Statistica Sinica Preprint No: SS-2023-0202	
Title	Addressing Label Noise in Causation Classification via Kernel Embeddings
Manuscript ID	SS-2023-0202
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0202
Complete List of Authors	Pingbo Hu and Grace Y. Yi
Corresponding Authors	Grace Y. Yi
E-mails	gyi5@uwo.ca
Notice: Accepted author version.	

Statistica Sinica

Addressing Label Noise in Causation Classification via Kernel Embeddings

Pingbo Hu[1] and Grace Y. Yi[1,2,*]

[1] *Department of Statistical and Actuarial Sciences, University of Western Ontario*

[2] *Department of Computer Science, University of Western Ontario*

[*] *the corresponding author*

E-mail: phu43@uwo.ca; gyi5@uwo.ca

Abstract: A basic task of causal inference is to infer whether there exists a cause-effect relationship between two sets of vectors of interest, akin to a binary classification problem. With a sequence of independent and identically distributed paired vectors, one may employ the *kernel mean embedding of probability distribution* to map the empirical distribution to a feature space, and then train a classifier in the feature space to infer the causation for a future pair of vectors. This strategy, however, is susceptible to mislabeling, a common challenge in causation studies. In this paper, we explore this issue and quantify mislabeling effects. We develop valid learning methods with the mislabeling effects accounted for and theoretically justify the validity of the proposed methods.

Key words: causation learning, classification, kernel mean embedding, label noise.

1. Introduction

Learning cause-effect relationships has attracted extensive attention in both statistical and machine learning communities. The *potential outcome* framework, originating from Neyman (1923), is a popular statistical approach to infer causality (Rubin 1974). Alternatively, learning causal relationships among variables can also be framed as a classification problem. We may, at least in principle, consider all associated variables, including causes, outcomes, and confounding factors, and exhaustively form distinct pairs of X and W. For each pair (X, W), we assign a binary label, 1 or -1, to represent whether X is the cause of W based on a specific method tailored to individual applications. Ultimately, this is a classification problem with the binary label, denoted l, as the output and (X, W) as the input.

A motivating example is the SUP3 dataset from the Kaggle competition (Guyon 2013), consisting of $n \triangleq 162$ variable pairs across diverse domains, such as chemistry, climatology, ecology, economy, engineering, epidemiology, genomics, medicine, physics, and sociology. Each pair is labeled as either 1 or -1, indicating the presence or absence of a causal relationships within the pair.

Lopez-Paz et al. (2015) framed learning cause-effect relationships for paired vectors as a classification problem using *kernel mean embeddings* in the *reproducing kernel Hilbert space* (RKHS), a method further explored by other authors, including Mooij et al. (2016), Monti, Zhang and Hyvärinen (2020), and Tagasovska, Chavez-Demoulin and Vatter (2020).

These methods typically require training data to be free of mismeasurement, but real-world applications often violate this assumption due to mismeasurement, including covariate error (aka input error) and response error (aka output error) (e.g., Carroll et al. 2006; Yi 2017; Yi, Delaigle and Gustafson 2021). In cases where the response variable represents class membership, response error is known as "label noise", "label corruption", and "mis-labeling" in machine learning (e.g., Guo, Wang and Yi 2023; Guo, Yi and Wang 2024).

Here, we focus on the scenario where input variables X and W are error-free but the output label l is subject to mislabeling, a common issue in learning causation relationships. Label noise arises from annotation errors due to ambiguity in labeling instructions, lack of experience, subjective judgement, uncertainty, imprecise answers to sensitive questions, or inaccurate measurement instruments. In particular, mislabeling is a significant concern in using observational data to determine causal relationships, which is obscured by hidden factors.

Building on the framework of Lopez-Paz et al. (2015) for classification learning without mislabeling, we examine mislabeling effects and contribute by (1) expanding the causal learning framework to account for mislabeled outputs, (2) analyzing the impact of ignoring label noise, (3) establishing theoretical properties that generalize existing results, (4) devising a correction method to accommodate the label noise effects, and (5) introducing new metrics to evaluate classifier performance under mislabeling.

The remainder of this article is organized as follows. Sections 2 and 3 consider the case of precisely measured variables. Sections 4 - 6 focus on label noise, examining its effects in Section 4, proposing our correction method in Section 5, and introducing new metrics with sensitivity analyses in Section 6. Finally, Section 7 provides discussions, with technical details and additional numerical studies deferred to the supplementary material.

2. Learning Framework

2.1 Notation and Data Format

Considering the causal learning framework considered by Lopez-Paz et al. (2015), suppose $Z_i \triangleq (X_i, W_i)$ are independent random variables for $i = 1, \cdots, n$, and for each i, l_i is a binary label, taking value 1 if X_i is the cause of W_i and value 0 otherwise. Here, X_i and W_i can be either vectors or univariate random variables. As an example with $n = 2$, X_1 and W_1 may represent respectively an individual's smoking status and lung cancer status, while X_2 and W_2 may respectively indicate the raining status and presence of clouds for a day. While Z_1 and Z_2 have distinct practical meanings, a common question may arise to examine the presence or absence of the causal relationship for the variables within them, which can be reflected by the value of their associated binary label.

Additionally, for each $i = 1, \cdots, n$, there is a random sample of measurements for paired input Z_i, denoted as $\mathcal{S}_i = \big\{Z_{ij} \triangleq (X_{ij}, W_{ij}) \bigm| j = 1, \cdots, m_i\big\}$, where the Z_{ij} with $j = 1, \cdots, m_i$ are independently and identically distributed (i.i.d) having the same joint probability distribution P_i of random vector Z_i, and m_i is a positive integer that may depend on i. These samples could, for example, represent measurements of smoking status and lung cancer status for m_i patients or measurements of raining status and the presence of clouds over m_i days. This framework was considered by Lopez-Paz et al. (2015), with the objective of training a binary classifier using the output data $\big\{l_i \bigm| i = 1, \cdots, n\big\}$, together with mapping the input $\mathcal{S}_i$ into a feature space. The goal is to predict the causation for a new pair of variables, say $(\widetilde{X}, \widetilde{W})$.

We make some comments here. As the practical meaning for each pair Z_i may differ across different indices i, analyzing them together might appear unnatural. However, those paired variables may share similarity in distribution, thus allowing us to examine them under the same framework.

While the Z_i differ for different index i, they can share some common elements or be related in nature, and the order of elements in Z_i matters. For example, in pair $Z_1 = (X_1, W_1)$ and $Z_2 = (X_2, W_2)$, X_1 may represent smoking status and W_1 may represent lung cancer status, whereas X_2 stands for chest pain status and W_2 can still represent lung cancer status. This setup yields identical W_1 and W_2, yet Z_1 differs from Z_2. Additionally, we may have $l_1 = 1$, showing that smoking is the cause of lung cancer, and $l_2 = -1$, indicating that chest pain is not the cause of lung cancer. On the other hand, if we interchange the roles of X_2 and W_2 such that X_2 represents lung cancer status and W_2 indicates chest pain status, then we may have $l_2 = 1$ to show that lung cancer causes chest pain (Potter and Higginson 2004).

Although variables in different pairs $Z_i = (X_i, W_i)$ for $i = 1, \cdots, n$ may share some elements or have practical connections, replicate measurements for Z_i, denoted $\left\{Z_{ij} \triangleq (X_{ij}, W_{ij}) \;\middle|\; j = 1, \cdots, m_i\right\}$, are assumed to be independently collected from m_i randomly selected subjects or units. Additionally, $\mathcal{S}_1, \cdots, \mathcal{S}_n$ are assumed to be independently formed.

More formally, let $(\mathcal{Z}, \tau_z)$ denote a separable topological space, with

τ_z representing the *topology* on the set $\mathcal{Z}$ (Armstrong 1983), and let $\sigma(\tau_z)$ denote the σ-algebra generated by τ_z. Let $\mathcal{P}$ denote the set of all Borel probability measures on the measurable space $(\mathcal{Z}, \sigma(\tau_z))$, and let $\mathcal{L} = \{-1, +1\}$. Let $\mathcal{M}$ denote a *mother distribution* defined on $\mathcal{P} \times \mathcal{L}$. For $i = 1, \cdots, n$, we assume that Z_i is a random variable mapping from a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ to the measurable space $(\mathcal{Z}, \sigma(\tau_z))$, with Ω, $\mathcal{E}$ and $\mathbb{P}$ representing a set, σ-algebra, and probability measure, respectively. We further assume that $\big\{\{P_i, l_i\} \bigm| i = 1, \cdots, n\big\}$ are independent and identically distributed (i.i.d.) from $\mathcal{M}$, where P_i is the probability measure of Z_i.

In summary, the data collection process involves *two-stage sampling*. First, n i.i.d. samples $\big\{\{P_i, l_i\} \bigm| i = 1, \cdots, n\big\}$ are generated from the *mother distribution* $\mathcal{M}$, and then for each i, m_i i.i.d samples $\mathcal{S}_i = \big\{Z_{ij} \bigm| j = 1, \cdots, m_i\big\}$ are generated from the probability measure P_i. This two-stage sampling framework is widely used in various domains, including *distribution learning* (e.g., Szabó et al. 2016) and *multi-instance learning* (e.g., Zhou and Xu 2007). In distribution learning, the *mother distribution* $\mathcal{M}$ is called a *Meta distribution*, where the i.i.d. assumption in the first stage sampling is typically imposed in both causal learning and distribution learning, although testing this assumption is difficult due to the unavailability of the probability measure P_i.

2.2 Training and Prediction Procedures

Lopez-Paz et al. (2015) developed the following *learning* algorithm:

- **Step 1**: for each i, we construct the probability measure:

  $$P_{\mathcal{S}_i}(A^*) \triangleq \frac{1}{m_i}\sum_{j=1}^{m_i} I\big\{Z_{ij} \in A^*\big\} \quad \text{for any } A^* \in \sigma(\tau_z), \tag{2.1}$$

  where $I(C^*)$ represents the indicator function, taking value 1 if the statement C^* is true and value 0 otherwise.

- **Step 2**: Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ denote a continuous, bounded, and positive-definite kernel function, and let $\mathcal{H}_k$ denote the induced *reproducing kernel Hilbert space* (RKHS) with the inner product, denoted $< \cdot, \cdot >_{\mathcal{H}_k}$, (Muandet et al. 2017, Section 2.2). For each i, use the *kernel mean embedding of probability distribution* to map $P_{\mathcal{S}_i}$ into $\mathcal{H}_k$ and let $\mu_k(P_{\mathcal{S}_i})$ denote its empirical kernel mean embedding, given by

  $$\mu_k(P_{\mathcal{S}_i}) = \frac{1}{m_i}\sum_{j=1}^{m_i} k(Z_{ij}, \cdot).$$

  As explained in Section S1 of the supplementary material, $\mu_k(P_{\mathcal{S}_i})$ is a random function from $\mathcal{Z}$ to $\mathbb{R}$ due to the randomness of $\mathcal{S}_i$; when the sample $\mathcal{S}_i$ is realized as s_i, the resulting $\mu_k(P_{s_i})$ becomes a deterministic function from $\mathcal{Z}$ to $\mathbb{R}$. Theorem S1 in the supplementary material establishes the convergence in mean of the empirical kernel mean embedding to the true kernel mean embedding. This mapping allows us to leverage the useful properties of the Hilbert space to analyze the data $\mathcal{S}_i$ through $\mu_k(P_{\mathcal{S}_i})$.

- **Step 3**: Using the data $\big\{\{\mu_k(P_{s_i}), l_i\} \bigm| i = 1, \cdots, n\big\}$, we train a nonlinear binary classifier with $\big\{\mu_k(P_{s_i}) \bigm| i = 1, \cdots, n\big\}$ and $\big\{l_i \bigm| i = 1, \cdots, n\big\}$ taken as the input and output, respectively.

With the trained classifier, we can predict whether a new vector, say $\widetilde{X}$, is the cause of another new vector, say $\widetilde{W}$, using realizations for a random sample $\widetilde{s} = \Big\{(\widetilde{x_j}, \widetilde{w_j}) \Bigm| j = 1, \cdots, \widetilde{m}\Big\}$ for the random vector $(\widetilde{X}, \widetilde{W})$.

3. Causation Learning Theory

For the kernel function k considered in Step 2 of Section 2.2 and any $P \in \mathcal{P}$, let $\mu_k(P)$ denote the *kernel mean embedding of probability distribution* that maps P into RKHS $\mathcal{H}_k$, which represents a function from $\mathcal{Z}$ to $\mathbb{R}$, as detailed in Section S1 of the supplementary material. Let $\mu_k(\mathcal{P}) = \big\{\mu_k(P) \bigm| P \in \mathcal{P}\big\}$, which is a subset of $\mathcal{H}_k$: $\mu_k(\mathcal{P}) \subseteq \mathcal{H}_k$. Let $\mathcal{M}_k$ denote a measure on $\mu_k(\mathcal{P}) \times \mathcal{L}$ induced by $\mathcal{M}$ (Lopez-Paz et al. 2015, Lemma 2). Then $\big\{\{\mu_k(P_i), l_i\} \bigm| i = 1, \cdots, n\big\}$ is a sequence of i.i.d copies drawn from $\mathcal{M}_k$, which are used to train a binary classifier in the space $\mathcal{H}_k$.

Let $\mathcal{G} = \Big\{g : \mathcal{H}_k \to \mathbb{R} \Big| \ g \text{ is a measurable functional}\Big\}$, where g in $\mathcal{G}$ is termed a *functional* because it maps a space of functions (i.e., $\mathcal{H}_k$) to $\mathbb{R}$, and g from $\mathcal{H}_k$ to $\mathbb{R}$ is called *measurable* if the preimage of any element in Borel σ-algebra in $\mathbb{R}$ belongs to a σ-algebra in $\mathcal{H}_k$. We follow Vapnik (1998) to find a suitable discriminant functional $f : \mathcal{H}_k \to \mathbb{R}$ such that the sign of $f(\mu_k(P))$ is used to predict the output l of $\mu_k(P)$. For the 0-1 loss function $L : \mathcal{L} \times \mathcal{L} \to \mathbb{R}^+$, given by $L(l_1, l_2) \triangleq \frac{|l_1 - l_2|}{2}$, we wish to minimize the *risk*:

$$R(f) \triangleq \mathbb{E}\{L(\text{sign}(f(\mu_k(P))), l)\}, \tag{3.2}$$

where the expectation is evaluated with respect to the joint distribution $\mathcal{M}_k$ for $\{\mu_k(P), l\}$, and sign(t) is given by sign$(t) = 1$ if $t \geq 0$, and sign$(t) = -1$ if $t < 0$. Letting $\ell(\alpha) = I\{\alpha \in [0, \infty)\}$, we re-write (3.2) as

$$R(f) = \mathbb{E}\{\ell(-lf(\mu_k(P)))\}. \tag{3.3}$$

The goal is to find the minimum value of $R(f)$, denoted R_0, over $\mathcal{G}$. However, computation of (3.3) is intractable due to the nonconvexity of $\ell(\cdot)$. As a remedy, one considers a surrogate function, say $\varphi : \mathbb{R} \to \mathbb{R}^+$, which is convex and well upper bound $\ell(\cdot)$, with $\ell(\alpha) \leq \varphi(\alpha)$ for any $\alpha \in \mathbb{R}$. Replacing $\ell(\cdot)$ in (3.3) with a convex surrogate $\varphi(\cdot)$, we define the φ-risk as $R_\varphi(f) \triangleq \mathbb{E}\{\varphi(-lf(\mu_k(P)))\}$.

Further, assessing $R_\varphi(f)$ for all $f \in \mathcal{G}$ is infeasible because $\mathcal{G}$ is too big. In practice, we usually consider a smaller set of $\mathcal{G}$, denoted $\mathcal{F}$. For example, $\mathcal{F}$ can be taken as the set of all bounded linear functionals on $\mathcal{H}_k$ (e.g., Conway 2019). We aim to find

$$f_0 = \text{argmin}_{f\in\mathcal{F}}\, R_{\varphi}(f). \tag{3.4}$$

The convexity of φ enables efficient convex optimization for solving (3.4). The φ-risk offers us a mathematically convenient measure to describe an upper bound for risk (3.3). While different surrogate functions may yield different upper bounds for (3.3), a well-calibrated surrogate function $\varphi(\cdot)$ can accurately approximate $\ell(\cdot)$ and allows us to identify meaningful upper bounds of risk (3.3), as discussed by Bartlett et al. (2006), who also explored a useful class of surrogate functions known as *classification-calibrated* convex surrogates, defined as follows.

Definition (Bartlett et al. 2006)**.** A convex function $\varphi : \mathbb{R} \to \mathbb{R}^+$ is called *classification-calibrated* if for any $\eta \neq \frac{1}{2}$,

$$\inf_{\alpha:\alpha(2\eta-1)\leq 0} \big\{\eta\varphi(\alpha) + (1-\eta)\varphi(-\alpha)\big\} > \inf_{\alpha\in\mathbb{R}} \big\{\eta\varphi(\alpha) + (1-\eta)\varphi(-\alpha)\big\}.$$

The class of classification-calibrated convex surrogate functions includes familiar functions such as $\varphi(u) = \log_2 \big\{1 + \exp(u)\big\}$ for the logistic loss $L(y, f(x)) = \log_2\, (1 + \exp(-yf(x)))$ used in logistic regression, $\varphi(u) = \max\{0, 1+u\}$ for the *hinge loss* $L(y, f(x)) = \max\{0, 1-yf(x)\}$ used in the *support vector machine (SVM)*, and $\varphi(u) = \exp(u)$ for the *exponential loss*

$L(y, f(x)) = \exp\{-yf(x)\}$ used in *Adaboost*, where $y \in \{-1, 1\}$, and $f(x)$ represents a predicted value.

Although using a convex surrogate function enables us to convert the intractable minimization problem (3.3) to a convex optimization problem, the unknown distribution $\mathcal{M}_k$ prevents us from obtaining f_0 directly from (3.4). To get around this difficulty, we replace $R_\varphi(f)$ in (3.4) with the empirical φ-risk:

$$\hat{R}_\varphi(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \varphi(-l_i f(\mu_k(P_{\mathcal{S}_i}))),$$

and aim to find

$$\hat{f} = \text{argmin}_{f\in\mathcal{F}}\ \hat{R}_\varphi(f). \tag{3.5}$$

The differences $R_\varphi(\hat{f}) - R_\varphi(f_0)$ and $R(\hat{f}) - R_0$ describe the performance of the classifier $\hat{f}$, where $R_\varphi(\hat{f})$ and $R(\hat{f})$ are random due to the involvement of data in $\hat{f}$. With a well-chosen φ function, in conjunction of $\mathcal{F}$ and kernel k, we expect $R_\varphi(\hat{f})$ and $R(\hat{f})$ to be close to or even identical to $R_\varphi(f_0)$ and R_0 in expectation, respectively. Let $m = \min\limits_{1\leq i\leq n} m_i$ and let $R(\mathcal{F})$ denote the *Rademacher complexity* of $\mathcal{F}$. Typically, the class $\mathcal{F}$ is chosen to ensure $R(\mathcal{F})$ is of order $\mathcal{O}(n^{-\frac{1}{2}})$, as considered in this paper (e.g., Lopez-Paz et al. 2015, Section 3.1).

Theorem 1. *Assume the following conditions hold:*

(R1). All elements in $\mathcal{F}$ are Lipschitz continuous with respect to the norm in $\mathcal{H}_k$, and there exists a common Lipschitz constant, denoted $L_{\mathcal{F}}$, for all elements in $\mathcal{F}$ such that for any $f \in \mathcal{F}$ and $h, h' \in \mathcal{H}_k$, $|f(h) - f(h')| \leq L_{\mathcal{F}} ||h - h'||_{\mathcal{H}_k}$;

(R2). There exists a positive constant B such that $\varphi(-lf(h)) \leq B$ for any $f \in \mathcal{F}$, $h \in \mathcal{H}_k$, and $l \in \mathcal{L}$;

(R3). $\varphi : \mathbb{R} \rightarrow \mathbb{R}^+$ is a Lipschitz continuous function with L_φ denoting a Lipschitz constant, such that $\varphi(\alpha) \geq \ell(\alpha)$;

(R4). The kernel function k associated with $\mathcal{H}_k$ satisfies $\sup_{z \in \mathcal{Z}} k(z, z) \leq 1$.

For any $0 < \delta < 1$, let

$$C(n, m, L_\varphi, L_{\mathcal{F}}, B) \triangleq 4L_\varphi R(\mathcal{F}) + 2B\sqrt{\frac{log(2n)}{2n}} + \frac{4L_\varphi L_{\mathcal{F}}}{n} \sum_{i=1}^{n} \left[\sqrt{\frac{\mathbb{E}\{k(Z_i, Z_i)\}}{m_i}} + \sqrt{\frac{log\left(2n^2\right)}{2m_i}} \right]. \tag{3.6}$$

Then for $\hat{f}$ in (3.5) and f_0 in (3.4),

(a). $0 \leq \mathbb{E}\{R_\varphi(\hat{f}) - R_\varphi(f_0)\} \leq C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{2B}{n}$;

(b). $\lim\limits_{n\rightarrow\infty} \lim\limits_{m\rightarrow\infty} \mathbb{E}\{R_\varphi(\hat{f}) - R_\varphi(f_0)\} = 0$;

(c). if φ is classification-calibrated and $\inf\limits_{h \in \mathcal{G}} R_\varphi(h) = \min\limits_{f \in \mathcal{F}} R_\varphi(f) = R_\varphi(f_0)$, then

(i) *there exists a nondecreasing continuous function* $\zeta_\varphi : \mathbb{R} \to [0,1]$ *with* $\zeta_\varphi(0) = 0$, *such that*

$$\mathbb{E}\{R(\hat{f}) - R_0\} \leq \zeta_\varphi\Big(C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{2B}{n}\Big);$$

(ii) $\lim_{n\to\infty} \lim_{m\to\infty} \mathbb{E}\{R(\hat{f}) - R_0\} = 0.$

(d). *If* $\inf_{h\in\mathcal{G}} R_\varphi(h) = \min_{f\in\mathcal{F}} R_\varphi(f) = R_\varphi(f_0)$, *then there exists a nonnegative, convex, continuous, and strictly increasing function* $\psi_\varphi : [0,1] \to \mathbb{R}$ *such that*

$$\psi_\varphi\Big(\mathbb{E}\{R(\hat{f}) - R_0\}\Big) \leq \mathbb{E}\{R_\varphi(\hat{f}) - R_\varphi(f_0)\}. \tag{3.7}$$

Furthermore, the following three conditions are equivalent:

(i) φ *is classification-calibrated;*

(ii) *For any sequence of real numbers,* $\big\{\theta_i \in [0,1] \big| i = 1, 2, \cdots\big\}$,

$$\lim_{i\to\infty} \psi_\varphi(\theta_i) = 0 \quad \textit{if and only if} \quad \lim_{i\to\infty} \theta_i = 0;$$

(iii) *For every sequence of measurable functional* $\Big\{f_i : \mathcal{H}_k \to \mathbb{R}\Big| i = 1, 2, \cdots\Big\}$,

$$\lim_{i\to\infty} R_\varphi(f_i) = R_\varphi(f_0) \quad \textit{implies} \quad \lim_{i\to\infty} R(f_i) = R_0.$$

The proof of Theorem 1 is presented in Section S1.3 of the supplementary material. Theorem 1 (a) is related to but differs from Theorem 3 of Lopez-Paz et al. (2015). Both theorems assume the same conditions and they describe upper bounds for $R_\varphi(\hat{f}) - R_\varphi(f_0)$. However, they focus on distinct perspectives. Theorem 3 of Lopez-Paz et al. (2015) presents a high probability upper bound for $R_\varphi(\hat{f}) - R_\varphi(f_0)$, whereas our result establishes an upper bound on its expectation. In addition, Theorem 1 (b) further strengthenes the result from the asymptotic viewpoint and shows that as n and m grow sufficiently large, the expected difference $R_\varphi(\hat{f}) - R_\varphi(f_0)$ approaches 0, demonstrating convergence. Furthermore, considering the excess risk $R(\hat{f}) - R_0$, Theorem 1 (c) identifies an upper bound for its expectation, both nonasymptotically and asymptotically. Notably, we present Theorem 1 (d) to connect $\mathbb{E}\{R(\hat{f}) - R_0\}$ with $\mathbb{E}\{R_\varphi(\hat{f}) - R_\varphi(f_0)\}$ through a strictly increasing, nonnegative, continuous and convex function ψ_φ. This connection offers us a guideline in choosing a suitable φ-surrogate function. When φ is chosen as a classification-calibrated convex surrogate, ψ_φ has desirable mathematical properties, as reflected by that $\lim_{n\to\infty} \lim_{m\to\infty} \mathbb{E}\{R_\varphi(\hat{f}) - R_\varphi(f_0)\} = 0$ implies $\lim_{n\to\infty} \lim_{m\to\infty} \mathbb{E}\{R(\hat{f}) - R_0\} = 0$. All these results offer multiple angles to describe how φ-surrogate functions may behave in comparison with the original 0-1 loss, which are, however, not covered in

Lopez-Paz et al. (2015).

Condition (R1) in Theorem 1 is commonly imposed on classifiers in machine learning contexts (e.g., Gouk et al. 2021). This condition can be easily met in practical cases, such as settings where $\mathcal{H}_k$ degenerates to the Euclidean space and $\mathcal{F}$ is specified as the set of linear functions with bounded coefficients. With a continuous $\varphi(\cdot)$ function, condition (R2) is met by considering the class $\mathcal{F}$ in which $|f(h)|$ is bounded by a common constant for all $f \in \mathcal{F}$. This follows from the property that any continuous function is bounded over a bounded closed set in $\mathbb{R}$. Condition (R3) holds for practically used loss functions such as the logistic loss and hinge loss, as shown in Section S2.4 of the supplementary material. Condition (R4) is satisfied by widely-used kernel functions such as the Gaussian kernel.

Theorem 1 describes the φ-risk for the minimizer $\hat{f}$ in (3.5) relative to the φ-risk for the minimizer f_0 in (3.4). More broadly, one may examine the φ-risk for any g in $\mathcal{F}$ relative to $R_\varphi(f_0)$ through the difference of g from $\hat{f}$, as shown in the following theorem whose proof is included in Section S1.4 of the supplementary material.

Theorem 2. *Assume the conditions of Theorem 1. Let* $g : \mathcal{H}_k \to \mathbb{R}$ *denote* *any measurable functional in $\mathcal{F}$, and let*

$$F(\hat{f}, g, L_\varphi) = \mathbb{E}\Big(L_\varphi \sup_{x\in\mathcal{H}_k} |g(x) - \hat{f}(x)|\Big).$$

Then the following results hold:

(a). $\mathbb{E}\big\{R_\varphi(g) - R_\varphi(f_0)\big\} \le C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{4B}{n} + F(\hat{f}, g, L_\varphi)$

(b). $\mathbb{E}\big\{R_\varphi(g) - R_\varphi(f_0)\big\} \le \limsup_{n\to\infty} \limsup_{m\to\infty} F(\hat{f}, g, L_\varphi)$

(c). If φ is classification-calibrated and $\inf_{h\in\mathcal{G}} R_\varphi(h) = \min_{f\in\mathcal{F}} R_\varphi(f) = R_\varphi(f_0)$, then

$$\mathbb{E}\{R(g) - R_0\} \le \zeta_\varphi\left[C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{4B}{n} + F(\hat{f}, g, L_\varphi)\right]$$

and

$$\mathbb{E}\{R(g) - R_0\} \le \zeta_\varphi\Big(\limsup_{n\to\infty} \limsup_{m\to\infty} F(\hat{f}, g, L_\varphi)\Big), \qquad (3.8)$$

where $\zeta_\varphi(\cdot)$ is as in Theorem 1.

(d). If $\inf_{h\in\mathcal{G}} R_\varphi(h) = \min_{f\in\mathcal{F}} R_\varphi(f) = R_\varphi(f_0)$, then

$$\psi_\varphi\Big(\mathbb{E}\{R(g) - R_0\}\Big) \le \mathbb{E}\{R_\varphi(g) - R_\varphi(f_0)\},$$

where ψ_φ is introduced in Theorem 1 (d).

Instead of comparing the minimizer $\hat{f}$ in (3.5) with the minimizer f_0 in (3.4), Theorem 2 extends Theorem 1 by comparing f_0 with any functional g in $\mathcal{F}$. The upper bound in Theorem 2 (a) retains the term $C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)$ from Theorem 1 (a) but extends the term $\frac{2B}{n}$ in Theorem 1 (a) to $\frac{4B}{n}$ in Theorem 2 (a), in addition to the inclusion of an extra term $F(\hat{f}, g, L_\varphi)$ to account for the comparison with an arbitrary functional g rather than just $\hat{f}$.

4. Impact of Mismeasured Output

Theorems 1 and 2 apply only for the case where the true labels l_i are available. Here we consider the setting where the true label l_i is unavailable but its observed version, denoted by $l_i^* \in \mathcal{L}$, is available for $i = 1, \cdots, n$.

To facilitate the relationship between l_i^* and l_i, one may consider

$$p_a^* \triangleq \mathbb{P}(l_i^* = a | \mathcal{S}_i, l_i = a) \quad \text{for } a = -1 \text{ or } 1, \tag{4.9}$$

which is often combined with the assumption that $\mathbb{P}(l_i^* = a|\mathcal{S}_i, l_i = a) = \mathbb{P}(l_i^* = a|l_i = a)$, also called *instance-independent label noise*, as done in this paper. Alternatively, swapping l_i^* and l_i in (4.9) gives

$$p_a \triangleq \mathbb{P}(l_i = a | \mathcal{S}_i, l_i^* = a) \quad \text{for } a = -1 \text{ or } 1, \tag{4.10}$$

for which one may assume that $\mathbb{P}(l_i = a|\mathcal{S}_i, l_i^* = a) = \mathbb{P}(l_i = a|l_i^* = a)$. Both (4.9) and (4.10) can equally describe the degrees of mislabeling, and they are called the (mis)classification and reclassification probabilities (Yi 2017, p.70).

Now we study the impact of mislabeling with either (4.9) or (4.10) used. To highlight the ideas, we assume that p^*_{-1} and p^*_1 (or p_{-1} and p_1) are known for now. The extension to accommodating scenarios with unknown misclassifications is included in the last section. Different from Section 3 with $\big\{\{\mathcal{S}_i, l_i\} \bigm| i = 1, \cdots, n\big\}$ available, here only the error-prone measurements $\big\{\{\mathcal{S}_i, l^*_i\} \bigm| i = 1, \cdots, n\big\}$ are accessible, with $\big\{\{P_i, l^*_i\} \bigm| i = 1, \cdots, n\big\}$ being i.i.d. following the distribution, denoted $\mathcal{M}^*$ on $\mathcal{P} \times \mathcal{L}$. Similar to the discussion in Section 3, let $\mathcal{M}^*_k$ denote the measure on $\mu_k(\mathcal{P}) \times \mathcal{L}$ induced from $\mathcal{M}^*$, then $\big\{\{\mu_k(P_i), l^*_i\} \bigm| i = 1, \cdots, n\big\}$ is a sequence of i.i.d copies from $\mathcal{M}^*_k$.

It may be tempting to train the classifier using the same process discussed in Section 2 by replacing l_i with l^*_i, i.e., use the error-prone samples $\big\{\{\mu_k(P_{s_i}), l^*_i\} \bigm| i = 1, \cdots, n\big\}$ for Step 3 in Section 2.2. We call such a trained classifier the *naive classifier*, given by

$$\hat{f}^* = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}^*_\varphi(f), \tag{4.11}$$

where $\hat{R}^*_\varphi(f) \triangleq \frac{1}{n} \sum_{i=1}^n \varphi(-l^*_i f(\mu_k(P_{s_i})))$ is a naive version of $\hat{R}_\varphi(f)$ in (3.5).

Let D denote the total degree of misclassification in the label, given by

$$D = \begin{cases} 2 - p_{-1}^* - p_1^*, & \text{if (4.9) is taken;} \\ 2 - p_{-1} - p_1, & \text{if (4.10) is taken.} \end{cases}$$

Theorem 3. *Assume the conditions in Theorem 1 and the following conditions:*

(R5). All elements in $\mathcal{F}$ are uniformly bounded. That is, there exists a constant $M > 0$ such that $|f(h)| \leq M||h||_{\mathcal{H}_k}$ for any $f \in \mathcal{F}$ and $h \in \mathcal{H}_k$;

(R6). There exists a constant $A > 0$ such that $k(z_1, z_2) \leq A$ for any $z_1, z_2 \in \mathcal{Z}$.

Then the following results hold:

(a). for any given data size n,

$$\mathbb{E}\{|R_\varphi(\hat{f}^*) - R_\varphi(\hat{f})|\} \leq C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{4B}{n} + 4ML_\varphi AD, \quad (4.12)$$

Furthermore, if φ is classification-calibrated and $\inf_{h \in \mathcal{G}} R_\varphi(h) = \min_{f \in \mathcal{F}} R_\varphi(f) = R(f_0)$, then

$$\mathbb{E}\{|R(\hat{f}^*) - R(\hat{f})|\} \leq 2\zeta_\varphi\Bigg(C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{4B}{n} + 2ML_\varphi AD\Bigg),$$

where $\zeta_\varphi(\cdot)$ is as in Theorem 1.

(b).

$$\limsup_{n\to\infty}\limsup_{m\to\infty}\mathbb{E}\{|R_\varphi(\hat{f}^*)-R_\varphi(\hat{f})|\}\le 4ML_\varphi AD. \tag{4.13}$$

Furthermore, if φ is classification-calibrated and $\inf_{h\in\mathcal{G}} R_\varphi(h)=\min_{f\in\mathcal{F}} R_\varphi(f)=R(f_0)$, then

$$\limsup_{n\to\infty}\limsup_{m\to\infty}\mathbb{E}\{|R(\hat{f}^*)-R(\hat{f})|\}\le 2\zeta_\varphi(2ML_\varphi AD),$$

where $\zeta_\varphi(\cdot)$ is as in Theorem 1.

The proof of this theorem is presented in Section S1.5 of the supplementary material. Conditions (R5) and (R6) share similarities to those in Theorem 1. When $f(0)=0$ for all $f\in\mathcal{F}$, condition (R1) in Theorem 1 implies condition (R5) in Theorem 3. If A in condition (R6) equals 1, condition (R4) in Theorem 1 evidently holds. Notably, Theorem 3 suggests that the empirical φ-risk derived from the naive classifier cannot indefinitely differ from that of the correct classifier. It describes upper bounds for the expected excess φ-risk $|R_\varphi(\hat{f}^*)-R_\varphi(\hat{f})|$ and the risk $\left|R(\hat{f}^*)-R(\hat{f})\right|$ for the naive classifier $\hat{f}^*$ in two different manners, nonasymptotically and asymptotically. Although the upper bound (4.12) is not necessarily sharp, it carries important implications. This bound is the sum of the asymptotic bound in (4.13) and $C(n,m,L_\varphi,L_\mathcal{F},B)+4Bn^{-1}$, where the latter term reflects the influence of the size n of data and the *Rademacher complexity* of $\mathcal{F}$. As $n \to \infty$ and $m \to \infty$, $C(n, m, L_\varphi, L_\mathcal{F}, B) \to 0$, and thus, Theorem 3 (a) leads to Theorem 3 (b). Further, applying Jensen's inequality to Theorem 3 (a) gives that

$$\left|\mathbb{E}\{R_\varphi(\hat{f}^*) - R_\varphi(\hat{f})\}\right| \leq C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big) + \frac{4B}{n} + 4ML_\varphi AD, \quad (4.14)$$

which characterizes a range for the difference between $R_\varphi(\hat{f}^*)$ and $R_\varphi(\hat{f})$ under finite settings, influenced by various factors such as M, L_φ, A, B, $R(\mathcal{F})$, and the total degree D of label misclassification. Theorem 3 (b) suggests that with a small degree of label noise, the upper bound (4.13) is close to zero, showing the practical utility of the naive classifier. Under such circumstances, even in the absence of precise measurements, using error-contaminated data can still aid in learning f_0 by increasing sample sizes m_i or n.

5. Correcting Mislabeling Effects

To correct mislabeling effects, we propose a new surrogate function by modifying the initial surrogate function φ introduced in Section 3 defined for true labels. For any $t \in \mathbb{R}$ and $l^* \in \mathcal{L}$, we define

$$\varphi^*(t, l^*) = \begin{cases} \frac{p^*_{-l^*}\varphi(-tl^*) - (1-p^*_{l^*})\varphi(tl^*)}{p^*_1 + p^*_{-1} - 1}, & \text{if (4.9) is taken;} \\ \varphi(-tl^*)p_{l^*} + \varphi(tl^*)(1 - p_{l^*}), & \text{if (4.10) is taken,} \end{cases} \quad (5.15)$$

and similar to (3.4), we define the φ^*-risk as

$$R_{\varphi^*}(f) \triangleq \mathbb{E}\{\varphi^*(f(\mu_k(P)), l^*)\}, \tag{5.16}$$

where the expectation is evaluated with respect to the joint distribution $\mathcal{M}_k^*$ of $\{\mu_k(P), l^*\}$, and f is a functional from $\mathcal{H}_k$ to $\mathbb{R}$.

By incorporating the misclassification probabilities p_1^* and p_{-1}^*, or the reclassification probabilities p_1 and p_{-1}, into the modified surrogate function $\varphi^*(\cdot, \cdot)$, we effectively mitigate the mislabeling effects. The adjustment ensures our original objective of minimizing the φ-risk to be preserved by minimizing the φ^*-risk, as demonstrated by the following Theorem 4, whose proof is presented in Section S1.6 of the supplementary material. Importantly, this new surrogate function $\varphi^*(t, l^*)$ can be directly applied to identify the optimal learner using the observed noisy labels.

Theorem 4. *For any $f \in \mathcal{F}$, we have that*

$$R_{\varphi^*}(f) = R_\varphi(f),$$

where $R_{\varphi^*}(f)$ and $R_\varphi(f)$ are defined in (5.16) and (3.4), respectively.

Similar to $\hat{R}_\varphi(\cdot)$ in (3.5), we define

$$\hat{R}_{\varphi^*}(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \varphi^*(f(\mu_k(P_{s_i})), l_i^*), \tag{5.17}$$

and determine the classifier based on using error-corrupted data:

$$\hat{f}^{correct} = \text{argmin}_{f \in \mathcal{F}} \hat{R}_{\varphi^*}(f). \tag{5.18}$$

When applying (4.9), φ^* in (5.15) may be a nonconvex function with respect to t due to the negative coefficient of $\varphi(tl^*)$, leading to a nonconvex optimization problem in (5.18). While nonconvex optimization presents a computational challenge, it is commonly encountered in classification tasks. In this case, a widely used approach is to relax nonconvex problems to be convex ones, similar to the idea of replacing the 0-1 loss function with a convex surrogate loss, as discussed in Section 3. Alternatively, one can directly solve nonconvex optimization problems using techniques such as projected gradient descent, alternating minimization, and stochastic optimization algorithms (Jain and Kar 2017).

Let

$$L_\varphi^* = \begin{cases} \frac{2L_\varphi}{|1-p_1^*-p_{-1}^*|}, & \text{if (4.9) is taken;} \\ L_\varphi, & \text{if (4.10) is taken.} \end{cases} \tag{5.19}$$

and

$$B^* = \begin{cases} \frac{2B}{|1-p_1^*-p_{-1}^*|}, & \text{if (4.9) is taken;} \\ B, & \text{if (4.10) is taken.} \end{cases} \tag{5.20}$$

Theorem 5. *Assume that the conditions of Theorem 1 hold. Then for* $\hat{f}^{correct}$ *in (5.18) and* f_0 *in (3.4),* *(a).* $0 \leq \mathbb{E}\{R_\varphi(\hat{f}^{correct}) - R_\varphi(f_0)\} \leq C\Big(n, m, L^*_\varphi, L_\mathcal{F}, B^*\Big) + \frac{2B^*}{n}$;

(b). $\lim_{n\to\infty}\lim_{m\to\infty} \mathbb{E}\{R_\varphi(\hat{f}^{correct}) - R_\varphi(f_0)\} = 0$;

(c). if φ *is classification-calibrated and* $\inf_{h\in\mathcal{G}} R_\varphi(h) = \min_{f\in\mathcal{F}} R_\varphi(f) = R(f_0)$, *then*

(i) $0 \leq \mathbb{E}\{R(\hat{f}^{correct}) - R_0\} \leq \zeta_\varphi\Big(C\Big(n, m, L^*_\varphi, L_\mathcal{F}, B^*\Big) + \frac{2B^*}{n}\Big)$, *where* $\zeta_\varphi(\cdot)$ *is as in Theorem 1.*

(ii) $\lim_{n\to\infty}\lim_{m\to\infty} \mathbb{E}\{R(\hat{f}^{correct}) - R_0\} = 0$.

The proof of Theorem 5 is presented in Section S1.7 of the supplementary material. The theorem states that $\mathbb{E}\{R_\varphi(\hat{f}^{correct}) - R_\varphi(f_0)\}$ and $\mathbb{E}\{R(\hat{f}^{correct}) - R(f_0)\}$ converge to zero as the sample sizes m and n approach infinity, which align with the convergence of $\mathbb{E}\{R_\varphi(\hat{f}) - R_\varphi(f_0)\}$ and $\mathbb{E}\{R(\hat{f}) - R_0\}$, respectively, as shown in Theorem 1. That is, like the empirically optimal classifier $\hat{f}$ obtained from precise measurements, the corrected classifier $\hat{f}^{correct}$ obtained from mismeasured data is asymptotically consistent for φ-risk in expectation.

We further comment on the performance of the classifier $\hat{f}^{correct}$. Relative to the classifier $\hat{f}$ trained from clean data, Theorem 5 is a counterpart of Theorem 1 (a)-(c), which incorporates the label noise effects through L^*_φ

and B^*. The upper bounds established for $\hat{f}^{correct}$ are identical to those for $\hat{f}$ when model (4.10) is used, but larger than those for $\hat{f}$ when model (4.9) is used, potentially indicating the price paid to train a valid classifier using noisy data relative to clean data. On the other hand, regarding the naive classifier $\hat{f}^*$ trained from noisy data without accounting for the label noise effects, though Theorems 3 and 5 do not compare $\hat{f}^*$ and $\hat{f}^{correct}$ relative to the same reference classifier, it is interesting to compare the upper bounds they identify. Specifically, comparing the upper bound in (4.12) and Theorem 5 (a), the resulting difference is

$$D_\varphi \triangleq 4\{\mathcal{R}(\mathcal{F}) + \frac{1}{n}L_{\mathcal{F}}\}(L_\varphi - L_\varphi^*) + 2\sqrt{\frac{\log\,(2n)}{n}}(B - B^*) + \frac{4B}{n} - \frac{2B^*}{n} \\ + 4ML_\varphi AD.$$

When model (4.10) is used, $D_\varphi = 4ML_\varphi AD + \frac{2B}{n}$, indicating that the upper bound for the classifier $\hat{f}^{correct}$ in Theorem 5 (a) is $4ML_\varphi AD + \frac{2B}{n}$ smaller than that for the naive classifier in (4.12). On the other hand, when model (4.9) is considered, $D_\varphi \leq 4ML_\varphi AD$ when n is large, as other terms in D_φ is close to 0.

The preceding development focuses on classification within the infinite-dimensional RKHS $\mathcal{H}_k$. While this provides a theoretical foundation, practical implementation often requires working within a finite-dimensional approximation of $\mathcal{H}_k$. To this end, we construct a finite-dimensional space that approximates $\mathcal{H}_k$, and provides the detail in Section S2 of the supplementary material, where we devise a classification method to address label noise within the finite-dimensional space approximating $\mathcal{H}_k$ and establish informative upper bounds for the φ-risk of the naive and correction classifiers relative to the true classifier in Theorems S2 and S3 of the supplementary material.

6. Sensitivity Analyses and Proposed Metrics

In this section, we propose assessment metrics to characterize the impact of mislabeling and examine the performance of the proposed correction method by using the SUP3 dataset discussed in Section 1, with the details deferred to Section S3 of the supplementary material. While the provided causal information is deemed to involve mislabeling, there is no validation dataset to quantify the degree of mislabeling. Consequently, we undertake sensitivity analyses to explore the impact of mislabeling and assess the performance of the proposed correction method, which involves examining (4.10) under various assumptions for the values of p_1 and p_{-1}.

6.1 Implementation Details

Causal learning is practically executed by transforming classification in the infinite-dimensional RKHS space $\mathcal{H}_k$ with kernel function k into an r-dimensional vector space that approximates $\mathcal{H}_k$, as also implemented in our study here, where we use the Gaussian kernel function, $k(v_1, v_2) = \exp(-\gamma||v_1 - v_2||_2^2)$, with hyper parameter γ. The parameter r is user-specified; a larger value r leads to a more accurate approximation but entails a higher computational cost. Further details on this approximation method, along with theoretical guarantees when using the Gaussian kernel, are provided in Section S2 of the supplementary material.

To assess the impact of different approximations, we consider different values for r and γ within specified ranges, denoted $[a_r, b_r]$ and $[a_\gamma, b_\gamma]$, respectively. We set $[a_r, b_r] = [100, 1000]$ by evenly dividing it into 10 subintervals and setting r to each of those cutpoint values; we take $[a_\gamma, b_\gamma] = [0.01, 10]$ by dividing it into 10 subintervals with equal length after taking the transformation of logarithm to the base ten and letting γ take each of the cutpoint values, that is, $10^{-2+\frac{j}{3}}$ with $j = 0, 1, \cdots, 9$.

In characterizing different degrees of label noise, we consider model (4.10) and allow p_1 and p_{-1} to take values in an interval, denoted $[a_p, b_p]$, where we set $[a_p, b_p] = [0.5, 1]$ by dividing it into 50 subintervals with equal length and let p_1 and p_{-1} take each of those cutpoint values except $(p_1, p_{-1}) = (0.5, 0.5)$ or $(1, 1)$. Let $\theta = (p_1, p_{-1}, r, \gamma)$.

The sensitivity analyses proceed in the following three steps:

1. With given values of p_1 and p_{-1}, independently generate values of l_i based on the reported value of l_i^* using (4.10) for $i = 1, \cdots, n$.

2. With the specified values for r in (S.49) and γ in (S.1) of the supplementary material, for $i = 1, \cdots, n$, we use the r-dimensional vector $\mu_{k,r}(P_{\mathcal{S}_i})$ discussed in Section S2 of the supplementary material to approximate $\mu_k(P_{\mathcal{S}_i})$ described in Section 2.2.

3. Given a value of θ, we consider three methods of using data, by respectively solving (3.5), (4.11), and (5.18), with $\mu_k(P_{\mathcal{S}_i})$ in $\hat{R}_\varphi(\cdot)$ of (3.5) replaced by $\mu_{k,r}(P_{\mathcal{S}_i})$ that is presented in (S.56) of the supplementary material. We call these as the *true*, *naive*, and *correction* methods, respectively, and for a given classification method, let $\text{sign}(f_\theta)$, $\text{sign}(f_\theta^*)$, and $\text{sign}(f_\theta^{correct})$ denote the *true*, *naive*, and *correction* classifiers, respectively, where f_θ, f_θ^*, and $f_\theta^{correct}$ represent the corresponding discriminant functions from $\mathbb{R}^r$ to $\mathbb{R}$ for an employed classification method.

   Here we consider two classification methods: logistic regression (LR) and Gaussian kernel-based support vector machine (SVM). In the LR method, we specify the convex surrogate function $\varphi(\cdot)$ to be $\varphi(u) = \log_2\big\{1 + \exp(u)\big\}$ for the logistic loss, and take the class $\mathcal{F}$ as $\mathcal{F}_r \triangleq \Big\{f \;\Big|\; f(x) = w^{\mathrm{T}}x + c,$ with $w \in \mathbb{R}^r$ and $c \in \mathbb{R}$ satisfying $||w||_2^2 \leq C_r$ and $|c| \leq C_r\Big\}$. For the SVM method, we set the convex surrogate function $\varphi(\cdot)$ to be $\varphi(u) = \max\{1, 1+u\}$ for the *hinge loss*, and let $\mathcal{F}_r \triangleq \Big\{f \;\Big|\; f(x) = \sum_{i=1}^{n} \alpha_i l_i k(\mu_{k,r}(P_{\mathcal{S}_i}), x) + b,$ with $|\alpha_i| \leq C_r$ for $i = 1, \cdots, n$ and $|b| \leq C_r\Big\}$. Here, C_r is a large constant, and k represents the Gaussian kernel (S.1) with $\gamma = 1$ (Section 6.3, Mohri, Rostamizadeh, and Talwalkar 2018), i.e., $k(z, z') = \exp(-||z - z'||_2^2)$. We employ the gradient decent (GD) method (Boyd and Vandenberghe 2004) to train a classifier.

When the convex surrogate φ is chosen for the logistic or hinge loss, and the class $\mathcal{F}$ of functionals is set to $\mathcal{F}_r$, we show in Section S2.4 of the supplementary material that the conditions of Theorem S3 are satisfied. Consequently, the theoretical results in Theorem S3 apply to the *correction* classifier $\mathrm{sign}(f_\theta^{correct})$.

6.2 Evaluation Metrics and Results

We compute the *accuracy* and *recall* of true classifier $\text{sign}(f_\theta)$, respectively given by $A(\theta) = 1 - \frac{\sum_{i=1}^{n} |l_i - \hat{l}_i|}{2n}$ and $R(\theta) = 1 - \frac{\sum_{i=1}^{n} I\{l_i=1\}|l_i - \hat{l}_i|}{2\sum_{i=1}^{n} I\{l_i=1\}}$, where $\hat{l}_i$ represents the predicted value for l_i using classifier $\text{sign}(f_\theta)$. Similarly, $A^*(\theta)$ and $R^*(\theta)$ are defined for the naive classifier $\text{sign}(f_\theta^*)$, and $A^{correct}(\theta)$ and $R^{correct}(\theta)$ are defined for corrected classifier $\text{sign}(f_\theta^{correct})$.

To quantify the mislabeling effects and assess the performance of the proposed correction method, we define

$$D_A(\theta) \triangleq A(\theta) - A^*(\theta) \quad \text{and} \quad D_R(\theta) \triangleq R(\theta) - R^*(\theta),$$

referred to as *accuracy-bias* and *recall-bias*, respectively, along with

$$D_A^{correct}(\theta) \triangleq A(\theta) - A^{correct}(\theta) \quad \text{and} \quad D_R^{correct}(\theta) \triangleq R(\theta) - R^{correct}(\theta),$$

termed *accuracy-correction* and *recall-correction*, respectively.

A large value of $D_A(\theta)$ or $D_R(\theta)$ indicates a substantial mislabeling effect, and a large value of $D_A^{correct}(\theta)$ or $D_R^{correct}(\theta)$ indicates a poor performance of the proposed correction method for a given value of θ.

To see how these measures vary with the degree of mislabeling, we divide $[0.5, 1]$ into N equal length subintervals with the cutpoints $0.5 = a_0 < a_1 < \cdots < a_{N-1} < a_N = 1$, and calculate these measures for $\theta = (a_i, a_j, r, \gamma)$ with $i, j = 1, \cdots, N$. To provide a comprehensive view, we construct a heatmap for $D_A(p_1, p_{-1}, r, \gamma)$, $D_R(p_1, p_{-1}, r, \gamma)$, $D_A^{correct}(p_1, p_{-1}, r, \gamma)$, and $D_R^{correct}(p_1, p_{-1}, r, \gamma)$ with given values of r and γ, where p_1 and p_{-1} take values of a_i and a_j for $i, j = 1, \cdots, N$, respectively, excluding $(p_1, p_{-1}) = (0.5, 0.5)$ or $(1, 1)$.

Furthermore, to assess the influence by r and γ, we calculate $T_X(N, r, \gamma) \triangleq \sum_{i=1}^{N}\sum_{j=1}^{N} D_X(a_i, a_j, r, \gamma)$; and $T_X^{correct}(N, r, \gamma) \triangleq \sum_{i=1}^{N}\sum_{j=1}^{N} D_X^{correct}(a_i, a_j, r, \gamma)$, with "$X$" representing "A" or "R". These metrics reflect the overall performance of the naive or proposed correction method in terms of accuracy and recall.

In our sensitivity analyses, we take $N = 50$, and display heatmaps for $D_X(p_1, p_{-1}, 500, 3)$ and $D_X^{correct}(p_1, p_{-1}, 500, 3)$ in the first and last two columns in Figure 1, respectively, where "X" represents "A" or "R". Clearly, $D_A(\theta)$ and $D_R(\theta)$ differs from zero for nearly all values of p_1 and p_{-1}, showing the existence of mismeasurement effects. As expected, such effects become more substantial as the degree of mislabeling increases regardless of whether the LR and SVM classifiers are used, although the impact varies with the classifier used. The proposed correction method outperforms the naive method in terms of accuracy and recall for both the LR and SVM classifiers.

(*insert Figure 1 about here*)

To assess how the mislabeling effects and the performance of the proposed correction method vary with r and γ, we consider $r = 100, 500$, or 1000, and $\gamma = 0.01, 0.1, 1, 3$, or 10, and report in Table 1 the results of $T_X(50, r, \gamma)$, and $T_X^{correct}(50, r, \gamma)$ obtained from the logistic regression and SVM classifiers, where "X" stands for "A" or "R". Additional results are reported in Figure S.1 of the supplementary material. Clearly, the mislabeling effects may be differently exhibited by different choices of a classifier. The choice of r and γ can impact the performance of both the naive and proposed methods. Overall, the proposed correction method outperforms the naive methods in all settings of r and γ.

(*insert Table 1 about here*)

7. Discussion

In this paper, we cast causal inference as a binary classification problem as in Lopez-Paz et al. (2015) but extend their framework to handle label noise. Further exploration can be considered for determining causal relationships for paired variables, which is inherently complex and contingent upon specific contexts. While one may consider all possible grouping combinations as noted in Section 1, this process, however, entails a myriad of possibilities when the number of variables is moderate or large.

Refining structures to better facilitate relationships among variables is an intriguing prospect. Instead of simply examining causal links between two vectors X_i and W_i, we may pool all components in X_i and W_i and use a directed acyclic graph (DAG) to represent causal relationships, where nodes represent variables and edges denote causal directions. We may also explore directed random graphs, where edge existence and direction are probabilistic. Labeling causal relationships would then involve probability components.

Our focus here is on settings with instance-independent label noise, also known as nondifferential response error, where all units have an equal probability of being mislabeled. In cases where subjects have varying probabilities of being mislabeled, we can refine our approach by forming two sets: one for subjects without label noise (using the usual classifier) and one with label noise (using the developed procedure).

As commented by a referee, when predicting labels for a new pair of variables, $(\tilde{X}, \tilde{W})$, with a sample of measurements, $\tilde{\mathcal{S}} \triangleq \{(\tilde{X}_k, \tilde{W}_k) \mid k = 1, \cdots, \tilde{m}\}$, it may be interesting to include the new data $\tilde{S}$ to the original dataset to retrain the classifier for possible performance enhancement. Techniques of handling missing outcomes may be useful in this regard.

Our development assumes knowledge of misclassification probabilities p^*_{-1} and p^*_1 (or p_{-1} and p_1), typically used in sensitivity analyses to assess classifier performance under varying degrees of label noise. Extending our method to handle unknown misclassifications is interesting. This extension can be achieved by utilizing validation data with measurements for both true labels and their surrogate versions and using a two-stage procedure: in the first stage, estimate misclassification probabilities using validation data, and in the second stage, apply our approach using these estimates.

Without validation data, an alternative is to construct a new loss function independent of misclassification probabilities. Using the minimax technique, we maximize the empirical φ^*-risk (5.17) with respect to misclassification probabilities p^*_{-1} and p^*_1 (or p_{-1} and p_1) over a user-specified set $\mathcal{B}$, and minimize this with respect to the classifier f over the class $\mathcal{F}$ of candidate classifiers. Ideally, $\mathcal{B}$ would contain the true misclassification probabilities, with a smaller $\mathcal{B}$ leading to better classifier performance.

Acknowledgments

Yi is the Canada Research Chair in Data Science (Tier 1). Her research was supported by the Canada Research Chairs Program and the National Sciences and Engineering Research Council of Canada (NSERC).

References

Armstrong, M. A. (1983). *Basic Topology.* New York: Springer.

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association 101*(473), 138–156.

Boyd, S. P. and L. Vandenberghe (2004). *Convex Optimization.* Cambridge University Press.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: a Modern Perspective.* Chapman and Hall/CRC.

Conway, J. B. (2019). *A Course in Functional Analysis.* New York: Springer.

Gouk, H., E. Frank, B. Pfahringer, and M. J. Cree (2021). Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning 110*, 393–416.

Guo, H., B. Wang, and G. Yi (2023). Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. In *Advances in Neural Information Processing Systems*, Volume 36, pp. 347–386.

Guo, H., G. Y. Yi, and B. Wang (2024). Learning from noisy labels via conditional distributionally robust optimization. In *Advances in Neural Information Processing Systems*, Volume 37, pp. 82627–82672.

Guyon, I. (2013). Cause-effect pairs kaggle competition, SUP1 data. `https://www.kaggle.com/c/cause-effect-pairs/data`.

Jain, P. and P. Kar (2017). Non-convex optimization for machine learning. *Foundations and*

Trends® in Machine Learning 10(3-4), 142–363.

Lopez-Paz, D., K. Muandet, B. Schölkopf, and I. Tolstikhin (2015). Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pp. 1452–1461.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018). *Foundations of Machine Learning*. MIT press.

Monti, R. P., K. Zhang, and A. Hyvärinen (2020). Causal discovery with general non-linear relationships using non-linear ICA. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pp. 186–195.

Mooij, J. M., J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research 17*(1), 1103–1204.

Muandet, K., K. Fukumizu, B. Sriperumbudur, and B. Schölkopf (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning 10*(1-2), 1–141.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science 5*(4), 465–480.

Potter, J. and I. J. Higginson (2004). Pain experienced by lung cancer patients: a review of prevalence, causes and pathophysiology. *Lung Cancer 43*(3), 247–257.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Szabó, Z., B. K. Sriperumbudur, B. Póczos, and A. Gretton (2016). Learning theory for distribution regression. *The Journal of Machine Learning Research 17*(1), 5272 – 5311.

Tagasovska, N., V. Chavez-Demoulin, and T. Vatter (2020). Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *Proceedings of the 37th International Conference on Machine Learning*, Volume 119, pp. 9311–9323.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. New York: Springer.

Yi, G. Y., A. Delaigle, and P. Gustafson (2021). *Handbook of Measurement Error Models*. CRC Press.

Zhou, Z.-H. and J.-M. Xu (2007). On the relation between multi-instance learning and semi-supervised learning. In *International Conference on Machine Learning*, pp. 1167–1174.

Table 1: *Sensitivity analyses of the SUP3 data: assessing the impact of different choices of r and γ on accuracy and recall*

γ	$T_A(50, 100, \gamma)$		$T_A(50, 500, \gamma)$		$T_A(50, 1000, \gamma)$		$T_R(50, 100, \gamma)$		$T_R(50, 500, \gamma)$		$T_R(50, 1000, \gamma)$	
	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM
0.01	263	330	570	597	612	599	861	827	1074	1082	1110	1076
0.1	299	377	563	600	640	600	895	948	1087	1096	1154	1068
1	332	403	596	598	617	602	949	995	1196	1076	1108	1072
3	307	401	591	603	635	600	939	1071	1148	1092	1151	1075
10	339	408	575	598	640	602	1004	1021	1108	1075	1165	1075
γ	$T_A^{correct}(50, 100, \gamma)$		$T_A^{correct}(50, 500, \gamma)$		$T_A^{correct}(50, 1000, \gamma)$		$T_R^{correct}(50, 100, \gamma)$		$T_R^{correct}(50, 500, \gamma)$		$T_R^{correct}(50, 1000, \gamma)$	
	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM
0.01	22	209	405	199	557	39	126	122	659	402	926	55
0.1	44	299	427	204	590	49	135	778	722	396	987	75
1	128	321	444	207	590	46	310	3	749	416	989	79
3	108	281	436	222	604	45	343	548	720	443	1031	78
10	99	357	455	202	585	46	252	646	759	388	997	77

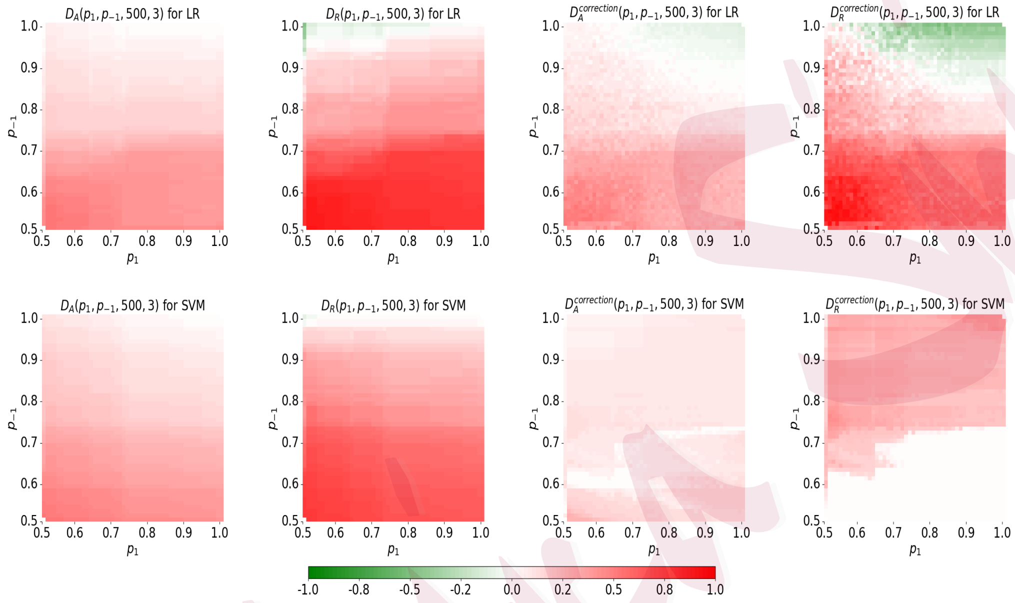

Figure 1: *Heatmaps generated from a naive method for* $D_A(p_1, p_{-1}, r, \gamma)$ *and* $D_R(p_1, p_{-1}, r, \gamma)$ *and the proposed correction method for* $D_A^{correct}(p_1, p_{-1}, r, \gamma)$ *and* $D_R^{correct}(p_1, p_{-1}, r, \gamma)$*, where the results for LR and SVM classifiers are reported in the top and bottom panels, respectively.*