step (ii), we derive the error bound for f given any fixed g (Section 4.2), providing theoretical insights into how

the estimation errors of g influence the error bound for f . By integrating the theoretical framework of WSL in (i), we further analyze how the properties of WFs influence f 's error bound, how the order of the error bound evolves with the learning of g , and the conditions under which sequential learning in LAC-dWFL (performing steps (i) and (ii) in sequence) achieves consistency.

4. For step (iii), we evaluate the generalization error of g given any fixed f (Section 4.3) and analyze how the properties and generalization performance of f affect the error bound for g . By integrating the results of Contributions 3 and 4, we establish the conditions under which iterative learning in LAC-dWFL (alternating steps (ii) and (iii)) achieves the consistency for $f \circ g$.

5. We validate the proposed theoretical framework on real-world datasets to evaluate how accurately the derived error bounds reflect actual learning behaviors (Section 5).

2. Related work

2.1. Learning Problems involving WFs

ItR and CFL are representative learning problems that serve as special cases of WFL, involving specific types of WFs. This section provides an overview of these frameworks.

In ItR, the focus lies on scenarios where input features contain missing values. The fundamental approach of ItR involves imputing the missing values and utilizing the resulting complete dataset to learn a label prediction model f for a downstream task. Sequential learning methods have been proposed, employing feature estimation models g that utilize techniques such as constant imputation (Josse et al., 2024) or machine learning-based approaches (Yoon et al., 2018; Mattei & Frellsen, 2019; Ipsen et al., 2021). Furthermore, joint or iterative optimization of g and f has also been explored in the literature (Le Morvan et al., 2020a; Ipsen et al., 2022).

Theoretical analyses of ItR have investigated whether combinations of g and f exist that equal Bayes rule, focusing on regression or classification as downstream tasks (Josse et al., 2024; Bertsimas et al., 2021; Le Morvan et al., 2021). However, these studies do not address the existence of learning algorithms capable of constructing optimal g and f . Moreover, analyses in finite-sample settings have been confined to restrictive cases where the true f is assumed to be a linear model, leaving more general problem settings unexplored (Le Morvan et al., 2020b). Additionally, these analyses often restrict downstream tasks to regression or classification and, in some cases, confine the loss function of f to mean squared error (Le Morvan et al., 2020b; 2021). In this paper, while we focus on discrete WFs, we achieve a generalization error analysis of WFL in finite-sample settings by assum-

ing only the boundedness and Lipschitz continuity of the loss function of f , without imposing any constraints on a downstream task and a data distribution.

In CFL, the primary focus lies on learning problems where input features include CFs (Sugiyama & Uchida, 2024). Within this framework, the feature estimation model g and the label prediction model f are defined as probabilistic models, and an objective function leveraging the Kullback-Leibler divergence has been derived to learn g and f . For learning g in CFL, complementary label learning (CLL) (Ishida et al., 2017; 2019; Yu et al., 2018; Lin & Lin, 2023; Ruan et al., 2024), a weakly supervised learning approach that predicts true labels from datasets labeled exclusively with incorrect labels, can be employed. Moreover, since partial label learning (PLL) (Cour et al., 2011; Feng et al., 2020; Xu et al., 2021; Tian et al., 2023), where supervision is provided in the form of a set containing the true label, can be interpreted as a generalized learning problem of CLL (Katsura & Uchida, 2020). Therefore, PLL is also applicable to the learning of g .

However, in CFL, the learning behavior under finite samples, as well as the conditions required to obtain asymptotically optimal g and f , have yet to be theoretically clarified. In this paper, we restrict g and f to deterministic models and perform a generalization error analysis of WFL for arbitrary downstream task and bounded, Lipschitz-continuous loss functions, thereby shedding light on the theoretical properties of CFL.

2.2. WFs Whose Exact Values Can Be Estimated

When constructing a feature estimation model g , it is natural to treat the observed values of each WF as WLs and employ the WSL methods. In fact, depending on the types and settings of WFs, g can be learned using WSL methods. In WSL, a variety of WL settings have been studied, including the aforementioned CLL and PLL. For instance, noisy label learning (Natarajan et al., 2013) deals with learning from data containing incorrect labels, while positive-unlabeled learning (Elkan & Noto, 2008) focuses on learning binary classifiers using only positive and unlabeled samples.

Various WSL methods have been theoretically formulated, and their generalization error has been analyzed under finite-sample conditions (Cour et al., 2011; Feng et al., 2020; Xu et al., 2021; Natarajan et al., 2013; Ishida et al., 2017; Yu et al., 2018). Many of these objective functions were defined using unbiased estimators of the expected risk in supervised learning, the expected risk’s upper bounds that are computable with WLs, or risks whose optimal solutions align those of the expected risk. Their theoretical analyses elucidated the conditions under which optimal hypotheses can be obtained by minimizing each objective function, as well as the relationship between WL settings and the error

bounds. Thus, if WSL is employed to learn g in WFL, the learning of g alone can be analyzed based on the WSL theories.

However, in WFL, it is necessary to consider the learning of both g and f . For example, in sequential learning, the learning of f depends on the output of g . In iterative learning, the learning of g depends on f , unlike the case where g is learned solely using WSL. Therefore, a theoretical discussion that establishes the relationship between g and f is essential. In this paper, we perform a generalization error analysis of WFL, explicitly considering the relationship between g and f .

3. Formulation

3.1. Review of ERM

In this paper, we formulate WFL from the perspective of risk minimization and adopt empirical risk minimization (ERM) as the learning framework. Below, we briefly review ERM in ordinary supervised learning (Shalev-Shwartz & Ben-David, 2014; Mohri et al., 2018). Let the input space be $\mathcal{X} \subseteq \mathbb{R}^d$ and the label space be $\mathcal{Y} \subseteq \mathbb{R}$. Here, $d \in \mathbb{N}_+$ represents an input dimension. We denote the random variables representing an instance by $\mathbf{X}$ and the random variable representing labels by Y , assuming that $(\mathbf{X}, Y)$ follows the true distribution $p_*(\mathbf{x}, y)$ over $\mathcal{X} \times \mathcal{Y}$ independently and identically distributed (i.i.d.). The goal in the ERM framework is to find a label prediction model $f : \mathcal{X} \rightarrow \mathbb{R} \in \mathcal{F}$ that minimizes the expected risk:

$$R_l(f) := \mathbb{E}_{p_*(\mathbf{x}, y)}[l(f(\mathbf{X}), Y)], \quad (3.1)$$

where $l : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ is a loss function, and $\mathcal{F}$ is the hypothesis set of label prediction models. Since only finite samples are available in practice, ERM approximates the expected risk with the empirical risk computed as a sample average and learns f by minimizing this empirical risk.

3.2. Formulation of discrete WFL

In this section, we formulate WFL based on risk minimization and employ ERM as the learning method. To account for the presence of WFs in instances, we decompose $\mathbf{X}$ representing an instance, into $\mathbf{X}^w$ representing the exact values of WFs and $\mathbf{X}^o$ representing the remaining ordinary features (OFs), such that $\mathbf{X} = (\mathbf{X}^w, \mathbf{X}^o)$. Let $\mathcal{X}^w$ and $\mathcal{X}^o$ denote the domains of $\mathbf{X}^w$ and $\mathbf{X}^o$, respectively, with $\mathcal{X}^w \times \mathcal{X}^o = \mathcal{X}$. Here, $\mathcal{X}^w$ is assumed to be a finite set. We denote the observed values of WFs as the random variables $\bar{\mathbf{X}}^w$, which follows the probability distribution $\bar{p}_*(\bar{\mathbf{x}}^w | \mathbf{x}, y)$.

We define the feature estimation models for estimating the exact values of WFs as $\mathbf{g} := (g_1, \dots, g_{F^w}) \in \mathcal{G} := \mathcal{G}_1 \times \dots \times \mathcal{G}_{F^w} : \mathcal{X}^o \rightarrow \mathcal{X}^w$, where F^w is the number of

WFs, and each $g_j \in \mathcal{G}_j : \mathcal{X}^o \rightarrow \mathcal{X}_j^w, \forall j \in [F^w]$. Here, $[F^w] := \{1, \dots, F^w\}$, and $\mathcal{G}_j$ represents the hypothesis set for estimating X_j^w . The probability mass function (PMF) representation of $\mathbf{g}$ is defined as $q_{\mathbf{g}}(\mathbf{x}^w | \mathbf{x}^o) := \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{x}^o)]}$. For simplicity, this paper primarily focuses on binary classification, but the proposed formulation and analyses can be easily extended to other prediction tasks such as multi-class classification or regression.

The primary objectives of WFL are to improve the generalization performance of a downstream task and to restore explainability lost due to WFs. The primary factor reducing explainability is the inaccuracy of information provided by WFs. The most natural approach to address this issue is to estimate the exact values of WFs accurately. Accordingly, WFL aims to learn f and $\mathbf{g}$ that minimize the following two risks. The first risk evaluates the generalization error of f :

$$\begin{aligned} R_{l,\mathbf{g}}(f) &:= \mathbb{E}_{p_*(\mathbf{x}^o, y)} q_{\mathbf{g}}(\mathbf{x}^w | \mathbf{x}^o) [l(f(\mathbf{X}), Y)] \\ &= \mathbb{E}_{p_*(\mathbf{x}^o, y)} [l(f(\mathbf{g}(\mathbf{X}^o), \mathbf{X}^o), Y)]. \end{aligned} \quad (3.2)$$

The second risk that assesses the estimation errors of $\mathbf{g}$:

$$R_{01,j}(g_j) := \mathbb{E}_{p_*(\mathbf{x})} [l_{01}(g_j(\mathbf{X}^o), X_j^w)], \forall j \in [F^w]. \quad (3.3)$$

Here, $l_{01}(y, y') := \mathbb{1}_{[y \neq y']}$ represents the 0-1 loss. Finally, the objective function for discrete WFL is defined as a linear combination of these risks:

$$R_{l,\lambda}^{\text{dWFL}}(\mathbf{g}, f) := R_{l,\mathbf{g}}(f) + \lambda \sum_{j \in [F^w]} R_{01,j}(g_j), \quad (3.4)$$

where $\lambda \in \mathbb{R}_+$ is a weighting parameter.

The objective function $R_{l,\lambda}^{\text{dWFL}}$ facilitates the unified treatment of any discrete WFs. This unification arises from representing the error of g_j via the risk $R_{01,j}$, which is aimed to be minimized regardless of the type of WF. In practice, for various types of WFs, WSL methods that learn g_j aim to minimize $R_{01,j}$, by minimizing risks that serve as upper-bound of $R_{01,j}$ or risks whose optimal solutions align with those of $R_{01,j}$ (Cour et al., 2011; Feng et al., 2020; Natarajan et al., 2013; Ishida et al., 2017; Yu et al., 2018). Therefore, such WSL methods can be utilized to learn g_j as part of minimizing $R_{l,\lambda}^{\text{dWFL}}$.

The validity of our formulation is demonstrated by the following theorem¹. Its proof is given in Appendix A.1.

Theorem 3.1. *For any $f \in \mathcal{F}$, $\mathbf{g} \in \mathcal{G}$, and l bounded by $U_l < \infty$, the following inequality holds:*

$$R_l(f) \leq R_{l,\mathbf{g}}(f) + U_l \sum_{j \in [F^w]} R_{01,j}(g_j). \quad (3.5)$$

¹Theorem 3.1 can also be derived from Lemma 4.1, which is introduced later in Section 4.1. However, while Lemma 4.1 is derived for the theoretical analysis in Section 4, Theorem 3.1 is intended to establish the validity of our formalization. We introduce this theorem to enhance the clarity of our discussion.

The RHS of Eq.(3.5) equals R_{l,U_l}^{dWFL} . By scaling l , U_l can be aligned with any λ , and minimizing $R_{l,\lambda}^{\text{dWFL}}$ is expected to yield an f that also minimizes R_l . In other words, by using $R_{l,\lambda}^{\text{dWFL}}$, f is learned to capture the true relationship between $\mathbf{X}$ and Y , despite relying on $\bar{\mathbf{X}}^w$.

This result of Theorem 3.1 is directly applicable to the following two scenarios. The first scenario occurs when test instances contain the exact values of WFs. For instance, practical scenarios arises when WFs are observed during training, but exact values are available during testing. In this scenario, minimizing R_l is essential, but it cannot be computed directly from the training data containing WFs. In contrast, since Theorem 3.1 ensures that minimizing $R_{l,\lambda}^{\text{dWFL}}$ indirectly minimizes R_l , our framework is a valuable approach for overcoming incomplete inputs during training while enhancing performance with complete inputs at test time.

The second scenario involves training data containing a mix of instances with exact values of WFs $\mathbf{X}^w$ and instances with WFs $\bar{\mathbf{X}}^w$. In practical applications, some portions of the data may be observed in detail, yielding exact values, while other portions may be only partially observed. Theorem 3.1 guarantees that minimizing $R_{l,\lambda}^{\text{dWFL}}$ contributes to minimizing R_l . Thus, it enables the simultaneous use of both data types during training. This approach facilitates the development of learning methods that seamlessly integrate both data types into a unified framework.

3.3. Learning Algorithm Class for discrete WFL

In this section, we introduce a class of learning algorithms that uniformly address not only any discrete WFs but also diverse methods within the discrete WFL framework. Building on the formulation proposed in Section 3.2, we define a class that encompasses numerous existing methods in ItR and CFL as follows:

Definition 3.2 (LAC-dWFL). *learning algorithm class for discrete WFL (LAC-dWFL) refers to the set of algorithms in discrete WFL that learn the feature estimation models $\mathbf{g}$ and the label prediction model f using one or a combination of the following three steps:*

- (i) Learning $\mathbf{g}$ by using $\bar{\mathbf{X}}^w$ as weak supervision and minimizing $\sum_{j \in [F^w]} R_{01,j}$, either directly or indirectly.
- (ii) Learning f with $\mathbf{g}$ fixed by minimizing $R_{l,\mathbf{g}}$.
- (iii) Learning $\mathbf{g}$ with f fixed by minimizing $R_{l,\lambda}^{\text{dWFL}}$.

The introduction of LAC-dWFL allows for a unified treatment of a wide range of methods for WFL. Most methods applicable to ItR and CFL fall under the category of *sequential learning*, where steps (i) and (ii) are executed in sequence

(Yoon et al., 2018; Mattei & Frellsen, 2019; Ipsen et al., 2021; Josse et al., 2024; Le Morvan et al., 2021; Sugiyama & Uchida, 2024). Additionally, methods that represent g and f as neural networks and combine them (Le Morvan et al., 2020a; Ipsen et al., 2022) can be regarded as *iterative learning*, where steps (ii) and (iii) are executed repeatedly, when these components are alternately optimized. Such methods are thus encompassed within LAC-dWFL.

4. Theoretical analysis

This section presents a theoretical analysis of the unified learning algorithm class, LAC-dWFL. Through this analysis, we elucidate the common properties of LAC-dWFL and establish a foundation for the theoretical exploration of various methods encompassed within this class. To achieve this, it is necessary to elucidate the properties of steps (i), (ii), and (iii) within LAC-dWFL. As established in Section 3.3, step (i) involves WSL methods, and its properties can therefore be analyzed using existing WSL theories. Consequently, our theoretical focus is directed toward steps (ii) and (iii). In Section 4.1, we derive a fundamental inequality to analyze steps (ii) and (iii). In Section 4.2, we establish an error bound for f learned via step (ii), given any g . In Section 4.3, we establish an error bound for g learned via step (iii), given any f .

4.1. Deriving an Analytical Tool

Our objective is to examine how the learning of f in step (ii) and g in step (iii) depend on the performance of g and f , respectively. To achieve this, the error bounds for f and g must be expressed in terms of $R_{0l,j}$ for any $j \in [F^w]$ and R_l , respectively. To achieve this requirement, we present the following lemma, with its proof provided in Appendix A.2.

Lemma 4.1. *For any measurable l bounded by $U_l < \infty$, $f \in \mathcal{F}$ and $g \in \mathcal{G}$, the following holds:*

$$|R_l(f) - R_{l,g}(f)| \leq (\sqrt{R_l(f)} + \sqrt{R_{l,g}(f)}) \left(2U_l \sum_{j \in [F^w]} R_{0l,j}(g_j) \right)^{\frac{1}{2}}. \quad (4.6)$$

Equation (4.6) shows that $|R_l(f) - R_{l,g}(f)| = 0$ is achieved when either $\sum_{j \in [F^w]} R_{0l,j}(g_j) = 0$ or $\sqrt{R_l(f)} + \sqrt{R_{l,g}(f)} = 0$. Although intuitive, this inequality serves a critical role in deriving subsequent error bounds. Lemma 4.1 enables the analysis of how f and g influence each other's learning processes.

4.2. Analysis of Learning Label Prediction Model f

In this section, we conduct a theoretical analysis of the learning process of f under LAC-dWFL's step (ii), where g

remains fixed. Since step (ii) involves learning f based on the output of g rather than X^w , the relationship between g and f cannot be analyzed using ordinary supervised learning frameworks (Mohri et al., 2018). We derive an error bound for f that captures how the errors of g influence the learning process of f , enabling an analysis of the step (ii).

To facilitate the analysis, we introduce the following definitions. Given $n \in \mathbb{N}_+$ training samples, we define the *ordinary dataset*, $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and the *weak dataset*, $\bar{S} = \{(\bar{\mathbf{x}}_i^w, \mathbf{x}_i^o, y_i)\}_{i=1}^n$. Here, $\mathbf{x}_i = (\mathbf{x}_i^o, \mathbf{x}_i^w)$, y_i and $\bar{\mathbf{x}}_i^w$ represent the realizations of $\mathbf{X} = (\mathbf{X}^w, \mathbf{X}^o)$, Y and $\bar{\mathbf{X}}^w$, respectively. Also, the i -th samples in S and $\bar{S}$ correspond to the same instance, for any $i \in [n]$. We assume that $\{(\mathbf{x}_i, y_i, \bar{\mathbf{x}}_i^w)\}_{i=1}^n$ are independently drawn from $p_*(\mathbf{x}, y) \bar{p}_*(\bar{\mathbf{x}}^w | \mathbf{x}, y)$. Let $\hat{R}_l$ and $\hat{R}_{l,g}$ denote the empirical risks calculated by the sample average over S and $\bar{S}$, respectively. For any $g \in \mathcal{G}$, the empirical risk minimizer obtained from LAC-dWFL's step (ii) is defined as follows:

$$f_{g,\bar{S}} := \arg \min_{f \in \mathcal{F}} \hat{R}_{l,g}(f).$$

Using Lemma 4.1, we establish the error bound for $f_{g,\bar{S}}$ learned via LAC-dWFL's step (ii) in the following theorem. The proof is provided in Appendix A.3.

Theorem 4.2. *Let S and $\bar{S}$ be the ordinary dataset and weak dataset, respectively, each containing n samples. For any measurable $g \in \mathcal{G}$, L_l -Lipschitz continuous l bounded by $U_l < \infty$ and $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$:*

$$\begin{aligned} R_{l,g}(f_{g,\bar{S}}) - R_l(f_{\mathcal{F}}) &\leq \\ &4 \left(L_l \mathfrak{R}_n^*(\mathcal{F}) + L_l \mathfrak{R}_n^g(\mathcal{F}) + U_l \sqrt{\frac{\log(4/\delta)}{2n}} \right) \\ &+ \left\{ 2 \left(R_l(f_{\mathcal{F}}) + 4L_l \mathfrak{R}_n^*(\mathcal{F}) + 2U_l \sqrt{\frac{\log(4/\delta)}{2n}} \right)^{\frac{1}{2}} \right. \\ &\quad \left. + \left(2U_l \sum_{j \in [F^w]} R_{0l,j}(g_j) \right)^{\frac{1}{2}} \right\} \\ &\times \left(2U_l \sum_{j \in [F^w]} R_{0l,j}(g_j) \right)^{\frac{1}{2}}. \end{aligned} \quad (4.7)$$

Here, $f_{\mathcal{F}} := \arg \min_{f \in \mathcal{F}} R_l(f)$ represents the true risk minimizer in ordinary supervised learning.

The terms $\mathfrak{R}_n^*(\mathcal{F})$ and $\mathfrak{R}_n^g(\mathcal{F})$ represent the Rademacher complexities of the hypothesis class $\mathcal{F}$ under the distributions $p_*(\mathbf{x})$ and $p_*(\mathbf{x}^o)q_g(\mathbf{x}^w | \mathbf{x}^o)$, respectively, and measure the complexity of $\mathcal{F}$. Equation (4.7) reveals that the errors of g combine with the result of virtual ordinary supervised learning $(R_l(f_{\mathcal{F}}) + \dots)^{1/2}$ to affect the error bound for f in WFL.

The first contribution of Theorem 4.2 is its ability to reveal the following property concerning the learning of f under LAC-dWFL's step (ii). Theorem 4.2 demonstrates how the

convergence rate of the error bound for f with respect to n depends on the estimation errors of g . Given the established result that the order of the Rademacher complexity's upper bound is $\mathcal{O}_p(1/n^{1/2})$ for kernel ridge regression and multilayer perceptrons (Mohri et al., 2018; Neyshabur et al., 2015), we assume the orders of Rademacher complexities about $\mathcal{F}$ or $\mathcal{G}$ are $\mathcal{O}_p(1/n^{1/2})$. Additionally, assume that $\mathcal{F}$ is sufficiently expressive and that $R_l(f_{\mathcal{F}}) = 0$. Under these assumptions, and with g fixed, the first and second terms in the error bound have orders of $\mathcal{O}_p(1/n^{1/2})$ and $\mathcal{O}_p(1/n^{1/4})$, respectively. Therefore, the second term, which decreases more slowly, becomes dominant when $\sum_{j \in [F^w]} R_{01,j}$ is large. This result suggests that when learning g to improve the generalization performance of f , maximizing the estimation accuracy of g is sufficient, rather than tailoring the outputs of g to be specifically suited for f . Furthermore, when WFs contain ambiguous information such as CFs, their values $\bar{X}^w$ can serve as inputs to f . The choice between $\bar{X}^w$ and the learned g is a critical decision, empirically validated in the context of CFL (Sugiyama & Uchida, 2024). Theorem 4.2 concludes that, if the learned g estimates X^w more accurately than $\bar{X}^w$, using g contributes more significantly to reducing the error bound for f compared to directly using $\bar{X}^w$.

The second contribution of Theorem 4.2 is its ability to integrate our WFL theory with the theories of WSL methods used to learn g under LAC-dWFL's step (i). This is because, when g_j is learned using certain WSL methods, their theoretical frameworks enable the derivation of error bounds for $R_{01,j}(g_j)$ in Eq. (4.7) (Cour et al., 2011; Feng et al., 2020; Xu et al., 2021; Natarajan et al., 2013; Ishida et al., 2017; Yu et al., 2018). For example, when $\bar{X}_j^w$ is a CF, applying the CLL method by Ishida et al. (Ishida et al., 2017) for learning g_j , and assuming that $\mathcal{G}_j$ is sufficiently large to satisfy $\min_{g_j \in \mathcal{G}_j} R_{01,j}(g_j) = 0$, the following holds for any L_l -Lipschitz continuous l and any $\delta \in (0, 1)$ with probability at least $1 - \delta$:

$$R_{01,j}(\hat{g}_j) \leq 4|\mathcal{X}_j^w|(|\mathcal{X}_j^w| - 1)L_l\mathfrak{R}_n^*(\mathcal{G}_j) + (|\mathcal{X}_j^w| - 1)\sqrt{\frac{8\log(2/\delta)}{n}}. \quad (4.8)$$

The combination of such error bound with Eq. (4.7) enables a unified generalization error analysis for the sequential learning of g and f under LAC-dWFL's steps (i) and (ii), elucidating the following three aspects: Firstly, the combination enables the analysis the influence of WFs' properties on the learning of f . For instance, since Eq. (4.8) depends on $|\mathcal{X}_j^w|$, combining it with our bound allows for analyzing how the number of possible values $|\mathcal{X}_j^w|$ influence f 's learning. Secondly, this combination elucidates the impact of whether g is learned or not on the learning of f . Applying Eq. (4.8) to Eq. (4.7) demonstrates that the order of the error bound for f is $\mathcal{O}_p(1/n^{1/2})$. Thus, when a constant

value or $\bar{X}^w$ is used as $g(X^o)$, the order of the error bound remains $\mathcal{O}_p(1/n^{1/4})$, whereas learning g improves this order to $\mathcal{O}_p(1/n^{1/2})$. In addition, Theorem 4.2 theoretically connects the error bounds of f and g , thereby elucidating the conditions under which sequential learning achieves consistency, as following theorem. The proof is shown in Appendix A.4.

Theorem 4.3. *Assume the existence of true deterministic functions $g_j^* : \mathcal{X}^o \rightarrow \mathcal{X}_j^w$ for all $j \in [F^w]$, such that $(g_1^*, \dots, g_{F^w}^*) \in \mathcal{G}$, and $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ such that $f^* \in \mathcal{F}$. Additionally, suppose l bounded by $U_l < \infty$ is L_l -Lipschitz continuous, and $\mathfrak{R}_n^*(\mathcal{F})$ and $\mathfrak{R}_n^g(\mathcal{F})$ asymptotically approach 0 as $n \rightarrow \infty$. If, for all $j \in [F^w]$, the number of samples available for learning g_j tends to infinity as $n \rightarrow \infty$, and a consistent method is employed to learn g_j , then sequential learning achieves consistency (i.e., as $n \rightarrow \infty$, $R_{l,g}(f_{g,S}) \rightarrow R_l(f_{\mathcal{F}})$).*

Thus, under the conditions stated in Theorem 4.3, an asymptotically optimal pair of g and f can be obtained through sequential learning alone. In contrast, in practical finite-sample scenarios, iterative learning involving LAC-dWFL's steps (ii) and (iii) may be necessary (Le Morvan et al., 2020b; 2021). In Section 4.3, we examines the step (iii) for a comprehensive understanding of LAC-dWFL.

4.3. Analysis of Learning Feature Estimation Models g

This section provides a theoretical analysis of the learning of g with f fixed in LAC-dWFL's step (iii). We aims to reveal how the learning of g using $R_{l,\lambda}$, with f fixed, is influenced by f .

The learning of g using $R_{l,\lambda}^{\text{dWFL}}$ in Eq. (3.4) involves the simultaneous minimization of two risks, which makes it challenging to conduct generalization error analysis for ERM directly. In contrast, in ordinary supervised learning, the simultaneous minimization of an expected risk and a regularization term is formulated as structural risk minimization (SRM), which has established methods for generalization error analysis (Mohri et al., 2018). SRM selects a hypothesis from a restricted hypothesis class in which the regularization term is below a certain threshold and derives error bounds for the selected hypothesis. Focusing on the influence of f on the learning of g , we treat $R_{l,f}$ as the expected risk for which bounds are derived and $\sum_{j \in [F^w]} R_{01,j}$ as the regularization term, and apply SRM theory.

To apply SRM theory, for any $j \in [F^w]$, we introduce the following definitions. Let l_j denote the loss function for g_j computed using $\bar{X}_j^w$. Define the datasets $\bar{S}_j := \{(\bar{x}_{ij}^w, \mathbf{x}_i^o)\}_{i=1}^n$. Let $\bar{R}_{l_j}(g_j) := \mathbb{E}_{p_*(\mathbf{x}, y) \bar{p}_*(\bar{x}_j^w | \mathbf{x}, y)}[l_j(g_j(\mathbf{X}^o), \bar{X}_j^w)]$ represents the expected risk of g_j computed using $\bar{X}_j^w$. Additionally, $\hat{\bar{R}}_{l_j}$ de-

notes the empirical risk, which approximates $\bar{R}_{l_j}$ by taking the sample average over $\bar{S}_j$. We assume that $\bar{R}_{l_j}$ satisfies either $R_{01,j}(g_j) = \bar{R}_{l_j}(g_j)$ or $R_{01,j}(g_j) \leq \bar{R}_{l_j}(g_j)$ for any g_j , or that the optimal solutions of $\bar{R}_{l_j}$ coincide with those of $R_{01,j}$. For any $\mathbf{r} = (r_1, \dots, r_{F^w}) \in \mathbb{R}_+^{F^w}$, define the following hypothesis class:

$$\mathcal{G}(\mathbf{r}, \bar{S}) := \mathcal{G}_1(r_1, \bar{S}_1) \times \dots \times \mathcal{G}_{F^w}(r_{F^w}, \bar{S}_{F^w}),$$

where, $\mathcal{G}_j(r_j, \bar{S}_j) := \{g_j | g_j \in \mathcal{G}_j \wedge \hat{R}_{l_j}(g_j) \leq r_j\}$, $\forall j \in [F^w]$. To explicitly indicate that $R_{l,g}(f)$ is part of the objective function of $\mathbf{g}$, we denote it as $R_{l,f}(\mathbf{g}) (\equiv R_{l,g}(f))$ and the empirical risk of $R_{l,f}(\mathbf{g})$ is defined as $\hat{R}_{l,f}(\mathbf{g}) (\equiv \hat{R}_{l,g}(f))$. By performing the learning of $\mathbf{g}$ in LAC-dWFL's step (iii) as outlined below, the analysis of minimizing $R_{l,f}$ while reducing $\sum_{j \in [F^w]} R_{01,j}$ becomes feasible:

$$\mathbf{g}_{f,\bar{S}}^{(r)} := \arg \min_{\mathbf{g} \in \mathcal{G}(\mathbf{r}, \bar{S})} \hat{R}_{l,f}(\mathbf{g}). \quad (4.9)$$

Based on the above definition, the assumptions and Lemma 4.1, the error bound for $\mathbf{g}_{f,\bar{S}}^{(r)}$ is presented in the following theorem. The proof is provided in Appendix A.5.

Theorem 4.4. *Suppose S and $\bar{S}$ represent an ordinary dataset and a weak dataset of n samples, respectively. Then, for any measurable $f \in \mathcal{F}$, l bounded by $U_l < \infty$ and $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$:*

$$\begin{aligned} R_{l,f}(\mathbf{g}_{f,\bar{S}}^{(r)}) - R_l(f) &\leq \\ &\left(4\mathfrak{R}_n^*(\tilde{\mathcal{G}}_{l,f}(\mathbf{r}, \bar{S})) + 2U_l \sqrt{\frac{\log(2/\delta)}{2n}} \right) \\ &+ \left\{ 2\sqrt{R_l(f)} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_{\bar{S},j}^{(r_j)}) \right)^{\frac{1}{2}} \right\} \\ &\times \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_{\bar{S},j}^{(r_j)}) \right)^{\frac{1}{2}} \end{aligned} \quad (4.10)$$

Here, $\tilde{\mathcal{G}}_{l,f}(\mathbf{r}, \bar{S}) := \{(\mathbf{x}^o, y) \mapsto l(f(\mathbf{g}(\mathbf{x}^o), \mathbf{x}^o), y) : \mathbf{g} \in \mathcal{G}(\mathbf{r}, \bar{S})\}$.

The term $R_{01,j}(g_{\bar{S},j}^{(r_j)})$ in Eq. (4.10) can be further upper-bounded by defining $\mathcal{G}_j(r_j, \bar{S}_j)$ as the set of empirical risk minimizers obtained via weakly supervised learning using $\bar{S}_j$. For instance, when $\bar{X}_j^w$ is a CF, and every hypothesis in $\mathcal{G}_j(r_j, \bar{S}_j)$ is an empirical risk minimizer obtained using the method of Ishida et al. (Ishida et al., 2017) and $\mathcal{G}_j$ is sufficiently large to satisfy $\min_{g_j \in \mathcal{G}_j} R_{01,j}(g_j) = 0$, then $R_{01,j}(g_{\bar{S},j}^{(r_j)})$ can be upper-bounded by Eq. (4.8). Therefore, when we define $\mathcal{G}_j(r_j, \bar{S}_j)$ as the class of empirical risk minimizers obtained by such a method with guaranteed consistency, and assume that the order of $\mathfrak{R}_n^*(\mathcal{G}_j)$ is $\mathcal{O}_p(1/n^{1/2})$, the order of the upper bound of $R_{01,j}(g_{\bar{S},j}^{(r_j)})$ is

also $\mathcal{O}_p(1/n^{1/2})$. Furthermore, assuming that $\mathfrak{R}_n^*(\mathcal{G}_j) \rightarrow 0$ as $n \rightarrow \infty$, it follows that $R_{01,j}(g_{\bar{S},j}^{(r_j)}) \rightarrow 0$ as $n \rightarrow \infty$.

In the following discussion, we assume that $R_{01,j}(g_{\bar{S},j}^{(r_j)})$ can be upper-bounded by a probability inequality of a form similar to that of Eq. (4.8).

Theorem 4.4 elucidates the influence of f 's prediction error and characteristics on the error bound for learning $\mathbf{g}$ via LAC-dWFL's step (iii), as well as the convergence of $\mathbf{g}$'s learning. First, Theorem 4.4 demonstrates that the convergence rate of the upper bound of $R_{l,f}(\mathbf{g}_{f,\bar{S}}^{(r)})$ with respect to n significantly depends on the expected risk $R_l(f)$. Specifically, assuming the orders of the Rademacher complexities in Eq. (4.10) are $\mathcal{O}_p(1/n^{1/2})$, the orders of the first and second terms on the RHS of Eq. (4.10) are $\mathcal{O}_p(1/n^{1/2})$ and $\mathcal{O}_p(1/n^{1/4})$, respectively. Consequently, the slower-decreasing second term becomes dominant when $R_l(f)$ is large. Although $R_l(f)$ cannot be directly minimized in WFL, it has been shown that minimizing $R_{l,\lambda}^{\text{dWFL}}(f, \cdot)$ helps reduce $R_l(f)$ (Theorem 3.1). Therefore $R_l(f)$ is expected to decrease through LAC-dWFL's step (ii). Additionally, if l is L_l -Lipschitz continuous and f is L_f -Lipschitz continuous, then $\mathfrak{R}_n^*(\tilde{\mathcal{G}}_{l,f}(\mathbf{r}, \bar{S})) \leq L_l L_f \mathfrak{R}_n^*(\mathcal{G}(\mathbf{r}, \bar{S}))$ holds (Mohri et al., 2018). This implies that the learning efficiency of $\mathbf{g}$ improves when f exhibits smoother variation with respect to its input. Furthermore, combining Theorem 4.4 with Theorem 4.2 reveals the conditions under which iterative learning in LAC-dWFL achieves consistency. The proof is shown in Appendix A.6.

Theorem 4.5. *In addition to the conditions stated in Theorem 4.3, assume f obtained via LAC-dWFL's step (ii) is Lipschitz continuous, and the Rademacher complexities about $\mathbf{g}$ asymptotically converge to 0 as n increases. Furthermore, for any $j \in [F^w]$, define $\mathcal{G}_j(r_j, \bar{S}_j)$ as the set of empirical risk minimizers obtained by methods that use $\bar{S}_j$ and are guaranteed to achieve consistency. Then iterative learning achieves consistency.*

5. Experiments

In Section 4 we revealed two key properties of LAC-dWFL: (1) the mutual influence between the feature estimation models $\mathbf{g}$ and the label prediction model f during their respective learning processes, and (2) the relationship between the generalization error and the number of training samples n in WFL. Furthermore, the theoretical analysis demonstrates that sequential learning alone suffices for WFL. To validate the critical aspect of WFL, namely the impact of the estimation error of $\mathbf{g}$ on the learning of f (Theorem 4.2), we evaluate how varying the estimation errors of $\mathbf{g}$ affects f 's learning performance.

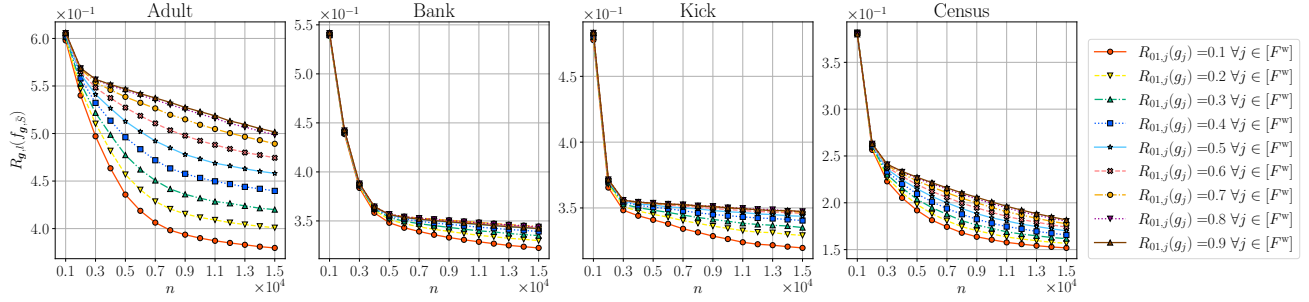


Figure 5.1. The relationship between the number of training samples n , $R_{l,g}(f_{g,\bar{S}})$, and various estimation errors of g .

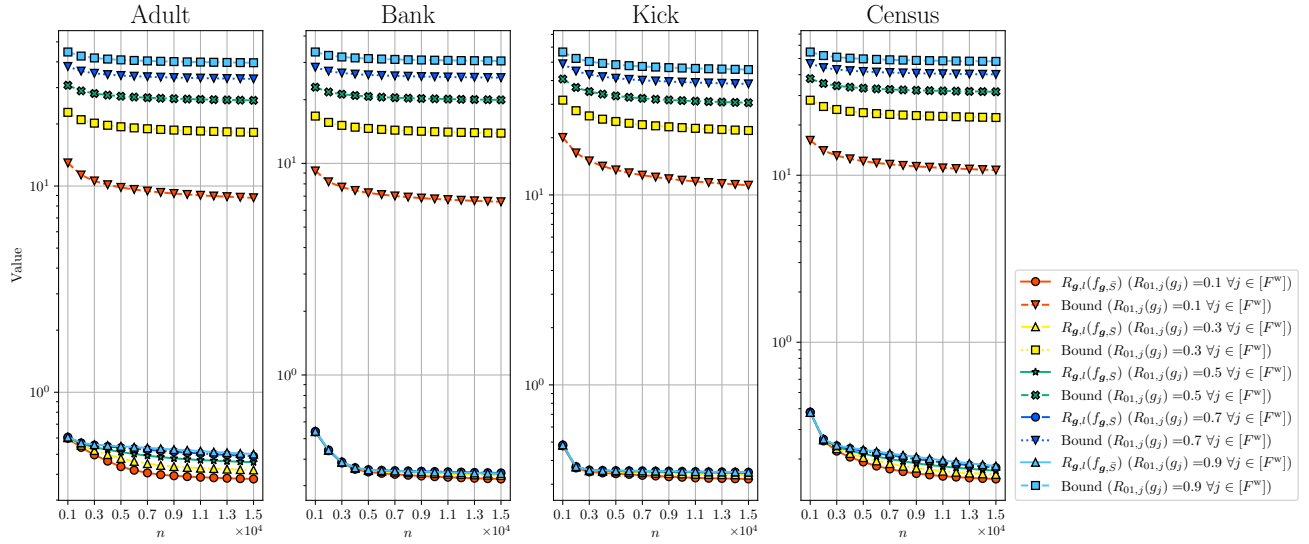


Figure 5.2. A comparison between $R_{l,g}(f_{g,\bar{S}})$ and the error bound derived in Theorem 4.2, for various estimation errors of g .

5.1. Experimental Settings

We used four real-world datasets: *Adult* (Becker & Kohavi, 1996), *Bank Marketing* (Moro & Cortez, 2014), *kick* (Vanschoren et al., 2013), and *Census-Income* (KDD) (cen, 2000; Dua & Graff, 2017). We refer to them as *Adult*, *Bank*, *Kick*, and *Census*, respectively. Details of these datasets are summarized in Appendix B.1. For each dataset, 50% of the samples were reserved as test data to estimate the generalization error. In this experiment, we focused on a representative case of WFs, where all categorical features are treated as CFs (Sugiyama & Uchida, 2024). Both the feature estimation models g and the label prediction model f were implemented using two-layer perceptrons with hidden layers of width 500 and ReLU as an activation function. Logistic loss was used as l . The Rademacher complexity, required for calculating the error bounds, was estimated using the method proposed by Neyshabur et al. (Neyshabur et al., 2015). Details of the experimen-

tal settings are summarized in Appendix B.2. The following results are the average of 5 trials. The experimental scripts used in this paper are available at the following URL: https://github.com/KOHsEMP/discrete_WFL

5.2. Impact of g on f 's learning

This section investigates how the estimation errors of feature estimation models g affects the learning of the label prediction model f . To perform this investigation, precise control over the estimation errors of g is required. However, achieving such fine-grained control through WSL methods is challenging. Importantly, for this experiment, the method used to obtain g is less relevant than the influence of its estimation errors on f . Thus, synthetic estimation functions for g are employed, which randomly misestimate with controlled error rates, enabling systematic examination of the impact of g 's estimation errors on f 's learning.

We vary the estimation error of g from 10% to 90% and train f under these settings. Figure 5.1 shows the relationship between n and $R_{l,g}(f_{g,\bar{s}})$. The results confirm that, as shown in Theorem 4.2, lower estimation errors of g lead to a higher reduction rate of $R_{l,g}(f_{g,\bar{s}})$ as n increases.

Additionally, Figure 5.2 compares the generalization error of f , shown in Figure 5.1, with the error bound in Theorem 4.2. Figure 5.2 shows that the decrease in $R_{l,g}(f_{g,\bar{s}})$ and its bound with increasing n exhibits a similar trend, with the reduction becoming more significant as the estimation error of g . Therefore, Theorem 4.2 effectively captures a fundamental characteristic of WFL, specifically the influence of g on the actual learning of f , which is consistent across various scenarios. The discrepancy between $R_{l,g}(f_{g,\bar{s}})$ and our bound in Figure 5.2 can be attributed to the fact that our bound does not account for the feature importance of WFs in predicting Y . This suggests a new research direction for deriving error bounds that incorporate the feature importance of WFs. Our results are considered to provide a critical foundation for such an approach.

6. Conclusion

This presented a unified formalization and theoretical analysis of discrete WFL. First, we proposed a formulation of WFL capable of handling arbitrary discrete WFs. We validated this formulation by demonstrating that the introduced objective function aids in learning a label prediction model f that captures the true input-output relationship. Within this framework, we performed a generalization error analysis for LAC-dWFL, a generalized learning algorithm class designed to learn both feature estimation models g and f . This analysis revealed the detailed influence of the estimation errors of g and f on the error bounds of f and g , respectively. Additionally, we identified theoretical conditions under which consistency can be achieved for the sequential and iterative learning approaches in LAC-dWFL. Finally, numerical experiments on real-world datasets verified that our theoretical results align with observed learning behavior. This study provides comprehensive theoretical insights into various problem settings, such as ItR and CF, involving discrete WFs.

Impact Statement

Our paper presents a formalization and theoretical analysis of discrete WFL that accommodates arbitrary discrete WFs. Understanding the impact of low-quality input features on the training of predictive models is crucial for the development and deployment of safe machine learning systems. This is because degraded input feature quality can potentially lead to predictive models that are vulnerable to adversarial attacks or that propagate socially undesirable

biases. The theoretical results presented in this work are expected to provide a foundational perspective for discussions concerning the interplay between input feature quality and the safety of machine learning.

Acknowledgment

This work was supported in part by the Japan Society for the Promotion of Science through Grants-in-Aid for Scientific Research (C) (23K11111).

References

- Census-Income (KDD). UCI Machine Learning Repository, 2000. DOI: <https://doi.org/10.24432/C5N30T>.
- Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996. DOI: <https://doi.org/10.24432/C5XW20>.
- Bertsimas, D., Delarue, A., and Pauphilet, J. Prediction with missing data. *stat*, 1050:7, 2021.
- Chapelle, O., Scholkopf, B., and Zien, A. Semi-supervised learning. 2006. *Cambridge, Massachusettes: The MIT Press View Article*, 2, 2006.
- Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *J. Mach. Learn. Res.*, 12(null):1501–1536, July 2011. ISSN 1532-4435.
- Dua, D. and Graff, C. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 213–220, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401920.
- Feng, L., Lv, J., Han, B., Xu, M., Niu, G., Geng, X., An, B., and Sugiyama, M. Provably consistent partial-label learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- Ipsen, N., Mattei, P.-A., and Frellsen, J. How to deal with missing data in supervised deep learning? In *Proceedings of 2022 International Conference on Learning Representations*, 2022. URL <https://iclr.cc/Conferences/2022>. 10th International Conference on Learning Representations, ICLR 2022 ; Conference date: 25-04-2022 Through 29-04-2022.

- Ipsen, N. B., Mattei, P.-A., and Frellsen, J. not- $\{miwae\}$: Deep generative modelling with missing not at random data. In *International Conference on Learning Representations*, 2021.
- Ishida, T., Niu, G., Hu, W., and Sugiyama, M. Learning from complementary labels. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- Ishida, T., Niu, G., Menon, A., and Sugiyama, M. Complementary-label learning for arbitrary losses and models. In *International Conference on Machine Learning*, pp. 2971–2980. PMLR, 2019.
- Josse, J., Chen, J. M., Prost, N., Varoquaux, G., and Scornet, E. On the consistency of supervised learning with missing values. *Statistical Papers*, 65(9):5447–5479, 2024.
- Kahn, M. Diabetes. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5T59G>.
- Katsura, Y. and Uchida, M. Bridging ordinary-label learning and complementary-label learning. In Pan, S. J. and Sugiyama, M. (eds.), *Proceedings of The 12th Asian Conference on Machine Learning*, volume 129 of *Proceedings of Machine Learning Research*, pp. 161–176. PMLR, 18–20 Nov 2020.
- Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.
- Le Morvan, M., Josse, J., Moreau, T., Scornet, E., and Varoquaux, G. Neumiss networks: differentiable programming for supervised learning with missing values. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5980–5990. Curran Associates, Inc., 2020a.
- Le Morvan, M., Prost, N., Josse, J., Scornet, E., and Varoquaux, G. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3165–3174. PMLR, 26–28 Aug 2020b.
- Le Morvan, M., Josse, J., Scornet, E., and Varoquaux, G. What’s a good imputation to predict with missing values? In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11530–11540. Curran Associates, Inc., 2021.
- Lin, W.-I. and Lin, H.-T. Reduction from complementary-label learning to probability estimates. In Kashima, H., Ide, T., and Peng, W.-C. (eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 469–481, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-33377-4.
- Mattei, P.-A. and Frellsen, J. MIWAE: Deep generative modelling and imputation of incomplete data sets. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4413–4423. PMLR, 09–15 Jun 2019.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018. ISBN 0262039400.
- Moro, S., R. P. and Cortez, P. Bank Marketing. UCI Machine Learning Repository, 2014. DOI: <https://doi.org/10.24432/C5K306>.
- Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, Paris, France, 03–06 Jul 2015. PMLR.
- Ruan, J., Zheng, Q., Zhao, R., and Dong, B. Biased complementary-label learning without true labels. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2616–2627, 2024. doi: 10.1109/TNNLS.2022.3190528.
- Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- Sugiyama, K. and Uchida, M. Learning from complementary features. *arXiv preprint arXiv:2408.14788*, 2024.
- Sugiyama, K. and Uchida, M. Unified analysis of continuous weak features learning with applications to learning from missing data. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.
- Tian, Y., Yu, X., and Fu, S. Partial label learning: Taxonomy, analysis and outlook. *Neural Networks*, 161:708–734, 2023. ISSN 0893-6080. doi: <https://doi.org/10.1016/j.neunet.2023.02.019>.

- Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. Openml: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198.
- Xu, N., Qiao, C., Geng, X., and Zhang, M.-L. Instance-dependent partial label learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27119–27130. Curran Associates, Inc., 2021.
- Yeh, I.-C. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: <https://doi.org/10.24432/C55S3H>.
- Yoon, J., Jordon, J., and van der Schaar, M. GAIN: Missing data imputation using generative adversarial nets. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5689–5698. PMLR, 10–15 Jul 2018.
- Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- Zaffran, M., Dieuleveut, A., Josse, J., and Romano, Y. Conformal prediction with missing values. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40578–40604. PMLR, 23–29 Jul 2023.

A. Proofs

A.1. Proof of Theorem 3.1

Proof of Theorem 3.1. For any $f \in \mathcal{F}$, $g \in \mathcal{G}$, and l bounded by $U_l < \infty$, the following inequality holds:

$$\begin{aligned}
 & R_l(f) - R_{l,g}(f) \\
 &= \mathbb{E}_{p_*(\mathbf{x}, y)}[l(f(\mathbf{X}), Y)] - \mathbb{E}_{p_*(\mathbf{x}^\circ, y)q_g(\mathbf{x}^\mathbf{w}|\mathbf{x}^\circ)}[l(f(\mathbf{X}), Y)] \\
 &= \mathbb{E}_{p_*(\mathbf{x}^\circ, y)} \left[\sum_{\mathbf{x}^\mathbf{w} \in \mathcal{X}^\mathbf{w}} l(f(\mathbf{X}^\circ, \mathbf{x}^\mathbf{w}), Y) (p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) - q_g(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ)) \right] \\
 &= \mathbb{E}_{p_*(\mathbf{x}^\circ, y)} \left[\sum_{\mathbf{x}^\mathbf{w} \in \mathcal{X}^\mathbf{w}} l(f(\mathbf{X}^\circ, \mathbf{x}^\mathbf{w}), Y) (p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) - \mathbb{1}_{[\mathbf{x}^\mathbf{w}=g(\mathbf{X}^\circ)])} \right] \\
 &\leq \mathbb{E}_{p_*(\mathbf{x}^\circ, y)} \left[\sum_{\mathbf{x}^\mathbf{w} \in \mathcal{X}^\mathbf{w}} l(f(\mathbf{X}^\circ, \mathbf{x}^\mathbf{w}), Y) (p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) - \mathbb{1}_{[\mathbf{x}^\mathbf{w}=g(\mathbf{X}^\circ)])} p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y)) \right] \\
 &= \mathbb{E}_{p_*(\mathbf{x}^\circ, y)} \left[\sum_{\mathbf{x}^\mathbf{w} \in \mathcal{X}^\mathbf{w}} l(f(\mathbf{X}^\circ, \mathbf{x}^\mathbf{w}), Y) p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) (1 - \mathbb{1}_{[\mathbf{x}^\mathbf{w}=g(\mathbf{X}^\circ)])} \right] \\
 &\leq \mathbb{E}_{p_*(\mathbf{x}, y)} \left[l(f(\mathbf{X}^\circ, \mathbf{X}^\mathbf{w}), Y) \sum_{j \in [F^\mathbf{w}]} (1 - \mathbb{1}_{[X_j^\mathbf{w}=g_j(\mathbf{X}^\circ)])} \right] \\
 &= \mathbb{E}_{p_*(\mathbf{x}, y)} \left[l(f(\mathbf{X}^\circ, \mathbf{X}^\mathbf{w}), Y) \sum_{j \in [F^\mathbf{w}]} l_{01}(g_j(\mathbf{X}^\circ), X_j^\mathbf{w}) \right] \\
 &\leq U_l \mathbb{E}_{p_*(\mathbf{x}, y)} \left[\sum_{j \in [F^\mathbf{w}]} l_{01}(g_j(\mathbf{X}^\circ), X_j^\mathbf{w}) \right].
 \end{aligned}$$

The second inequality arises from the decomposition of the 0-1 loss. This decomposition is intended to derive the risk for each g_j . The final inequality uses the assumption that the maximum value of the loss function l is U_l . Therefore, Theorem 3.1 is proved. $\square$

A.2. Proof of Lemma 4.1

Proof of Lemma 4.1. The LHS of Eq. (4.6) can be rewritten as follows:

$$\begin{aligned}
 & |R_l(f) - R_{l,g}(f)| \\
 &= |\mathbb{E}_{p_*(\mathbf{x}, y)}[l(f(\mathbf{X}), Y)] - \mathbb{E}_{p_*(\mathbf{x}^\circ, y)q_g(\mathbf{x}^\mathbf{w}|\mathbf{x}^\circ)}[l(f(\mathbf{X}), Y)]| \\
 &= \left| \mathbb{E}_{p_*(\mathbf{x}^\circ, y)} \left[\sum_{\mathbf{x}^\mathbf{w} \in \mathcal{X}^\mathbf{w}} l(f(\mathbf{x}^\mathbf{w}, \mathbf{X}^\circ), Y) \{p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) - q_g(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ)\} \right] \right| \\
 &= \left| \mathbb{E}_{p_*(\mathbf{x}^\circ, y)} \left[\sum_{\mathbf{x}^\mathbf{w} \in \mathcal{X}^\mathbf{w}} l(f(\mathbf{x}^\mathbf{w}, \mathbf{X}^\circ), Y) \{p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) - \mathbb{1}_{[\mathbf{x}^\mathbf{w}=g(\mathbf{X}^\circ)]}\} \right] \right| \\
 &= \left| \mathbb{E}_{p_*(\mathbf{x}^\circ, y)} \left[\sum_{\mathbf{x}^\mathbf{w} \in \mathcal{X}^\mathbf{w}} \left\{ l(f(\mathbf{x}^\mathbf{w}, \mathbf{X}^\circ), Y) p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) - l(f(\mathbf{x}^\mathbf{w}, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^\mathbf{w}=g(\mathbf{X}^\circ)]} \right\} \right] \right| \\
 &= \left| \mathbb{E}_{p_*(\mathbf{x}^\circ, y)} \left[\sum_{\mathbf{x}^\mathbf{w} \in \mathcal{X}^\mathbf{w}} \left\{ l(f(\mathbf{x}^\mathbf{w}, \mathbf{X}^\circ), Y) p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) - l(f(\mathbf{x}^\mathbf{w}, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^\mathbf{w}=g(\mathbf{X}^\circ)]} p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) \right. \right. \right. \\
 &\quad \left. \left. \left. + l(f(\mathbf{x}^\mathbf{w}, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^\mathbf{w}=g(\mathbf{X}^\circ)]} p_*(\mathbf{x}^\mathbf{w}|\mathbf{X}^\circ, Y) - l(f(\mathbf{x}^\mathbf{w}, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^\mathbf{w}=g(\mathbf{X}^\circ)]} \right\} \right] \right|
 \end{aligned}$$

$$\begin{aligned}
 &\leq \mathbb{E}_{p_*(\mathbf{x}^\circ, Y)} \left[\underbrace{\sum_{\mathbf{x}^w \in \mathcal{X}^w} \left| l(f(\mathbf{x}^w, \mathbf{X}^\circ), Y) p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y) - l(f(\mathbf{x}^w, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{X}^\circ)]} p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y) \right|}_{(a1)} \right. \\
 &\quad \left. + \underbrace{\sum_{\mathbf{x}^w \in \mathcal{X}^w} \left| l(f(\mathbf{x}^w, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{X}^\circ)]} p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y) - l(f(\mathbf{x}^w, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{X}^\circ)]} \right|}_{(a2)} \right]. \quad (A.11)
 \end{aligned}$$

The term (a1) in Eq. (A.11) can be expressed as:

$$\begin{aligned}
 (a1) &= \sum_{\mathbf{x}^w \in \mathcal{X}^w} l(f(\mathbf{x}^w, \mathbf{X}^\circ), Y) p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y) (1 - \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{X}^\circ)]}) \\
 &= \mathbb{E}_{p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y)} \left[l(f(\mathbf{X}^w, \mathbf{X}^\circ), Y) l_{01}(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^w) \right].
 \end{aligned}$$

Here, $l_{01}(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^w) := 1 - \mathbb{1}_{[\mathbf{X}^w = \mathbf{g}(\mathbf{X}^\circ)]}$. The term (a2) in Eq. (A.11) can be expressed as:

$$\begin{aligned}
 (a2) &= \sum_{\mathbf{x}^w \in \mathcal{X}^w} \{ l(f(\mathbf{x}^w, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{X}^\circ)]} - l(f(\mathbf{x}^w, \mathbf{X}^\circ), Y) \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{X}^\circ)]} p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y) \} \\
 &= l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y) - l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y) p_*(\mathbf{g}(\mathbf{X}^\circ) | \mathbf{X}^\circ, Y) \\
 &= l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y) \left(\sum_{\mathbf{x}^w \in \mathcal{X}^w} p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y) \right) \\
 &\quad - l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y) \left(\sum_{\mathbf{x}^w \in \mathcal{X}^w} p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y) \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{X}^\circ)]} \right) \\
 &= l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y) \left(\sum_{\mathbf{x}^w \in \mathcal{X}^w} p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y) (1 - \mathbb{1}_{[\mathbf{x}^w = \mathbf{g}(\mathbf{X}^\circ)]}) \right) \\
 &= l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y) \mathbb{E}_{p_*(\mathbf{x}^w | \mathbf{X}^\circ, Y)} \left[l_{01}(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^w) \right].
 \end{aligned}$$

By substituting these results into (a1) and (a2) of Eq. (A.11), Eq. (A.11) can be rewritten as follows:

$$|R_l(f) - R_{l,g}(f)| \leq \mathbb{E}_{p_*(\mathbf{x}, Y)} \left[(l(f(\mathbf{X}^w, \mathbf{X}^\circ), Y) + l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y)) l_{01}(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^w) \right]. \quad (A.12)$$

Since l , l_{01} , f and $\mathbf{g}$ are all measurable functions, applying the Cauchy-Schwarz inequality to the RHS of Eq. (A.12), $|R_l(f) - R_{l,g}(f)|$ can be upper-bounded as follows:

$$\begin{aligned}
 &|R_l(f) - R_{l,g}(f)| \\
 &\leq \mathbb{E}_{p_*(\mathbf{x}, Y)} \left[(l(f(\mathbf{X}^w, \mathbf{X}^\circ), Y) + l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y)) l_{01}(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^w) \right] \\
 &\leq \left(\underbrace{\mathbb{E}_{p_*(\mathbf{x}, Y)} \left[(l(f(\mathbf{X}^w, \mathbf{X}^\circ), Y))^2 \right]}_{(b1)} \times \underbrace{\mathbb{E}_{p_*(\mathbf{x})} \left[(l_{01}(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^w))^2 \right]}_{(b2)} \right)^{\frac{1}{2}} \\
 &\quad + \left(\underbrace{\mathbb{E}_{p_*(\mathbf{x}^\circ, Y)} \left[(l(f(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^\circ), Y))^2 \right]}_{(b3)} \times \underbrace{\mathbb{E}_{p_*(\mathbf{x})} \left[(l_{01}(\mathbf{g}(\mathbf{X}^\circ), \mathbf{X}^w))^2 \right]}_{(b2)} \right)^{\frac{1}{2}}. \quad (A.13)
 \end{aligned}$$

The terms (b1)–(b3) of Eq. (A.13) can be expressed as:

$$(b1) = \mathbb{E}_{p_*(x,y)} \left[(l(f(\mathbf{X}), Y))^2 - 0^2 \right] \leq 2U_l \mathbb{E}_{p_*(x,y)} [l(f(\mathbf{X}), Y)] = 2U_l R_l(f),$$

$$(b2) = \mathbb{E}_{p_*(x)} [l_{01}(\mathbf{g}(\mathbf{X}^o), \mathbf{X}^w)] \leq \sum_{j \in [F^w]} \mathbb{E}_{p_*(x)} [l_{01}(g_j(\mathbf{X}^o), X_j^w)] = \sum_{j \in [F^w]} R_{01,j}(g_j),$$

$$(b3) = \mathbb{E}_{p_*(x^o,y)} \left[(l(f(\mathbf{g}(\mathbf{X}^o), \mathbf{X}^o), Y))^2 - 0^2 \right] \leq 2U_l \mathbb{E}_{p_*(x^o,y)} [l(f(\mathbf{g}(\mathbf{X}^o), \mathbf{X}^o), Y)] = 2U_l R_{l,g}(f).$$

Here, in (b1) and (b3), the fact that the function $x \mapsto x^2$ is $2U_l$ -Lipschitz continuous on the interval $[0, U_l]$ was utilized.

Applying the above inequalities related to (b1)–(b3) to the RHS of Eq. (A.13), $|R_l(f) - R_{l,g}(f)|$ can be upper-bounded as follows:

$$\begin{aligned} & |R_l(f) - R_{l,g}(f)| \\ & \leq \left\{ 2U_l R_l(f) \sum_{j \in [F^w]} R_{01,j}(g_j) \right\}^{\frac{1}{2}} + \left\{ 2U_l R_{l,g}(f) \sum_{j \in [F^w]} R_{01,j}(g_j) \right\}^{\frac{1}{2}} \\ & = \left(\sqrt{R_l(f)} + \sqrt{R_{l,g}(f)} \right) \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \end{aligned} \quad (A.14)$$

Thus, Lemma 4.1 is proven. $\square$

A.3. Proof of Theorem 4.2

From Lemma 4.1, the following lemma holds:

Lemma A.1. *For any $f \in \mathcal{F}$, $\mathbf{g} \in \mathcal{G}$ and l bounded by $U_l < \infty$, the following inequality holds:*

$$|R_l(f) - R_{l,g}(f)| \leq \left(2\sqrt{R_l(f)} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}} \right) \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \quad (A.15)$$

Proof of Lemma A.1. From Lemma 4.1, for any $f \in \mathcal{F}$, $\mathbf{g} \in \mathcal{G}$ and l bounded by $U_l < \infty$, the following inequality holds:

$$\left| \sqrt{R_l(f)} - \sqrt{R_{l,g}(f)} \right| \leq \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \quad (A.16)$$

Hence, $\sqrt{R_{l,g}(f)}$ can be upper-bounded as follows:

$$\begin{aligned} \sqrt{R_{l,g}(f)} &= \sqrt{R_{l,g}(f)} + \sqrt{R_l(f)} - \sqrt{R_l(f)} \\ &\leq \sqrt{R_l(f)} + \left| \sqrt{R_l(f)} - \sqrt{R_{l,g}(f)} \right| \\ &\leq \sqrt{R_l(f)} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \end{aligned} \quad (A.17)$$

By applying the above inequality to the RHS of Eq. (A.14), $|R_l(f) - R_{l,g}(f)|$ can be upper-bounded as follows:

$$|R_l(f) - R_{l,g}(f)| \leq \left(2\sqrt{R_l(f)} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}} \right) \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \quad (\text{A.18})$$

□

By leveraging Lemma A.1, Theorem 4.2 is proven as follows.

Proof of Theorem 4.2. From Section 3.1, we define the empirical risk minimizer in ordinary supervised learning as follows:

$$f_S := \arg \min_{f \in \mathcal{F}} \hat{R}_l(f).$$

The LHS of Eq. (4.7) can be rewritten as:

$$\begin{aligned} & R_{l,g}(f_{g,\bar{S}}) - R_l(f_{\mathcal{F}}) \\ &= \underbrace{R_{l,g}(f_{g,\bar{S}}) - \hat{R}_{l,g}(f_{g,\bar{S}})}_{(\text{a1})} + \underbrace{\hat{R}_{l,g}(f_{g,\bar{S}}) - R_{l,g}(f_S)}_{(\text{a2})} \\ & \quad + \underbrace{R_{l,g}(f_S) - R_l(f_S)}_{(\text{a3})} + \underbrace{R_l(f_S) - R_l(f_{\mathcal{F}})}_{(\text{a4})}. \end{aligned} \quad (\text{A.19})$$

The terms (a1) and (a2) in Eq. (A.19) can be upper-bounded as follows:

$$\begin{aligned} (\text{a1}) &\leq \max_{f \in \mathcal{F}} |R_{l,g}(f) - \hat{R}_{l,g}(f)|, \\ (\text{a2}) &\leq \hat{R}_{l,g}(f_S) - R_{l,g}(f_S) \leq \max_{f \in \mathcal{F}} |R_{l,g}(f) - \hat{R}_{l,g}(f)|. \end{aligned}$$

The term (a3) in Eq. (A.19) can be upper-bounded using Lemma A.1 as follows:

$$(\text{a3}) \leq \left\{ 2\sqrt{R_l(f_S)} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}} \right\} \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \quad (\text{A.20})$$

Additionally, $R_l(f_S)$ can be upper-bounded as follows:

$$\begin{aligned} R_l(f_S) &= R_l(f_S) - \hat{R}_l(f_S) + \hat{R}_l(f_S) - R_l(f_{\mathcal{F}}) + R_l(f_{\mathcal{F}}) \\ &\leq R_l(f_S) - \hat{R}_l(f_S) + \hat{R}_l(f_{\mathcal{F}}) - R_l(f_{\mathcal{F}}) + R_l(f_{\mathcal{F}}) \\ &\leq R_l(f_{\mathcal{F}}) + 2 \max_{f \in \mathcal{F}} |R_l(f) - \hat{R}_l(f)|. \end{aligned} \quad (\text{A.21})$$

Hence, (a3) in Eq. (A.19) can be upper-bounded as follows:

$$(\text{a3}) \leq \left(2 \left(R_l(f_{\mathcal{F}}) + 2 \max_{f \in \mathcal{F}} |R_l(f) - \hat{R}_l(f)| \right)^{\frac{1}{2}} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}} \right) \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \quad (\text{A.22})$$

Similarly, the term (a4) in Eq. (A.19) can be upper-bounded as follows:

$$R_l(f_S) - R_l(f_{\mathcal{F}}) \leq 2 \max_{f \in \mathcal{F}} |R_l(f) - \hat{R}_l(f)|.$$

By applying the above inequalities regarding (a1)–(a4) to the RHS of Eq. (A.19), it can be upper-bounded as follows:

$$\begin{aligned} & R_{l,g}(f_{g,\bar{S}}) - R_l(f_{\mathcal{F}}) \\ &\leq 2 \max_{f \in \mathcal{F}} |R_{l,g}(f) - \hat{R}_{l,g}(f)| + 2 \max_{f \in \mathcal{F}} |R_l(f) - \hat{R}_l(f)| \\ & \quad + \left\{ 2 \left(R_l(f_{\mathcal{F}}) + 2 \max_{f \in \mathcal{F}} |R_l(f) - \hat{R}_l(f)| \right)^{\frac{1}{2}} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}} \right\} \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \end{aligned} \quad (\text{A.23})$$

From the uniform law of large numbers (Mohri et al., 2018), for any $\delta \in (0, 1)$, the following holds with a probability of at least $1 - \delta/2$:

$$\begin{aligned} \max_{f \in \mathcal{F}} |R_{l,g}(f) - \widehat{R}_{l,g}(f)| &\leq 2\mathfrak{R}_n^g(\widetilde{\mathcal{F}}_l) + U_l \sqrt{\frac{\log(4/\delta)}{2n}}, \\ \max_{f \in \mathcal{F}} |R_l(f) - \widehat{R}_l(f)| &\leq 2\mathfrak{R}_n^*(\widetilde{\mathcal{F}}_l) + U_l \sqrt{\frac{\log(4/\delta)}{2n}}. \end{aligned}$$

Here, $\widetilde{\mathcal{F}}_l := \{(\mathbf{x}, y) \mapsto l(f(\mathbf{x}), y) : f \in \mathcal{F}\}$.

From the assumption that l is L_l -Lipschitz continuous, it follows that $\mathfrak{R}_n^*(\widetilde{\mathcal{F}}_l) \leq L_l \mathfrak{R}_n^*(\mathcal{F})$ and $\mathfrak{R}_n^g(\widetilde{\mathcal{F}}_l) \leq L_l \mathfrak{R}_n^g(\mathcal{F})$ (Lemma 26.9 in (Shalev-Shwartz & Ben-David, 2014)).

Thus, for any $\delta \in (0, 1)$, the following holds with a probability of at least $1 - \delta$:

$$\begin{aligned} &R_{l,g}(f_{g,\bar{S}}) - R_l(f_{\mathcal{F}}) \\ &\leq 4 \left(L_l \mathfrak{R}_n^*(\mathcal{F}) + L_l \mathfrak{R}_n^g(\mathcal{F}) + U_l \sqrt{\frac{\log(4/\delta)}{2n}} \right) \\ &\quad + \left\{ 2 \left(R_l(f_{\mathcal{F}}) + 4L_l \mathfrak{R}_n^*(\mathcal{F}) + 2U_l \sqrt{\frac{\log(4/\delta)}{2n}} \right)^{\frac{1}{2}} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}} \right\} \left(2U_l \sum_{j \in [F^w]} R_{01,j}(g_j) \right)^{\frac{1}{2}}. \end{aligned} \quad (\text{A.24})$$

□

A.4. Proof of Theorem 4.3

Proof of Theorem 4.3. By assumption, there exist true deterministic functions $g_j^* : \mathcal{X}^o \rightarrow \mathcal{X}_j^w$ for any $j \in [F^w]$, and $(g_1^*, \dots, g_{F^w}^*) \in \mathcal{G}$. Therefore, when $\mathbf{g}_{\bar{S}} = (g_{\bar{S},1}, \dots, g_{\bar{S},F^w})$ is obtained by the methods that achieve consistency (Cour et al., 2011; Feng et al., 2020; Xu et al., 2021; Natarajan et al., 2013; Ishida et al., 2017; Yu et al., 2018), the following holds:

$$n \rightarrow \infty, \quad R_{01,j}(g_{\bar{S},j}) \rightarrow 0, \quad \forall j \in [F^w]. \quad (\text{A.25})$$

Additionally, by assumption, there exists a true deterministic function $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ for label prediction, and $f^* \in \mathcal{F}$. Hence, the following holds:

$$R_l(f_{\mathcal{F}}) = 0. \quad (\text{A.26})$$

Thus, if $\mathfrak{R}_n^*(\mathcal{F})$ and $\mathfrak{R}_n^g(\mathcal{F})$ are monotonically decreasing with respect to n and converge to 0, the error bound established in Theorem 4.2 converges to 0 as $n \rightarrow \infty$.

□

A.5. Proof of Theorem 4.4

For a weak dataset $\bar{S}$ and a positive real-valued vector $\mathbf{r}$, define the feature estimation models $\mathbf{g}_{\bar{S}}^{(\mathbf{r})} = (g_{\bar{S},1}^{(\mathbf{r})}, \dots, g_{\bar{S},F^w}^{(\mathbf{r})})$ as follows:

$$g_{\bar{S},j}^{(\mathbf{r})} := \arg \min_{g \in \mathcal{G}(\mathbf{r}_j, \bar{S})} \widehat{R}_{l,j}(g), \quad \forall j \in [F^w]. \quad (\text{A.27})$$

Using Lemma A.1, Theorem 4.4 is proven as follows.

Proof of Theorem 4.4. The LHS of Eq. (4.10) can be rewritten as follows:

$$R_{l,f}(\mathbf{g}_{f,\bar{S}}^{(\mathbf{r})}) - R_l(f) = \underbrace{R_{l,f}(\mathbf{g}_{f,\bar{S}}^{(\mathbf{r})}) - \widehat{R}_{l,f}(\mathbf{g}_{f,\bar{S}}^{(\mathbf{r})})}_{(\text{a1})} + \underbrace{\widehat{R}_{l,f}(\mathbf{g}_{f,\bar{S}}^{(\mathbf{r})}) - R_{l,f}(\mathbf{g}_{\bar{S}}^{(\mathbf{r})})}_{(\text{a2})} + \underbrace{R_{l,f}(\mathbf{g}_{\bar{S}}^{(\mathbf{r})}) - R_l(f)}_{(\text{a3})}. \quad (\text{A.28})$$

The terms (a1) and (a2) of Eq. (A.28) can be upper-bounded as follows:

$$(a1) \leq \max_{\mathbf{g} \in \mathcal{G}(\mathbf{r}, \bar{\mathbf{S}})} |R_{l,f}(\mathbf{g}) - \hat{R}_{l,f}(\mathbf{g})|, \quad (\text{A.29})$$

$$(a2) \leq \hat{R}_{l,f}(\mathbf{g}_{\bar{\mathbf{S}}}^{(r)}) - R_{l,f}(\mathbf{g}_{\bar{\mathbf{S}}}^{(r)}) \leq \max_{\mathbf{g} \in \mathcal{G}(\mathbf{r}, \bar{\mathbf{S}})} |R_{l,f}(\mathbf{g}) - \hat{R}_{l,f}(\mathbf{g})|. \quad (\text{A.30})$$

The term (a3) of Eq. (A.28) can be upper-bounded using Lemma A.1 as follows:

$$\begin{aligned} (a3) &\leq |R_{l,f}(\mathbf{g}_{\bar{\mathbf{S}}}^{(r)}) - R_l(f)| \\ &\leq \left\{ 2\sqrt{R_l(f)} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(\mathbf{g}_{\bar{\mathbf{S}},j}^{(r_j)}) \right)^{\frac{1}{2}} \right\} \left(2U_l \sum_{j \in [F^w]} R_{01,j}(\mathbf{g}_{\bar{\mathbf{S}},j}^{(r_j)}) \right)^{\frac{1}{2}}. \end{aligned} \quad (\text{A.31})$$

Therefore, by applying Eqs. (A.29), (A.30), and (A.31) to Eq. (A.28), we obtain:

$$\begin{aligned} R_{l,f}(\mathbf{g}_{\bar{\mathbf{S}}}^{(r)}) - R_l(f) &\leq 2 \max_{\mathbf{g} \in \mathcal{G}(\mathbf{r}, \bar{\mathbf{S}})} |R_{l,f}(\mathbf{g}) - \hat{R}_{l,f}(\mathbf{g})| \\ &\quad + \left\{ 2\sqrt{R_l(f)} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(\mathbf{g}_{\bar{\mathbf{S}},j}^{(r_j)}) \right)^{\frac{1}{2}} \right\} \left(2U_l \sum_{j \in [F^w]} R_{01,j}(\mathbf{g}_{\bar{\mathbf{S}},j}^{(r_j)}) \right)^{\frac{1}{2}}. \end{aligned} \quad (\text{A.32})$$

From the uniform law of large numbers (Mohri et al., 2018), for any $\delta \in (0, 1)$, the following holds with a probability of at least $1 - \delta$:

$$\max_{\mathbf{g} \in \mathcal{G}(\mathbf{r}, \bar{\mathbf{S}})} |R_{l,f}(\mathbf{g}) - \hat{R}_{l,f}(\mathbf{g})| \leq 2\mathfrak{R}_n^*(\tilde{\mathcal{G}}_{l,f}(\mathbf{r}, \bar{\mathbf{S}})) + U_l \sqrt{\frac{\log(2/\delta)}{2n}}. \quad (\text{A.33})$$

Furthermore, by applying Eq. (A.33) to Eq. (A.32), we obtain that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, Eq. (4.10) holds:

$$\begin{aligned} R_{l,f}(\mathbf{g}_{\bar{\mathbf{S}}}^{(r)}) - R_l(f) &\leq 4\mathfrak{R}_n^*(\tilde{\mathcal{G}}_{l,f}(\mathbf{r}, \bar{\mathbf{S}})) + 2U_l \sqrt{\frac{\log(2/\delta)}{2n}} \\ &\quad + \left\{ 2\sqrt{R_l(f)} + \left(2U_l \sum_{j \in [F^w]} R_{01,j}(\mathbf{g}_{\bar{\mathbf{S}},j}^{(r_j)}) \right)^{\frac{1}{2}} \right\} \left(2U_l \sum_{j \in [F^w]} R_{01,j}(\mathbf{g}_{\bar{\mathbf{S}},j}^{(r_j)}) \right)^{\frac{1}{2}}. \end{aligned} \quad (\text{A.34})$$

□

A.6. Proof of Theorem 4.5

Proof of Theorem 4.5. By assumption, there exist true deterministic functions $g_j^* : \mathcal{X}^o \rightarrow \mathcal{X}_j^w$ for any $j \in [F^w]$, and $(g_1^*, \dots, g_{F^w}^*) \in \mathcal{G}$. In this case, for any $\mathbf{r}$ and $\bar{\mathbf{S}}$, it holds that $\mathbf{g}^* \in \mathcal{G}(\mathbf{r}, \bar{\mathbf{S}})$. Therefore, the following holds:

$$R_{01,j}(g_{\mathcal{G}(\mathbf{r}, \bar{\mathbf{S}}),j}^*) = 0, \quad \forall j \in [F^w]. \quad (\text{A.35})$$

For any $j \in [F^w]$, define $\mathcal{G}_j(r_j, \bar{\mathbf{S}}_j)$ as the set of hypotheses that satisfy the following two conditions: (i) each element is a solution obtained by methods that are guaranteed to achieve consistency (Cour et al., 2011; Feng et al., 2020; Xu et al., 2021; Natarajan et al., 2013; Ishida et al., 2017; Yu et al., 2018), and (ii) its empirical risk is at most r_j . As n increases and $r_j \rightarrow 0$, the assumptions on $\bar{R}_{l,j}$ and the theoretical guarantees of these consistent methods for g_j imply the following:

$$R_{01,j}(g_j) \rightarrow 0, \quad \forall g_j \in \mathcal{G}_j(r_j, \bar{\mathbf{S}}_j), \quad \forall j \in [F^w]. \quad (\text{A.36})$$

Additionally, by assumption, there exists a true deterministic function $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ for label prediction, and $f^* \in \mathcal{F}$. Hence, the following holds:

$$R_l(f_{\mathcal{F}}) = 0. \quad (\text{A.37})$$

Thus, under the conditions of Theorem 4.3, the following holds for $f_{g,\bar{S}}$ obtained through LAC-dWFL’s step (ii):

$$n \rightarrow \infty, R_{l,g}(f_{g,\bar{S}}) \rightarrow 0. \quad (\text{A.38})$$

Furthermore, using Theorem 3.1, and additionally letting $r_j \rightarrow 0$ as n increases for any $j \in [F^w]$, the following holds:

$$n \rightarrow \infty, R_l(f_{g,\bar{S}}) \rightarrow 0, \forall g \in \mathcal{G}(\mathbf{r}, \bar{S}). \quad (\text{A.39})$$

Since l is L_l -Lipschitz continuous and $f_{g,\bar{S}}$ is L_f -Lipschitz continuous, using Talagrand’s lemma (Shalev-Shwartz & Ben-David, 2014), the following holds:

$$\mathfrak{R}_n^*(\tilde{\mathcal{G}}_{l,f}(\mathbf{r}, \mathbf{S})) \leq L_l L_f \mathfrak{R}_n^*(\mathcal{G}(\mathbf{r}, \mathbf{S})). \quad (\text{A.40})$$

Consequently, if $\mathfrak{R}_n^*(\mathcal{G}(\mathbf{r}, \mathbf{S}))$ and $\mathfrak{R}_n^*(\mathcal{G}_j(r_j, \bar{S}_j))$ for any $j \in [F^w]$ are monotonically decreasing with respect to n and converge to 0, the error bound established in Theorem 4.4 converges to 0 as $n \rightarrow \infty$. $\square$

B. Detail Information of Experiments

B.1. Details of Datasets

We used four real-world datasets: *Adult* (Becker & Kohavi, 1996), *Bank Marketing* (Moro & Cortez, 2014), *kick* (Vanschoren et al., 2013), and *Census-Income (KDD)* (cen, 2000; Dua & Graff, 2017) in Section 5. In Section C.1, we additionally used two datasets: *Default of Credit Card Clients* (Yeh, 2009) and *Diabetes 130-US Hospitals for Years 1999–2008* (Kahn). We will refer to them as *Adult*, *Bank*, *Kick*, *Census*, *Default*, and *Diabetes*, respectively. These datasets can be downloaded from UCI Machine Learning Repository (Dua & Graff, 2017) or OpenML (Vanschoren et al., 2013). Table B.1 summarizes the characteristics of these datasets. All binary features were set to take values of either 0 or 1, all categorical features were encoded using one-hot encoding, and all continuous features were scaled to fall within the range $[0, 1]$. For the *Adult*, *Bank*, *Kick*, *Default*, and *Diabetes* datasets, all available samples were used. For the *Census* dataset, experiments were conducted using 50,000 randomly sampled data points.

Table B.1. Outline of datasets. binary, categorical, and numerical represent the number of features of each type, respectively.

dataset	Adult	Bank	Kick	Census	Default	Diabetes
data size	48842	45211	72983	299285	30000	101766
binary	1	3	3	3	1	10
categorical	7	5	12	22	2	12
numerical	6	8	16	10	20	10
target	binary	binary	binary	binary	binary	3 classes

B.2. Details of Experimental setup

We summarize the settings of feature estimation models g and label prediction model f . Both g and f were implemented using two-layer perceptrons with hidden layers of width 500 and ReLU as an activation function. The optimization algorithm used across all models is Adam (Kingma & Ba, 2014), with the following hyperparameters: a learning rate of 0.0005, batch size of 512, 100 epochs, and a weight decay of 0.0002. Logistic loss was used for training f .

We summarize the method for calculating the error bound presented in Theorem 4.2 for this experiment. The parameter δ was set to 0.01. Additionally, assuming that $\mathcal{F}$ is sufficiently expressive, we set $R_l(f_{\mathcal{F}}) = 0$. During inference, the predicted label from f was determined based on the largest output value of f . Consequently, scaling the outputs of f does not affect the inference results. Thus, we assumed that the maximum value of each element in f ’s output is 1 and set $U_l = 2.0$. Since the loss function l for label prediction is the logistic loss, which is 1-Lipschitz continuous, it follows that $\mathfrak{R}_n^*(\tilde{\mathcal{F}}_l) = \mathfrak{R}_n^*(\mathcal{F})$ and $\mathfrak{R}_n^g(\tilde{\mathcal{F}}_l) = \mathfrak{R}_n^g(\mathcal{F})$ (Lemma 26.9 in (Shalev-Shwartz & Ben-David, 2014)). The terms $\mathfrak{R}_n^*(\mathcal{F})$ and $\mathfrak{R}_n^g(\mathcal{F})$ were computed using the upper bound on the Rademacher complexity of a multilayer perceptron derived by Neyshabur et al. (Theorem 1 in (Neyshabur et al., 2015)). To compute this upper bound, it is necessary to determine μ , which represents the upper bound on the l_p -norm of all parameters of f , as well as the value of p . In this experiment, $p = 2$ was chosen. Furthermore, since it is possible to scale all parameter values of f without affecting the inference results for predicting a single label, we set $\mu = 1$.

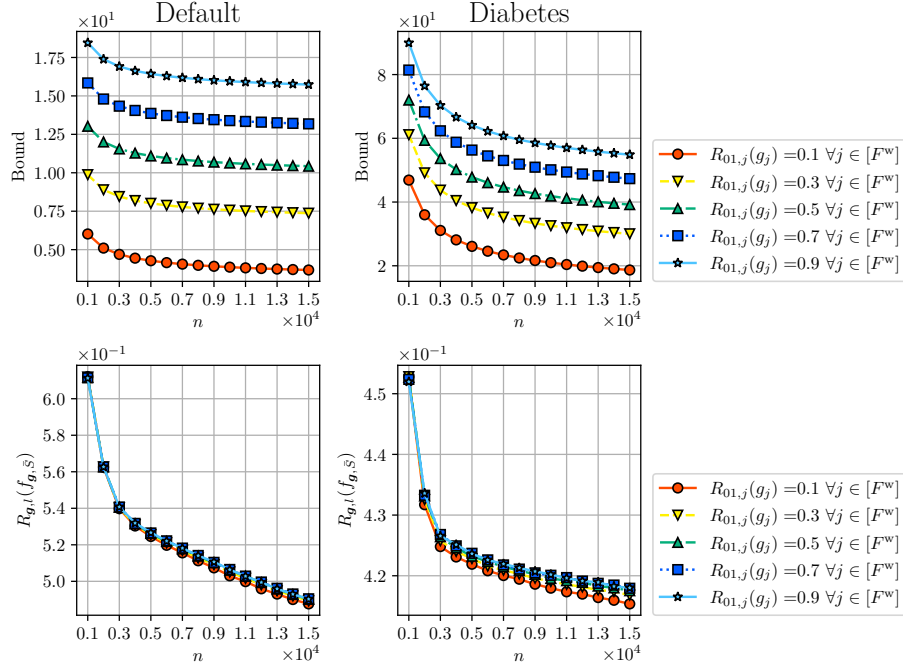


Figure C.3. A comparison between $R_{l,g}(f_{g,\bar{S}})$ and the error bound derived in Theorem 4.2, for various estimation errors of g using Default and Diabetes datasets.

C. Additional Experiments

C.1. Additional Datasets

In Section 5, we validated our theoretical results through numerical experiments using real-world datasets. In this section, we further examine the validity of our findings on two additional datasets that were not used in Section 5.2. The datasets used here are Default (Yeh, 2009) and Diabetes (Kahn), with details provided in Appendix B.1. Figure C.3 presents the results of the same experimental procedure applied to these datasets, following the methodology outlined in Section 5.2. In these datasets, variations in the estimation accuracy of the exact values of WFs resulted in only minor changes in the risk of the downstream task. As a consequence, visualizations similar to those in Figure 5.2 significantly compromised readability. To address this, we separately plot the observed risks and the theoretical bound derived from Theorem 4.7.

From Figure C.3, we observe that a smaller value of $R_{01,j}(g_j)$ leads to a greater reduction in $R_{g,l}(f_{g,\bar{S}})$ as n increases. This trend is consistent with the behavior of the error bounds illustrated in the same figure. Regarding the rate of decrease in both $R_{g,l}(f_{g,\bar{S}})$ and the bound with respect to n , we find that their sensitivity to changes in $R_{01,j}(g_j)$ is less pronounced in the Default dataset than in the Diabetes dataset. This difference can be attributed to the fact that the Default dataset contains only two WFs, whereas the Diabetes dataset includes four. Consequently, the influence of WFs on downstream tasks is inherently smaller in the Default dataset. Therefore, from the results on these two additional datasets, we further confirm that our derived error bound in Theorem 4.7 successfully capture the relationship between the rate of decrease in $R_{g,l}(f_{g,\bar{S}})$ with increasing n and the value of $R_{01,j}(g_j)$.

C.2. Comparison with the Case WFs are Used for the Inputs of f

In this section, we compared the risk of the f obtained by directly using $\bar{X}^w$ for training (i.e., $g(X^o) = \bar{X}^w$) with the experimental results on risks presented in Section 5. We focused on the four datasets used in Section 5, as the two additional datasets examined in Section C.1 exhibited a relatively weaker dependency of the risk of f on the estimation accuracy of g . The training procedure for f remained identical to that described in Section 5. By conducting this comparison, we aim to provide empirical insight into how accurately g must estimate the exact values of WFs in order to improve the generalization performance of f .

Figure C.4 shows the results under the setting where all WFs are complementary features (CFs). Each observation of a

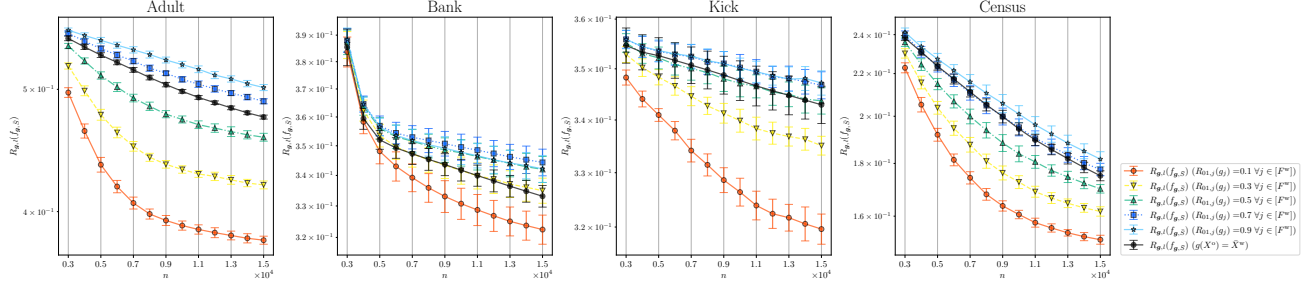


Figure C.4. Comparison between the risks in Section 5 and the risk of f when $g(X^o) = \bar{X}^w$. The results correspond to the case where all WFs are CFs, and both the mean and standard deviation over five trials are reported.

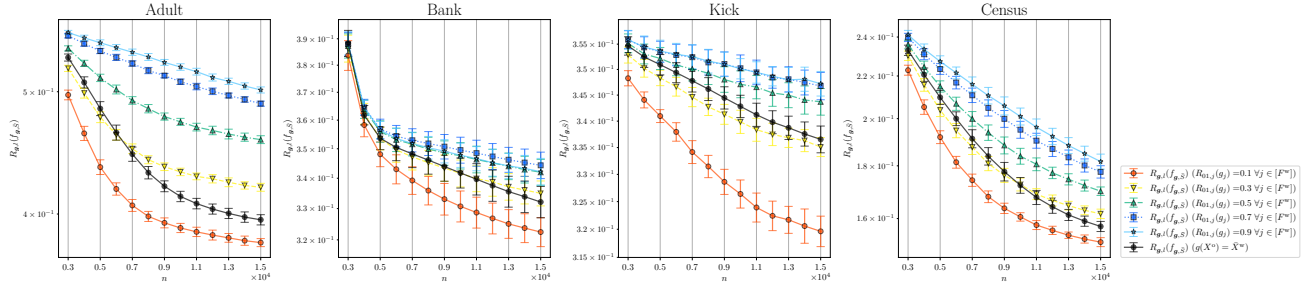


Figure C.5. Comparison between the risks in Section 5 and the risk of f when $g(X^o) = \bar{X}^w$. The results correspond to the case where each WF is observed as a set of size two including the exact value, with mean and standard deviation over five runs reported.

CF is sampled uniformly from all values except the exact one. From this figure, we observe that for the Adult, Kick, and Census datasets, a classification error of g_j below 0.5 is sufficient to achieve a model f that outperforms the baseline where $g(X^o) = \bar{X}^w$. In contrast, for the Bank dataset, the classification error of g_j must be below at least 0.3 to achieve similar improvement.

Figure C.5 presents the results for the setting in which each observed value of every WF is represented as a set of size two that includes the exact value. The additional value in the set, apart from the exact value, is uniformly sampled at random. When using such $\bar{X}^w$ as input to f , each WF is encoded as a one-hot vector, where the entries corresponding to the values in the observed set are set to one. These vectors are then used as inputs to f . From Figure C.5, in contrast to the case where WFs are CFs, we observe that g yielding better performance than the case where $g(X^o) = \bar{X}^w$ depends on n . This difference is attributed to the fact that $\bar{X}^w$ always includes the exact value, whereas estimation via g inevitably produces instances with incorrect values due to estimation errors. These results suggest that, for estimating the exact values of WFs, it may be more effective to output a probability distribution over possible values, rather than predicting a single deterministic value. Accordingly, developing methods where g produces a distribution as output, along with establishing a theoretical framework that accommodates such g , are important future directions for WFL. The findings in this paper are considered to provide a fundamental basis for such extensions.

D. Limitation

In this paper, we consider the basic framework of WFL, where a feature estimation model is constructed for each WF independently, using OFs as inputs. Accordingly, our framework does not consider approaches that incorporate dependencies among WFs into the construction of feature estimation models. For example, we do not address methods that first estimate the exact values of a given WF and then use those estimates as input features to infer the exact values of other WFs, or approaches that jointly estimate the exact values of multiple WFs within a single model. Developing such approaches is of practical importance and is left for future work. Furthermore, analyzing such methods would require a theoretical framework for quantifying how well the dependencies among WFs are captured, as well as for modeling these dependencies.

Theoretical insights presented in this paper are considered to provide a foundational basis for such future investigations.

For similar reasons, methods that construct f and g as a single unified model and train it jointly are not yet covered by the analysis presented in this paper. Developing and analyzing such methods also constitutes an important direction for future research.

Our analysis does not account for the feature importance of each WF in the downstream prediction task. Intuitively, the estimation accuracy for WFs that are more strongly related to the downstream target variable should have a greater impact on the error bound for learning f , compared to those WFs that are weakly related or irrelevant. However, the error bound we derived does not account for such feature importance and thus cannot differentiate the relative contributions of individual WFs. Therefore, deriving an error bound that incorporates feature importance remains an important direction for future work. It is considered that the results presented in this paper provide a solid foundation for such an extension.