Supplementary Material for

"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, Canada*

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

[*] *the corresponding author*

This supplementary material includes new theorems, technical derivations, numerical studies, and additional material for the manuscript entitled above. In Section S1, we describe the *kernel mean embeddings of probability distributions* and provide the proofs of all the theorems presented in the main text. In Section S2, we describe the use of a finite-dimensional space approximation to the infinite-dimensional RKHS to save computation costs, and then theoretically investigate the impact of ignoring mislabeling of outcomes and the performance of the proposed correction method in this finite-dimensional approximate space. In Section S3, we describe the SUP3 dataset in details and report additional numerical results to the sensitivity analyses presented in Section 6.2 of the main text. In Section S4, we report simulation studies to assess the impact of ignoring mislabeling of outcomes and the performance of the proposed correction method.

S1 Kernel Mean Embedding of Probability Distributions and Proofs of Theorems in the Main Text

S1.1 Empirical Kernel Mean Embedding

For completeness, here we describe basics about *kernel mean embeddings of probability distributions.* Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ denote a continuous, positive-definite kernel function, and let $\mathcal{H}_k$ denote the *reproducing kernel Hilbert space* (RKHS) induced from the kernel k; all elements in $\mathcal{H}_k$ are functions from $\mathcal{Z}$ to $\mathbb{R}$. A common choice of the kernel function k is the Gaussian kernel function

$$k(v_1, v_2) = \exp(-\gamma ||v_1 - v_2||_2^2), \tag{S.1}$$

For details, see Muandet et al. (2017).

As in Lopez-Paz et al. (2015), consider a separable topological space $(\mathcal{Z}, \tau_z)$, where τ_z represents the topology on the set $\mathcal{Z}$ (Armstrong 1983). Suppose that Q is the probability distribution of a random variable V, defined as a function from a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ to a measurable space $(\mathcal{Z}, \sigma(\tau_z))$, where Ω is the sample space, $\mathcal{E}$ represents a σ-algebra in Ω, $\mathbb{P}$ denotes a probability measure, and $\sigma(\tau_z)$ is the σ-algebra generated by τ_z. The kernel mean embedding method is to project the probability distribution Q into RKHS $\mathcal{H}_k$ via a mapping, denoted μ_k. Specifically, let $\mu_k(Q)$ denote the corresponding element in $\mathcal{H}_k$, which is a function mapping from $\mathcal{Z}$ to $\mathbb{R}$, defined by

$$\mu_k(Q)(x) \triangleq \int_{z \in \mathcal{Z}} k(z, x) dQ(z) \quad \text{for any } x \in \mathcal{Z}, \tag{S.2}$$

where the integral refers to the Bochner integral (e.g., Diestel and Uhl 1977, Chapter 2).

Evaluation of (S.2) may be difficult, and often, we consider its *empirical kernel mean embedding* by approximating Q with an empirical distribution. Specifically, let $S \triangleq \{V_j \mid j = 1, \cdots, J\}$ denote a sequence of independent and identically distributed (i.i.d) variables from the probability space $(\Omega, \mathcal{E}, \mathbb{P})$ to the measurable space $(\mathcal{Z}, \sigma(\tau_z))$, generated from the probability distribution Q, where J is a large positive integer specified by users. Define

$$Q_S \triangleq \frac{1}{J} \sum_{j=1}^{J} \delta_{(V_j)},$$

where $\delta_{(x_0)}$ represents a Dirac measure centered at x_0, defined by $\delta_{(x_0)}(A^*) = 1$ if $x_0 \in A^*$; and $\delta_{(x_0)}(A^*) = 0$ if $x_0 \notin A^*$. Here, x_0 is a point in $\mathbb{R}$ and A^* is any subset of $\mathbb{R}$.

Consequently, an *empirical kernel mean embedding* for Q is defined as the function $\mu_k(Q_S) : \mathcal{Z} \to \mathbb{R}$, given by

$$\mu_k(Q_S)(x) = \frac{1}{J}\sum_{j=1}^{J} k(V_j, x) \quad \text{for } x \in \mathcal{Z}, \tag{S.3}$$

which, for each argument $x \in \mathcal{Z}$, is essentially a random variable from $(\Omega, \mathcal{E}, \mathbb{P})$ to $\mathbb{R}$ due to the randomness induced from the V_j in S; when the sample S is realized as s, the resultant $\mu_k(Q_s)$ becomes a deterministic function.

S1.2 Convergence in Mean

To make the empirical kernel mean embedding (S.3) useful to approximate (S.2), we want to identify conditions to ensure $\mu_k(Q_S)$ to be close to $\mu_k(Q)$ in some sense. For $f \in \mathcal{H}_k$, let $||f||_\infty = \sup_{z \in \mathcal{Z}} |f(z)|$ and $||f||_{\mathcal{H}_k} = \sqrt{< f, f >_{\mathcal{H}_k}}$, where $< \cdot, \cdot >_{\mathcal{H}_k}$ is the inner product in $\mathcal{H}_k$; see Muandet et al. (2017, Chapter 2) for details. Lopez-Paz et al. (2015) showed that under certain conditions, $||\mu_k(Q_S) - \mu_k(Q)||_{\mathcal{H}_k}$ converges in probability to zero as $J \to \infty$. Here, we examine $\mathbb{E}||\mu_k(Q_S) - \mu_k(Q)||_{\mathcal{H}_k}$ and establish the following convergence result.

Theorem S1. *Assume that the kernel function k is bounded on $\mathcal{Z}$ and that* *for any $f \in \mathcal{H}_k$ with $||f||_{\mathcal{H}_k} \leq 1$, $||f||_\infty \leq 1$. Then for $\mu_k(Q_S)$ in (S.3),*

$$\lim_{J\to\infty} \mathbb{E}||\mu_k(Q_S) - \mu_k(Q)||_{\mathcal{H}_k} = 0,$$

where J is the size of S and the expectation is evaluated with respect to the joint distribution of S.

Proof. By Theorem 28 of Song (2008), we have that for any random variable V following the probability distribution Q,

$$||\mu_k(Q_S) - \mu_k(Q)||_{\mathcal{H}_k} = \sup_{||f||_{\mathcal{H}_k}\leq 1} [\mathbb{E}\{f(V)\} - \frac{1}{J}\sum_{j=1}^{J} f(V_j)], \tag{S.4}$$

where the expectation is evaluated with respect to Q. Then by the proof of Theorem 1 in Lopez-Paz et al. (2015), we have that

$$\mathbb{E}\Big\{ \sup_{||f||_{\mathcal{H}_k}\leq 1} [\mathbb{E}\{f(V)\} - \frac{1}{J}\sum_{j=1}^{J} f(V_j)]\Big\} \leq 2\sqrt{\frac{\mathbb{E}[k(V,V)]}{J}}. \tag{S.5}$$

Since $k(z,z)$ is bounded on $\mathcal{Z}$, $\mathbb{E}[k(V,V)] < \infty$. Then

$$\lim_{J\to\infty} \sqrt{\frac{\mathbb{E}[k(V,V)]}{J}} = 0. \tag{S.6}$$

Furthermore, by the definition of norm, $||\mu_k(Q_S) - \mu_k(Q)||_{\mathcal{H}_k} \geq 0$. Therefore, combining (S.4), (S.5) and (S.6) gives that

$$\lim_{J\to\infty} \mathbb{E}\big\{||\mu_k(Q_S) - \mu_k(Q)||_{\mathcal{H}_k}\big\} = 0.$$

□

The conditions in Theorem S1 are satisfied by useful kernel functions, such as the Gaussian kernel in (S.1). Theorem S1 shows the convergence in the mean of $||\mu_k(Q_S) - \mu_k(Q)||_{\mathcal{H}_k}$ to zero as $J \to \infty$, implying that the mean distance between $\mu_k(Q_S)$ and $\mu_k(Q)$ approaches zero as the sample size J of S approaches infinity, where the distance is measured by the norm in Hilbert space $\mathcal{H}_k$. This result is stronger than the convergence in probability of $||\mu_k(Q_S) - \mu_k(Q)||_{\mathcal{H}_k}$ to zero as $J \to \infty$, implied by Theorem 1 of Lopez-Paz et al. (2015).

Remark 1. It is known that convergence in mean implies convergence in probability (Geiss and Geiss 2004, Proposition 4.1.3), but not vice versa.

As a counterexample, let $\mathcal{B}(\mathbb{R})$ denote the Borel σ-algebra on $\mathbb{R}$ (see Definition 1.1.8 of Geiss and Geiss 2004). Consider the probability space $(\Omega, \mathcal{E}, \mathbb{P})$, where $\Omega = [0,1]$, $\mathcal{E} \triangleq \big\{G \cap [0,1] \bigm| G \in \mathcal{B}(\mathbb{R})\big\}$, and $\mathbb{P}$ is the Lebesgue measure on $[0,1]$ (see Definition 1.3.1 of Geiss and Geiss 2004).

For $n = 1, 2, \cdots$, let $p(n)$ denote the unique nonnegative integer such that $n \in [2^{p(n)+1} - 1, 2^{p(n)+2} - 2]$. Define

$$U_n(\omega) = 2^{p(n)+1} I\left\{\omega \in \left[\frac{n - 2^{p(n)+1} + 1}{2^{p(n)+1}}, \frac{n - 2^{p(n)+1} + 2}{2^{p(n)+1}}\right]\right\},$$

where $I\left\{\omega \in \left[\frac{n-2^{p(n)+1}+1}{2^{p(n)+1}}, \frac{n-2^{p(n)+1}+2}{2^{p(n)+1}}\right]\right\}$ is the indicator function representing whether ω is in the interval of $\left[\frac{n-2^{p(n)+1}+1}{2^{p(n)+1}}, \frac{n-2^{p(n)+1}+2}{2^{p(n)+1}}\right]$, taking value 1 if yes and value 0 otherwise. That is, for $n = 1, 2, \cdots$, U_n is a random variable defined over the probability space $(\Omega, \mathcal{E}, \mathbb{P})$.

Then for any $1 > \varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\big\{|U_n| \geq \varepsilon\big\} &\leq \mathbb{P}\bigg\{\omega \in \Big[\frac{n - 2^{p(n)+1} + 1}{2^{p(n)+1}}, \frac{n - 2^{p(n)+1} + 2}{2^{p(n)+1}}\Big]\bigg\} \\ &= \frac{1}{2^{p(n)+1}},\end{aligned}$$

implying that

$$\lim_{n\to\infty} \mathbb{P}\big\{|U_n| \geq \varepsilon\big\} = 0.$$

That is, $\big\{U_n \bigm| n = 1, 2, \cdots\big\}$ converges to zero in probability as n approaches to infinity.

On the other hand,

$$\begin{aligned}\mathbb{E}\big\{U_n\big\} &= 2^{p(n)+1} \cdot \mathbb{P}\bigg\{\omega \in \Big[\frac{n - 2^{p(n)+1} + 1}{2^{p(n)+1}}, \frac{n - 2^{p(n)+1} + 2}{2^{p(n)+1}}\Big]\bigg\} \\ &= 2^{p(n)+1} \cdot \frac{1}{2^{p(n)+1}} \\ &= 1,\end{aligned}$$

implying that $\big\{U_n \bigm| n = 1, 2, \cdots\big\}$ does not converge in mean to zero as n approaches to infinity.

Remark 2. Back to our setting in Section 2.2 in the main text, applying (S.3) to each $i = 1, \cdots, n$, we define empirical kernel mean embedding

$$\mu_k(P_{\mathcal{S}_i}) = \frac{1}{m_i} \sum_{j=1}^{m_i} k(Z_{ij}, \cdot), \tag{S.7}$$

and then application of Theorem S1 gives that for each $i = 1, \cdots, n$,

$$\lim_{m_i \to \infty} \mathbb{E}||\mu_k(P_{\mathcal{S}_i}) - \mu_k(P_i)||_{\mathcal{H}_k} = 0,$$

where the expectation is evaluated with respect to the joint distribution of $\mathcal{S}_i$.

S1.3 Proof of Theorem 1 in the Main Text

Proof of Theorem 1 (a):

Applying Theorem 3 of Lopez-Paz et al. (2015) with δ set to $\frac{1}{n}$ and using the notation $C(n, m, L_\varphi, L_{\mathcal{F}}, B)$ in (3.6) in the main text, we have that

$$\mathbb{P}\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0) \le C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big)\Big\} \ge 1 - \frac{1}{n}. \tag{S.8}$$

Since $0 \le \varphi(-f(h)l) \le B$ for every $f \in \mathcal{F}$, $h \in \mathcal{H}_k$ and $l \in \mathcal{L}$,

$$\begin{aligned} R_\varphi(\hat{f}) - R_\varphi(f_0) \le |R_\varphi(\hat{f})| + |R_\varphi(f_0)| \\ &= \mathbb{E}\{\varphi(-\hat{f}(\mu_k(P)(\cdot))l)\} + \mathbb{E}\{\varphi(-f_0(\mu_k(P)(\cdot))l)\} \\ &\le 2B. \end{aligned} \tag{S.9}$$

Using (S.8) with (S.9), we now examine that for any n and m_i,

$$\begin{aligned} &\mathbb{E}\{R_\varphi(\hat{f}) - R_\varphi(f_0)\} \\ = &\mathbb{E}\Big[\big\{R_\varphi(\hat{f}) - R_\varphi(f_0)\big\} I\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0) \le C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big)\Big\} \end{aligned}$$

$$
\begin{aligned}
&+ \{R_\varphi(\hat{f}) - R_\varphi(f_0)\} I\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0) > C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big)\Big\}\Big] \\
&\leq \mathbb{E}\Big[C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) I\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0) \leq C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big)\Big\} \\
&\quad + 2BI\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0) > C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big)\Big\}\Big] \\
&= C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) \mathbb{P}\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0) \leq C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big)\Big\} \\
&\quad + 2B\mathbb{P}\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0) > C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big)\Big\} \\
&\leq C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + 2B\mathbb{P}\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0) > C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big)\Big\} \\
&\leq C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{2B}{n}, \qquad \text{(S.10)}
\end{aligned}
$$

where the first inequality comes from (S.9) as well as the definition of the indicator function, and the last inequality is due to (S.8). This completes the proof of Theorem 1 (a).

Proof of Theorem 1 (b):

By the remark below Theorem 3 of Lopez-Paz et al. (2015), the order of $R(\mathcal{F})$ is $O(n^{-\frac{1}{2}})$, showing that $\lim_{n\to\infty} R(\mathcal{F}) = 0$. Since $k(z,z)$ is bounded on $\mathcal{Z}$, $\mathbb{E}\{k(Z_i, Z_i)\} < \infty$, we obtain that

$$
\lim_{n\to\infty} \lim_{\min\limits_{1\leq i\leq n} m_i \to\infty} \frac{4L_\varphi L_{\mathcal{F}}}{n} \sum_{i=1}^{n} \left[\sqrt{\frac{\mathbb{E}\{k(Z_i, Z_i)\}}{m_i}} + \sqrt{\frac{\log 2n}{m_i}}\right] = 0,
$$

and thus, by the assumption $R(\mathcal{F}) = \mathcal{O}(n^{-\frac{1}{2}})$,

$$
\lim_{n\to\infty} \left\{C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{2B}{n}\right\} = 0. \qquad \text{(S.11)}
$$

Since f_0 is the minima of $R_\varphi(\cdot)$, we have that

$$R_\varphi(\hat{f}) - R_\varphi(f_0) \geq 0.$$

Therefore, by Theorem 1 (a) in the main text, we have that

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

Proof of Theorem 1 (c):

Part (ii) of Theorem 1 (c) is immediate by taking the limit on both sides of the inequality in part (i) of Theorem 1 (c), where we utilize the proof of part (b) and the fact that $\zeta_\varphi(\cdot)$ is a continuous function with $\zeta_\varphi(0) = 0$.

Now it remains to show part (i) of Theorem 1 (c). By Theorem 1 and Lemma 2 of Bartlett et al. (2006), for the convex surrogate $\varphi(\cdot)$, there exists a nonnegative continuous convex function $\tilde{\psi}_\varphi(\cdot)$ from $[-1, 1]$ to $\mathbb{R}$ with $\tilde{\psi}_\varphi(0) = 0$ such that

$$\begin{aligned}\tilde{\psi}_\varphi(R(\hat{f}) - R_0) &\leq R_\varphi(\hat{f}) - \inf_{h\in\mathcal{G}} R_\varphi(h) \\ &= R_\varphi(\hat{f}) - R_\varphi(f_0), \qquad \text{(S.12)}\end{aligned}$$

where the equality is due to $\inf_{h\in\mathcal{G}} R_\varphi(h) = \min_{f\in\mathcal{F}} R_\varphi(f) = R(f_0)$.

If the convex surrogate $\varphi(\cdot)$ is classification-calibration, then by the comment after Theorem 1 of Bartlett et al. (2006), $\tilde{\psi}_\varphi(\cdot)$ is invertible on $[0, 1]$. Thus, we consider the restricted version of $\tilde{\psi}_\varphi(\cdot)$ on $[0, 1]$, and let $\psi_\varphi(\cdot)$

denote it. That is, $\psi_\varphi(\cdot)$ maps $[0,1]$ to $\mathbb{R}$, satisfying $\psi_\varphi(x) = \tilde{\psi}_\varphi(x)$ for all $x \in [0,1]$. Then $\psi_\varphi(\cdot)$ is nonnegative, convex, invertible, and continuous over $[0,1]$, where continuity at the end points 0 and 1 refers to the right-continuous at 0 and left-continuous at 1, respectively. Further, ψ_φ is strictly increasing over $[0,1]$. Indeed, by part 9 of Lemma 2 of Bartlett et al. (2006), for all $x \in (0,1]$, we have that $\psi_\varphi(x) > 0$, i.e., $\psi_\varphi(x) > \psi_\varphi(0)$ because $\psi_\varphi(0) = \tilde{\psi}_\varphi(0) = 0$; by part 2 of Lemma 1 of Bartlett et al. (2006), we have that for all $0 < y < x \le 1$, $\psi_\varphi(y) \le \frac{y}{x}\psi_\varphi(x) < \psi_\varphi(x)$. Therefore, $\psi_\varphi(\cdot)$ is nonnegative, convex, continuous, strictly increasing, and invertible with $\psi_\varphi(0) = 0$.

As the domain $[0,1]$ of $\psi_\varphi(\cdot)$ is compact and $\mathbb{R}$ is a Hausdorff space (Kelly 2017), by the result that the inverse of a continuous bijection from a compact space onto a Hausdorff space is also continuous (Hoffmann 2015), the inverse of $\psi_\varphi(\cdot)$, denoted $\zeta_\varphi(\cdot)$, is continuous. In addition, because $\psi_\varphi(\cdot)$ is strictly increasing with $\psi_\varphi(0) = 0$, its inverse $\zeta_\varphi(\cdot)$ is also strictly increasing with $\psi_\varphi(0) = 0$.

Because R_0 is the minimum value of $R(h)$ over $\mathcal{G}$, $\mathcal{F}$ is a subset of $\mathcal{G}$, and $R(\cdot)$ is always between 0 and 1 by (3.3) with $\ell(\cdot)$ being between 0 and 1, we obtain that

$$0 \le R(\hat{f}) - R_0 \le R(\hat{f}) \le 1.$$

Then by (S.12) and the definition of ψ_φ, we have that

$$\psi_\varphi(R(\hat{f}) - R_0) \le R_\varphi(\hat{f}) - R_\varphi(f_0). \tag{S.13}$$

Then by the property of ψ_φ and Jensen's inequality together with (S.13), we have that

$$\begin{aligned}\psi_\varphi\big(\mathbb{E}\{R(\hat{f}) - R_0\}\big) &\le \mathbb{E}\Big\{\psi_\varphi\big(R(\hat{f}) - R_0\big)\Big\}\\ &\le \mathbb{E}\big\{R_\varphi(\hat{f}) - R_\varphi(f_0)\big\},\end{aligned}$$

yielding that by the monotonicity of the inverse $\zeta_\varphi(\cdot)$ of $\psi_\varphi(\cdot)$,

$$\mathbb{E}\{R(\hat{f}) - R_0\} \le \zeta_\varphi\Big(\mathbb{E}\big\{R_\varphi(\hat{f}) - R_\varphi(f_0)\big\}\Big). \tag{S.14}$$

Then combining (S.14) with the result in part (a), we prove part (i) of Theorem 1 (c).

Proof of Theorem 1 (d): By (S.13), there exists a nonnegative, convex, continuous, and strictly increasing function $\psi_\varphi : [0,1] \to \mathbb{R}$ with $\psi_\varphi(0) = 0$ such that

$$\psi_\varphi\big(R(\hat{f}) - R_0\big) \le R_\varphi(\hat{f}) - R_\varphi(f_0).$$

Then taking the expectation on both sides of this inequality and utilizing Jensen's inequality yield

$$\begin{aligned}\psi_\varphi\Big(\mathbb{E}\{R(\hat{f}) - R_0\}\Big) &\le \mathbb{E}\Big\{\psi_\varphi\big(R(\hat{f}) - R_0\big)\Big\}\\ &\le \mathbb{E}\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0)\Big\},\end{aligned}$$

which shows (3.7).

Furthermore, applying Theorem 1.3 of Bartlett et al. (2006), we prove the equivalence among parts (i), (ii), and (iii).

S1.4 Proof of Theorem 2 in the Main Text

Proof of Part (a):

We first examine $R_\varphi(g) - R_\varphi(f_0)$ by connecting it with $\hat{R}_\varphi(\cdot)$ and $\mathcal{F}$:

$$\begin{aligned}
&R_\varphi(g) - R_\varphi(f_0)\\
&= \{R_\varphi(g) - \hat{R}_\varphi(g)\} + \{\hat{R}_\varphi(g) - \hat{R}_\varphi(\hat{f})\} + \{\hat{R}_\varphi(\hat{f}) - \hat{R}_\varphi(f_0)\} + \{\hat{R}_\varphi(f_0) - R_\varphi(f_0)\}\\
&\leq \{R_\varphi(g) - \hat{R}_\varphi(g)\} + \{\hat{R}_\varphi(g) - \hat{R}_\varphi(\hat{f})\} + \{\hat{R}_\varphi(f_0) - R_\varphi(f_0)\}\\
&\leq 2\sup_{f\in\mathcal{F}} |R_\varphi(f) - \hat{R}_\varphi(f)| + \{\hat{R}_\varphi(g) - \hat{R}_\varphi(\hat{f})\}\\
&\leq 2\sup_{f\in\mathcal{F}} |R_\varphi(f) - \hat{R}_\varphi(f)| + \left|\hat{R}_\varphi(g) - \hat{R}_\varphi(\hat{f})\right|, \qquad \text{(S.15)}
\end{aligned}$$

where the first inequality holds since $\hat{f}$ is the minimum point of the functional $\hat{R}_\varphi(f)$, and the second inequality comes from the definition of supremum.

By the proof of Theorem 3 of Lopez-Paz et al. (2015) and setting their δ to $\frac{1}{n}$, we have that

$$\mathbb{P}\Big\{2\sup_{f\in\mathcal{F}} |R_\varphi(f) - \hat{R}_\varphi(f)| \leq C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\} \geq 1 - \frac{1}{n}. \qquad \text{(S.16)}$$

Since $0 \leq \varphi(-f(h)l) \leq B$ for every $f \in \mathcal{F}$, $h \in \mathcal{H}_k$ and $l \in \mathcal{L}$,

$$\begin{aligned}|R_\varphi(f) - \hat{R}_\varphi(f)| &\leq |R_\varphi(f)| + |\hat{R}_\varphi(f)| \\ &= \Big|\mathbb{E}\{\varphi(-\hat{f}(\mu_k(P)(\cdot))l)\}\Big| + \Big|\frac{1}{n}\sum_{i=1}^{n}\varphi(-l_i f(\mu_k(P_{\mathcal{S}_i})))\Big| \\ &\leq 2B,\end{aligned}$$

where the second step is due to the definition of $R_\varphi(\cdot)$ at the end of the paragraph before (3.4) in the main text and the definition of $\hat{R}_\varphi(\cdot)$ before (3.5) in the main text. Consequently,

$$2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| \leq 4B \tag{S.17}$$

Then, for any n and m_i,

$$\begin{aligned}&\mathbb{E}\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)|\} \\ &= \mathbb{E}\Big[\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)|\}I\Big\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| \leq C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\} \\ &\quad + \{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)|\}I\Big\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| > C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\}\Big] \\ &\leq \mathbb{E}\Big[C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)I\Big\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| \leq C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\} \\ &\quad + 4BI\Big\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| > C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\}\Big] \\ &= C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\mathbb{P}\Big\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| \leq C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\} \\ &\quad + 4B\mathbb{P}\Big\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| > C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\}\end{aligned}$$

$$
\begin{aligned}
&\leq C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big) + 4B\mathbb{P}\Big\{2\sup_{f\in\mathcal{F}} |R_\varphi(f) - \hat{R}_\varphi(f)| > C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\} \\
&\leq C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big) + \frac{4B}{n}, \qquad \text{(S.18)}
\end{aligned}
$$

where the first inequality comes from (S.17), the third step is due to the definition of the indicator function, and the last inequality is due to (S.16).

Now we examine the second term of (S.15). By the definition of $\hat{R}_\varphi(\cdot)$ before (3.5) in the main text, we have that

$$
\begin{aligned}
&\big|\hat{R}_\varphi(g) - \hat{R}_\varphi(\hat{f})\big| \\
&= \Big|\frac{1}{n}\sum_{i=1}^{n} \varphi(-l_i g(\mu_k(P_{\mathcal{S}_i}))) - \frac{1}{n}\sum_{i=1}^{n} \varphi(-l_i \hat{f}(\mu_k(P_{\mathcal{S}_i})))\Big| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \big|\varphi(-l_i g(\mu_k(P_{\mathcal{S}_i}))) - \varphi(-l_i \hat{f}(\mu_k(P_{\mathcal{S}_i})))\big| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} L_\varphi |g(\mu_k(P_{\mathcal{S}_i})) - \hat{f}(\mu_k(P_{\mathcal{S}_i}))| \cdot |l_i| \\
&= \frac{1}{n}\sum_{i=1}^{n} L_\varphi |g(\mu_k(P_{\mathcal{S}_i})) - \hat{f}(\mu_k(P_{\mathcal{S}_i}))| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} L_\varphi \sup_{x\in\mathcal{H}_k} |g(x) - \hat{f}(x)| \\
&= L_\varphi \sup_{x\in\mathcal{H}_k} |g(x) - \hat{f}(x)|, \qquad \text{(S.19)}
\end{aligned}
$$

where the second step is due to the triangle inequality of absolute value, the third step comes from condition (R3) of Theorem 1, the fourth step is due to the fact that $|l_i| = 1$ for any i, and the fifth step is due to the definition of supremum.

Taking expectation on both sides of (S.15) and combining (S.18) and (S.19) yield part (a), where the definition of $F(\hat{f}, g, L_\varphi)$ is used.

Proof of Part (b):

Taking limsup on the both sides of the inequality in part (a) in the main text and utilizing that

$$\begin{aligned}
&\limsup_{n\to\infty}\limsup_{m\to\infty}\left\{C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{4B}{n}\right\} \\
= &\lim_{n\to\infty}\lim_{m\to\infty}\left\{C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{4B}{n}\right. \\
= &\ 0,
\end{aligned}$$

we prove part (b) in the main text.

Proof of Part (c):

Replacing $\hat{f}$ with g and replacing the upper bound in Theorem 1 (a) with the upper bound in part (a), we repeat the proof of Part (i) in Theorem 1 (c) and then we can prove the first inequality in part (c).

Taking limsup with respect to m and n on both sides of the first inequality in Part (c) and utilizing the fact that $\zeta_\varphi(\cdot)$ is continuous, we can show the second inequality in part (c).

Proof of Part (d): Repeating the proof of Theorem 1 (d) by replacing $\hat{f}$ with g yields part (d).

S1.5 Proof of Theorem 3 in the Main Text

Lemma 1: *For any l_i and l_i^*, we have that* $\mathbb{E}|l_i - l_i^*| \leq 2D$.

Proof. First, it can be easily shown that for any random variable U and a binary variable V taking values v_1 and v_2,

$$\mathbb{E}(U) = \mathbb{E}(U|V = v_1)\mathbb{P}(V = v_1) + \mathbb{E}(U|V = v_2)\mathbb{P}(V = v_2) \tag{S.20}$$

Next, we examine $\mathbb{E}|l_i - l_i^*|$ using (S.20) with U set as $|l_i - l_i^*|$ and V taken different forms for (4.9) and (4.10) in the main text. If (4.9) in the main text is used to describe the misclassification in labels, then we set $V = l_i$ and applying (S.20) gives that

$$\begin{aligned}
\mathbb{E}|l_i - l_i^*| &= \mathbb{E}\{|l_i - l_i^*|\big|l_i = 1\} \times \mathbb{P}(l_i = 1) + \mathbb{E}\{|l_i - l_i^*|\big|l_i = -1\} \times \mathbb{P}(l_i = -1) \\
&\leq \mathbb{E}\{|l_i - l_i^*|\big|l_i = 1\} + \mathbb{E}\{|l_i - l_i^*|\big|l_i = -1\} \\
&= 2(1 - p_1^*) + 2(1 - p_{-1}^*) \\
&= 2D.
\end{aligned}$$

When (4.10) in the main text is used to describe the misclassification in labels, we take $V_i = l_i^*$ and applying (S.20) gives that

$$\begin{aligned}
\mathbb{E}|l_i - l_i^*| &= \mathbb{E}\{|l_i - l_i^*|\big|l_i^* = 1\} \times \mathbb{P}(l_i^* = 1) + \mathbb{E}\{|l_i - l_i^*|\big|l_i^* = -1\} \times \mathbb{P}(l_i^* = -1) \\
&\leq \mathbb{E}\{|l_i - l_i^*|\big|l_i^* = 1\} + \mathbb{E}\{|l_i - l_i^*|\big|l_i^* = -1\} \\
&= 2(1 - p_1) + 2(1 - p_{-1})
\end{aligned}$$

$$= 2D.$$

Therefore, the conclusion in the lemma follows. □

Proof of Part (a):

The proof of the first inequality consists of the following three steps.

Step 1: First, we examine the absolute difference $|R_\varphi(\hat{f}^*) - R_\varphi(\hat{f})|$ by connecting it with $\hat{R}(\cdot)$ and $\mathcal{F}$, where we examine two cases by the sign of $R_\varphi(\hat{f}^*) - R_\varphi(\hat{f})$.

Case 1: when $R_\varphi(\hat{f}^*) - R_\varphi(\hat{f}) \geq 0$:

$$\begin{aligned}
|R_\varphi(\hat{f}^*) - R_\varphi(\hat{f})| &= R_\varphi(\hat{f}^*) - R_\varphi(\hat{f}) \\
&= \big\{R_\varphi(\hat{f}^*) - \hat{R}^*_\varphi(\hat{f}^*)\big\} + \big\{\hat{R}^*_\varphi(\hat{f}^*) - \hat{R}^*_\varphi(\hat{f})\big\} + \big\{\hat{R}^*_\varphi(\hat{f}) - R_\varphi(\hat{f})\big\} \\
&\leq \big\{R_\varphi(\hat{f}^*) - \hat{R}^*_\varphi(\hat{f}^*)\big\} + \big\{\hat{R}^*_\varphi(\hat{f}) - R_\varphi(\hat{f})\big\} \\
&\leq 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}^*_\varphi(f)| \qquad \text{(S.21)} \\
&= 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f) + \hat{R}_\varphi(f) - \hat{R}^*_\varphi(f)| \\
&\leq 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}}|\hat{R}_\varphi(f) - \hat{R}^*_\varphi(f)|, \qquad \text{(S.22)}
\end{aligned}$$

where the first inequality holds since $\hat{R}^*_\varphi(\hat{f}^*) - \hat{R}^*_\varphi(\hat{f}) \leq 0$ by that $\hat{f}^*$ is the minimum point of the functional $\hat{R}^*_\varphi(f)$, and the last inequality comes from the triangle inequality and the definition of supremum.

Case 2: when $R_\varphi(\hat{f}^*) - R_\varphi(\hat{f}) < 0$:

$$\begin{aligned}
|R_\varphi(\hat{f}^*) - R_\varphi(\hat{f})| &= R_\varphi(\hat{f}) - R_\varphi(\hat{f}^*) \\
&= \big\{R_\varphi(\hat{f}) - \hat{R}_\varphi(\hat{f})\big\} + \big\{\hat{R}_\varphi(\hat{f}) - \hat{R}_\varphi(\hat{f}^*)\big\} + \big\{\hat{R}_\varphi(\hat{f}^*) - R_\varphi(\hat{f}^*)\big\} \\
&\le \big\{R_\varphi(\hat{f}) - \hat{R}_\varphi(\hat{f})\big\} + \big\{\hat{R}_\varphi(\hat{f}^*) - R_\varphi(\hat{f}^*)\big\} \\
&\le 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| \\
&\le 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}}|\hat{R}_\varphi(f) - \hat{R}^*_\varphi(f)|, \quad \text{(S.23)}
\end{aligned}$$

where the first inequality holds since $\hat{R}_\varphi(\hat{f}) - \hat{R}_\varphi(\hat{f}^*) \le 0$ by that $\hat{f}$ is the minimum point of the functional $\hat{R}_\varphi(f)$.

By combining (S.22) and (S.23), we have that

$$|R_\varphi(\hat{f}^*) - R_\varphi(\hat{f})| \le 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}}|\hat{R}_\varphi(f) - \hat{R}^*_\varphi(f)|. \quad \text{(S.24)}$$

Step 2: Next, we examine the first term of (S.24). By the proof of Theorem 3 in Lopez-Paz et al. (2015) and setting their δ to $\frac{1}{n}$, we have that

$$\mathbb{P}\Big\{2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| \le C\Big(n, m, L_\varphi, L_\mathcal{F}, B\Big)\Big\} \ge 1 - \frac{1}{n}. \quad \text{(S.25)}$$

In addition,

$$2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)|$$

$$
\begin{aligned}
&\leq 2\sup_{f\in\mathcal{F}}|R_{\varphi}(f)| + 2\sup_{f\in\mathcal{F}}|\hat{R}_{\varphi}(f)| \\
&= 2\sup_{f\in\mathcal{F}}\Big|\mathbb{E}\{\varphi(-f(\mu_k(P)(\cdot))l)\}\Big| + 2\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varphi(-l_i f(\mu_k(P_{\mathcal{S}_i})))\Big| \\
&\leq 4B \qquad \text{(S.26)}
\end{aligned}
$$

where the first step comes from the triangle inequality and the definition of supremum, the second step is due to the definition of $R_{\varphi}(\cdot)$ at the end of the paragraph before (3.4) in the main text and the definition of $\hat{R}_{\varphi}(\cdot)$ before (3.5) in the main text, and the third step comes from the condition that $0 \leq \varphi(-f(h)l) \leq B$ for every $f \in \mathcal{F}$, $h \in \mathcal{H}_k$ and $l \in \mathcal{L}$.

Next, we evaluate $\mathbb{E}\{2\sup_{f\in\mathcal{F}}|R_{\varphi}(f)-\hat{R}_{\varphi}(f)|\}$ by following the same derivations as in (S.10), except replacing $R_{\varphi}(\hat{f})-R_{\varphi}(f_0)$ in (S.10) with $2\sup_{f\in\mathcal{F}}|R_{\varphi}(f)-\hat{R}_{\varphi}(f)|$, (S.8) with (S.25), and (S.9) with (S.26). Then we obtain

$$
\mathbb{E}\{2\sup_{f\in\mathcal{F}}|R_{\varphi}(f) - \hat{R}_{\varphi}(f)|\} \leq C\Big(n, m, L_{\varphi}, L_{\mathcal{F}}, B\Big) + \frac{4B}{n}. \qquad \text{(S.27)}
$$

Step 3: Finally, we examine the expectation of the second term in (S.24).

$$
\begin{aligned}
&\mathbb{E}\Big\{2\sup_{f\in\mathcal{F}}|\hat{R}_{\varphi}(f) - \hat{R}^{*}_{\varphi}(f)|\Big\} \\
&= \mathbb{E}\Big[2\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\{\varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i) - \varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i^{*})\}\Big|\Big] \\
&\leq \mathbb{E}\Big[2\sup_{f\in\mathcal{F}}\Big\{\frac{1}{n}\sum_{i=1}^{n}|\varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i) - \varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i^{*})|\Big\}\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\{\sup_{f\in\mathcal{F}}|\varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i)-\varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i^*)|\} \\
&\leq \frac{2L_\varphi}{n}\sum_{i=1}^{n}\mathbb{E}\{\sup_{f\in\mathcal{F}}|f(\mu_k(P_{\mathcal{S}_i}))l_i-f(\mu_k(P_{\mathcal{S}_i}))l_i^*|\} \\
&= \frac{2L_\varphi}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big\{\sup_{f\in\mathcal{F}}|f(\mu_k(P_{\mathcal{S}_i}))|\big\}|l_i-l_i^*|\Big], \qquad \text{(S.28)}
\end{aligned}
$$

where the first inequality is due to Jensen's inequality, the second inequality comes from that $\sup_x\{|g_1(x)|+|g_2(x)|\}\leq\sup_x|g_1(x)|+\sup_x|g_2(x)|$ for any functions g_1 and g_2, and the third inequality is due to the Lipschitzness of φ.

Since for any $f\in\mathcal{F}$ and $h\in\mathcal{H}_k$, $|f(h)|\leq M||h||_{\mathcal{H}_k}$, then

$$
|f(\mu_k(P_{\mathcal{S}_i}))|\leq M||\mu_k(P_{\mathcal{S}_i})||_{\mathcal{H}_k}. \qquad \text{(S.29)}
$$

Furthermore,

$$
\begin{aligned}
||\mu_k(P_{\mathcal{S}_i})||_{\mathcal{H}_k} &= \left\langle \frac{1}{m_i}\sum_{j=1}^{m_i}k(Z_{ij},\cdot),\frac{1}{m_i}\sum_{j=1}^{m_i}k(Z_{ij},\cdot)\right\rangle_{\mathcal{H}_k} \\
&= \frac{1}{m_i^2}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}k(Z_{ij},Z_{ik}) \\
&\leq A, \qquad \text{(S.30)}
\end{aligned}
$$

where the last step comes from that $k(z_1,z_2)\leq A$ for any $z_1,z_2\in\mathcal{Z}$. Combining (S.29) and (S.30) gives that

$$
\sup_{f\in\mathcal{F}}|f(\mu_k(P_{\mathcal{S}_i}))|\leq MA. \qquad \text{(S.31)}
$$

Combining (S.28), (S.31) and Lemma 1 gives that

$$\mathbb{E}\{2\sup_{f\in\mathcal{F}}|\hat{R}_\varphi(f) - \hat{R}^*_\varphi(f)|\} \leq 4ML_\varphi AD. \tag{S.32}$$

Then applying (S.27) and (S.32) to (S.24) proves the first inequality in part (a).

Now we prove the second inequality in part (a). By the triangle inequality, we have that

$$\begin{aligned}\frac{\mathbb{E}\{|R(\hat{f}^*) - R(\hat{f})|\}}{2} &= \frac{\mathbb{E}\{|R(\hat{f}^*) - R_0 + R_0 - R(\hat{f})|\}}{2} \\ &\leq \frac{\mathbb{E}\{|R(\hat{f}^*) - R_0|\}}{2} + \frac{\mathbb{E}\{|R_0 - R(\hat{f})|\}}{2} \\ &= \frac{\mathbb{E}\{R(\hat{f}^*) - R_0\}}{2} + \frac{\mathbb{E}\{R(\hat{f}) - R_0\}}{2} \end{aligned} \tag{S.33}$$

where the last equality is because both $R(\hat{f}^*)$ and $R(\hat{f})$ are always greater than or equal to R_0.

By repeating the derivation of (S.13) with replacing $\hat{f}$ by $\hat{f}^*$, we have that

$$\psi_\varphi(R(\hat{f}^*) - R_0) \leq R_\varphi(\hat{f}^*) - R_\varphi(f_0). \tag{S.34}$$

where $\psi_\varphi(\cdot)$ is defined in the proof of Theorem 1 and is a nonnegative, convex, continuous, and strictly increasing function.

Then applying $\psi_\varphi(\cdot)$ to both sides of (S.33) yields that

$$\psi_\varphi\Big(\frac{1}{2}\mathbb{E}\{|R(\hat{f}^*) - R(\hat{f})|\}\Big) \leq \psi_\varphi\Big(\frac{\mathbb{E}\{R(\hat{f}^*) - R_0\}}{2} + \frac{\mathbb{E}\{R(\hat{f}) - R_0\}}{2}\Big)$$

$$
\begin{aligned}
&\leq \frac{\psi_\varphi\Big(\mathbb{E}\{R(\hat{f}^*) - R_0\}\Big) + \psi_\varphi\Big(\mathbb{E}\{R(\hat{f}) - R_0\}\Big)}{2} \\
&\leq \frac{\mathbb{E}\Big\{\psi_\varphi\big(R(\hat{f}^*) - R_0\big)\Big\} + \mathbb{E}\Big\{\psi_\varphi\big(R(\hat{f}) - R_0\big)\Big\}}{2} \\
&\leq \frac{\mathbb{E}\Big\{R_\varphi(\hat{f}^*) - R_\varphi(f_0)\Big\} + \mathbb{E}\Big\{R_\varphi(\hat{f}) - R_\varphi(f_0)\Big\}}{2} \\
&= \frac{\mathbb{E}\Big\{R_\varphi(\hat{f}^*) - R_\varphi(f_0) + R_\varphi(\hat{f}) - R_\varphi(f_0)\Big\}}{2},
\end{aligned}
\tag{S.35}
$$

where the second inequality is due to the convexity of ψ_φ, the third inequality is due to Jensen's inequality, and the fourth inequality is due to (S.13) and (S.34).

Now we examine $R_\varphi(\hat{f}^*) - R_\varphi(f_0) + R_\varphi(\hat{f}) - R_\varphi(f_0)$ in (S.35) by introducing $\hat{R}^*_\varphi(\hat{f}^*)$, $\hat{R}^*_\varphi(f_0)$, $\hat{R}_\varphi(\hat{f})$, and $\hat{R}_\varphi(f_0)$ as bridging components:

$$
\begin{aligned}
&R_\varphi(\hat{f}^*) - R_\varphi(f_0) + R_\varphi(\hat{f}) - R_\varphi(f_0) \\
&= R_\varphi(\hat{f}^*) - \hat{R}^*_\varphi(\hat{f}^*) + \hat{R}^*_\varphi(\hat{f}^*) - \hat{R}^*_\varphi(f_0) + \hat{R}^*_\varphi(f_0) - R_\varphi(f_0) + R_\varphi(\hat{f}) - \hat{R}_\varphi(\hat{f}) \\
&\quad + \hat{R}_\varphi(\hat{f}) - \hat{R}_\varphi(f_0) + \hat{R}_\varphi(f_0) - R_\varphi(f_0) \\
&\leq R_\varphi(\hat{f}^*) - \hat{R}^*_\varphi(\hat{f}^*) + \hat{R}^*_\varphi(f_0) - R_\varphi(f_0) + R_\varphi(\hat{f}) - \hat{R}_\varphi(\hat{f}) + \hat{R}_\varphi(f_0) - R_\varphi(f_0) \\
&\leq 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}}|\hat{R}^*_\varphi(f) - R_\varphi(f)| \\
&= 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}}\big|\hat{R}^*_\varphi(f) - \hat{R}_\varphi(f) + \hat{R}_\varphi(f) - R_\varphi(f)\big| \\
&\leq 2\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}}|\hat{R}^*_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}}|\hat{R}_\varphi(f) - R_\varphi(f)|
\end{aligned}
$$

$$= 4\sup_{f\in\mathcal{F}}|R_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}}|\hat{R}_\varphi(f) - \hat{R}^*_\varphi(f)|, \tag{S.36}$$

where the first inequality holds since $\hat{R}^*_\varphi(\hat{f}^*) - \hat{R}^*_\varphi(f_0) \leq 0$ and $\hat{R}_\varphi(\hat{f}) - \hat{R}_\varphi(f_0) \leq 0$ by that $\hat{f}^*$ and $\hat{f}$ are the minimum points of the functional $\hat{R}^*_\varphi(f)$ and $\hat{R}_\varphi(f)$, respectively; the second inequality comes from the definition of supremum; and the last inequality is due to the triangle inequality and the property of the supremum.

Taking the expectation on both sides of (S.36) and utilizing (S.27), (S.32), and (S.35), we obtain that

$$\psi_\varphi\Big(\frac{1}{2}\mathbb{E}\{|R(\hat{f}^*) - R(\hat{f})|\}\Big) \leq C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{4B}{n} + 2ML_\varphi AD. \tag{S.37}$$

By the proof of Theorem 1, the inverse $\zeta_\varphi(\cdot)$ of $\psi_\varphi(\cdot)$ is a nondecreasing continuous function with $\zeta_\varphi(0) = 0$. Therefore, applying $\zeta_\varphi(\cdot)$ to the both sides of (S.37), we obtain that

$$\mathbb{E}\{|R(\hat{f}^*) - R(\hat{f})|\} \leq 2\zeta_\varphi\Bigg(C\Big(n, m, L_\varphi, L_{\mathcal{F}}, B\Big) + \frac{4B}{n} + 2ML_\varphi AD\Bigg). \tag{S.38}$$

Then we prove the second inequality in part (a).

Proof of Part (b):

By taking limsup as both m and n approach infinity on both sides of the first inequality in part (a), as well as utilizing (S.11), we prove the first inequality in part (b).

Similarly, taking limsup as both m and n approach infinity on both sides of (S.38) and utilizing (S.11), we prove the second inequality in part (b).

S1.6 Proof of Theorem 4 in the Main Text

First, we show the following lemma.

Lemma 2: *For any $t \in \mathbb{R}$, if the misclassification probability (4.9) in the main text is taken, we have that*

$$\mathbb{E}\{\varphi^*(t, l^*)|l\} = \varphi(-tl); \tag{S.39}$$

if the reclassification probability (4.10) in the main text is taken, we have that

$$\mathbb{E}\{\varphi(-tl)|l^*\} = \varphi^*(t, l^*). \tag{S.40}$$

Proof. We first prove (S.39). Under the misclassification probability (4.9) in the main text, by definition (5.15) in the main text, we have that

$$\begin{aligned}
&\mathbb{E}\{\varphi^*(t, l^*)|l = 1\} \\
&= \mathbb{E}\Big\{\frac{p^*_{-l^*}\varphi(-tl^*) - (1 - p^*_{l^*})\varphi(tl^*)}{p^*_1 + p^*_{-1} - 1}\Big| l = 1\Big\} \\
&= \frac{p^*_{-1}\varphi(-t) - (1 - p^*_1)\varphi(t)}{p^*_1 + p^*_{-1} - 1} \times \mathbb{P}(l^* = 1|l = 1) \\
&\quad + \frac{p^*_1\varphi(t) - (1 - p^*_{-1})\varphi(-t)}{p^*_1 + p^*_{-1} - 1} \times \mathbb{P}(l^* = -1|l = 1) \\
&= \frac{p^*_{-1}\varphi(-t) - (1 - p^*_1)\varphi(t)}{p^*_1 + p^*_{-1} - 1} \times p^*_1 + \frac{p^*_1\varphi(t) - (1 - p^*_{-1})\varphi(-t)}{p^*_1 + p^*_{-1} - 1} \times (1 - p^*_1)
\end{aligned}$$

$$= \varphi(-t)$$

and

$$\begin{aligned}
&\mathbb{E}\{\varphi^*(t,l^*)|l=-1\}\\
&=\mathbb{E}\Big\{\frac{p^*_{-l^*}\varphi(-tl^*)-(1-p^*_{l^*})\varphi(tl^*)}{p^*_1+p^*_{-1}-1}\Big|l=-1\Big\}\\
&=\frac{p^*_{-1}\varphi(-t)-(1-p^*_1)\varphi(t)}{p^*_1+p^*_{-1}-1}\times\mathbb{P}(l^*=1|l=-1)\\
&\quad+\frac{p^*_1\varphi(t)-(1-p^*_{-1})\varphi(-t)}{p^*_1+p^*_{-1}-1}\times\mathbb{P}(l^*=-1|l=-1)\\
&=\frac{p^*_{-1}\varphi(-t)-(1-p^*_1)\varphi(t)}{p^*_1+p^*_{-1}-1}\times(1-p^*_{-1})+\frac{p^*_1\varphi(t)-(1-p^*_{-1})\varphi(-t)}{p^*_1+p^*_{-1}-1}\times p^*_{-1}\\
&=\varphi(t).
\end{aligned}$$

Thus, (S.39) follows.

Next, we prove (S.40). Under the reclassification probability (4.10) in the main text, we have that

$$\begin{aligned}
\mathbb{E}\{\varphi(-tl)|l^*=1\}&=\varphi(-t)\times\mathbb{P}(l=1|l^*=1)+\varphi(t)\times\mathbb{P}(l=-1|l^*=1)\\
&=\varphi(-t)\times p_1+\varphi(t)\times(1-p_1);
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\{\varphi(-tl)|l^*=-1\}&=\varphi(-t)\times\mathbb{P}(l=1|l^*=-1)+\varphi(t)\times\mathbb{P}(l=-1|l^*=-1)\\
&=\varphi(-t)\times(1-p_{-1})+\varphi(t)\times p_{-1}.
\end{aligned}$$

Then by definition (5.15) in the main text, we conclude (S.40). □

Now we prove Theorem 4 by taking a different start point, either from $R_{\varphi^*}(f)$ or $R_\varphi(f)$, depending on whether (4.9) or (4.10) in the main text is under consideration. The derivations for (4.9) and (4.10) in the main text are of opposite order, with intermediate steps differing in whether l^* or l is taken as the conditioning variable. Specifically, under the misclassification probability (4.9) in the main text, we have that

$$\begin{aligned} R_{\varphi^*}(f) &= \mathbb{E}\{\varphi^*(f(\mu_k(P)(\cdot)), l^*)\} \\ &= \mathbb{E}\big[\mathbb{E}\{\varphi^*(f(\mu_k(P)(\cdot)), l^*)|l\}\big] \\ &= \mathbb{E}\{\varphi(-lf(\mu_k(P)(\cdot)))\} \\ &= R_\varphi(f) \end{aligned} \tag{S.41}$$

where the first equality is by definition of $R_{\varphi^*}(f)$ in Section 5 of the main text, the second equality is due to the property of conditional expectation, the third equality is due to (S.39) of Lemma 2 in this appendix, and the last equality is due to the definition of $R_\varphi(\cdot)$ at the end of the paragraph before (3.4) in the main text.

In contrast, under the misclassification probability (4.10) in the main text, we have that

$$\begin{aligned} R_\varphi(f) &= \mathbb{E}\{\varphi(-f(\mu_k(P)(\cdot))l)\} \\ &= \mathbb{E}\big[\mathbb{E}\{\varphi(-lf(\mu_k(P)(\cdot)))|l^*\}\big] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\{\varphi^*(f(\mu_k(P)(\cdot)), l^*)\} \\
&= R_{\varphi^*}(f) \qquad \text{(S.42)}
\end{aligned}
$$

where the first equality is due to the definition of $R_{\varphi}(\cdot)$ at the end of the paragraph before (3.4) in the main text, the second equality is due to the property of conditional expectation, the third equality is due to (S.40) of Lemma 2 in this appendix, and the last equality comes from the definition of $R_{\varphi^*}(f)$ in Section 5 of the main text.

Thus, (S.41) and (S.42) lead to Theorem 4 in the main text.

S1.7 Proof of Theorem 5 in the Main Text

To show Theorem 5, we first present two lemmas.

Lemma 3: For any $t_1, t_2 \in \mathbb{R}$ and $l^* \in \{-1, 1\}$, we have that

$$\big|\varphi^*(t_1, l^*) - \varphi^*(t_2, l^*)\big| \leq L_{\varphi}^* |t_1 - t_2|,$$

where L_{φ}^* is defined in (5.19) of the main text.

Proof. We first consider the misclassification model (4.9) in the main text:

$$
\begin{aligned}
&\big|\varphi^*(t_1, l^*) - \varphi^*(t_2, l^*)\big| \\
&= \Big| \frac{p_{-l^*}^* \varphi(-t_1 l^*) - (1 - p_{l^*}^*)\varphi(t_1 l^*)}{p_1^* + p_{-1}^* - 1} - \frac{p_{-l^*}^* \varphi(-t_2 l^*) - (1 - p_{l^*}^*)\varphi(t_2 l^*)}{p_1^* + p_{-1}^* - 1} \Big| \\
&\leq \frac{p_{-l^*}^*}{|p_1^* + p_{-1}^* - 1|} \big|\varphi(-t_1 l^*) - \varphi(-t_2 l^*)\big| + \frac{1 - p_{l^*}^*}{|p_1^* + p_{-1}^* - 1|} \big|\varphi(t_1 l^*) - \varphi(t_2 l^*)\big|
\end{aligned}
$$

$$\begin{aligned}
&\leq \frac{(1-p^*_{l^*}+p^*_{-l^*})L_\varphi}{|p^*_1+p^*_{-1}-1|}|t_1-t_2| \\
&\leq \frac{2L_\varphi}{|p^*_1+p^*_{-1}-1|}|t_1-t_2| \\
&= L^*_\varphi|t_1-t_2|, \qquad \text{(S.43)}
\end{aligned}$$

where the first equality is due to the expression of $\varphi^*(\cdot,\cdot)$ in (5.15) of the main text, the second step is due to the triangle inequality of absolute value, the third step holds because $\varphi(\cdot)$ is L_φ-Lipschitz continuous and $|l^*|=1$ for any $l^*\in\{-1,1\}$, the fourth step is due to the fact that $p^*_{-l^*}\leq 1+p^*_{l^*}$, and the last step is due to (5.19) in the main text.

Similarly, we evaluate the reclassification model (4.10) in the main text:

$$\begin{aligned}
\left|\varphi^*(t_1,l^*)-\varphi^*(t_2,l^*)\right| &= \left|p_{l^*}(\varphi(-t_1l^*)-\varphi(-t_2l^*))+(1-p_{l^*})(\varphi(t_1l^*)-\varphi(t_2l^*))\right| \\
&\leq p_{l^*}|\varphi(-t_1l^*)-\varphi(-t_2l^*)|+(1-p_{l^*})|\varphi(t_1l^*)-\varphi(t_2l^*)| \\
&\leq p_{l^*}L_\varphi|t_1-t_2|+(1-p_{l^*})L_\varphi|t_1-t_2| \\
&= L_\varphi|t_1-t_2| \\
&= L^*_\varphi|t_1-t_2|, \qquad \text{(S.44)}
\end{aligned}$$

where the first equality is due to the expression of $\varphi^*(\cdot,\cdot)$ in (5.15) of the main text, the first inequality is due to the triangle inequality of absolute value, and the second inequality is because $\varphi(\cdot)$ is L_φ-Lipschitz continuous and $|l^*|=1$ for any $l^*\in\{-1,1\}$.

By combining (S.43) and (S.44), we prove Lemma 3. □

Lemma 4: For any $f \in \mathcal{F}$, $h \in \mathcal{H}_k$, and $l \in \mathcal{L}$, we have that

$$\left|\varphi^*(f(h), l)\right| \leq B^*,$$

where B^* is defined in (5.20) in the main text.

Proof. If the reclassification model (4.10) in the main text is taken, the proof is straightforward by using condition (R2) in Theorem 1 and the definition of B^* in (5.20) in the main text.

Then we examine the misclassification model (4.9) in the main text. For any $f \in \mathcal{F}$, $h \in \mathcal{H}_k$, and $l \in \mathcal{L}$, we have that

$$\begin{aligned} \left|\varphi^*(f(h), l)\right| &= \left|\frac{p_{-l}^*\varphi(-f(h)l) - (1-p_l^*)\varphi(f(h)l)}{p_1^* + p_{-1}^* - 1}\right| \\ &\leq \frac{p_{-l}^*}{|p_1^* + p_{-1}^* - 1|}\left|\varphi(-f(h)l)\right| + \frac{1-p_l^*}{|p_1^* + p_{-1}^* - 1|}\left|\varphi(f(h)l)\right| \\ &\leq \frac{Bp_{-l}^*}{|p_1^* + p_{-1}^* - 1|} + \frac{B(1-p_l^*)}{|p_1^* + p_{-1}^* - 1|} \\ &\leq \frac{2B}{|p_1^* + p_{-1}^* - 1|} \end{aligned} \qquad \text{(S.45)}$$

where the first inequality is due to the triangle inequality of absolute value, the second inequality is due to condition (R2) presented in Theorem 1, and the last inequality holds because both p_{-l}^* and $1 - p_l^*$ are less than or equal to one for any $l \in \mathcal{L}$.

By the definition of B^* in (5.20) in the main text, (S.45) implies Lemma 4. □

Proof of Part (a):

By Theorem 4 in the main text, f_0 is also the minimum of the φ^*-risk functional $R_{\varphi^*}(\cdot)$. We further have

$$R_\varphi(\hat{f}^{correct}) - R_\varphi(f_0) = R_{\varphi^*}(\hat{f}^{correct}) - R_{\varphi^*}(f_0). \tag{S.46}$$

With $C(\cdot,\cdot,\cdot,\cdot,\cdot)$ defined in (3.6) in the main text, by applying Lemmas 3 and 4 to the proof of Theorem 3 of Lopez-Paz et al. (2015) with their δ set to $\frac{1}{n}$, we have that

$$\mathbb{P}\Big\{R_{\varphi^*}(\hat{f}^{correct}) - R_{\varphi^*}(f_0) \leq C(n, m, L^*_\varphi, L_{\mathcal{F}}, B)\Big\} \geq 1 - \frac{1}{n}.$$

Then by (S.46) and repeating the proof of Theorem 1 (a) in Section S1.3 with B replaced by B^*, we prove Theorem 5 (a) in the main text.

Proof of Part (b): Repeating the proof of (S.11) by replacing L_φ and B with L^*_φ and B^*, respectively, we have that

$$\lim_{n\to\infty}\left\{C\Big(n, m, L^*_\varphi, L_{\mathcal{F}}, B^*\Big) + \frac{2B^*}{n}\right\} = 0.$$

Then by taking limits as both m and n approach infinity in both sides of part (a), we prove part (b).

Proof of Part (c):

By repeating the proof of part (i) in Theorem 1 (c) with $\hat{f}$ and the upper bound in Theorem 1 (a) replaced by $\hat{f}^{correct}$ and the upper bound in part (a), respectively, we show the upper bound in part (c) (i). The lower bound in part (c) (i) is directly derived from the fact that R_0 is the minimum value of $R(\cdot)$ over $\mathcal{G}$ and $\hat{f}^{correct} \in \mathcal{G}$.

Part (c) (ii) is immediate by taking the limit on both sides of the inequality in part (c) (i), where we utilize the proof of part (b) and the fact that $\zeta_\varphi(\cdot)$ is a continuous function with $\zeta_\varphi(0) = 0$.

S2 Finite-dimensional Approximation Space of the Infinite-dimensional RKHS

S2.1 Approximation of Empirical Kernel Mean Embedding (S.7)

Because RKHS $\mathcal{H}_k$ is usually infinite-dimensional, it is practically difficult or even impossible to implement exact classification algorithms in $\mathcal{H}_k$ due to huge or infinite computation and memory costs. For instance, if the support vector machine (SVM) is used to classify $\mu_k(P_i)$ for $i = 1, 2, \cdots, n$, then one needs to compute the matrix

$$\begin{bmatrix} \langle \mu_k(P_1)(\cdot), \mu_k(P_1)(\cdot)\rangle_{\mathcal{H}_k} & \dots & \langle \mu_k(P_1)(\cdot), \mu_k(P_n)(\cdot)\rangle_{\mathcal{H}_k} \\ \vdots & \ddots & \vdots \\ \langle \mu_k(P_n)(\cdot), \mu_k(P_1)(\cdot)\rangle_{\mathcal{H}_k} & \dots & \langle \mu_k(P_n)(\cdot), \mu_k(P_n)(\cdot)\rangle_{\mathcal{H}_k} \end{bmatrix},$$

which requires $O(n^2)$ computation and memory costs, and this can be prohibitive when the size n is extremely large.

As a viable solution, we use a finite-dimensional space to approximate infinite-dimensional $\mathcal{H}_k$ so that $\mu_k(P_i)$ for $i = 1, 2, \cdots, n$ can be reasonably approximated by elements in the finite-dimensional space.

S2.1.1 Approximation with a Shift-Invariant Kernel

To see how to approximate the infinite-dimensional $\mathcal{H}_k$ with a finite-dimensional space, we consider the method of Lopez-Paz et al. (2015), where the kernel function k is assumed shift-invariant. That is, there exists a function $k' : \mathcal{Z} \to \mathbb{R}$ such that

$$k(z, z') = k^{'}(z - z') \quad \text{for any } z, z' \in \mathcal{Z}.$$

Without loss of generality, assume that $\mathcal{Z} = \mathbb{R}^d$ with d being a positive integer.

For a real-valued and shift-invariant kernel function k, Bochner's Theorem (Rudin, 1962) showed that the kernel function k can be expressed as:

$$k(z, z') = 2C_k\mathbb{E}[cos(\langle \omega, z\rangle + b)cos(\langle \omega, z'\rangle + b)] \tag{S.47}$$

for any $z, z' \in \mathcal{Z}$, where ω and b are independent random variables having the distribution $\omega \sim \frac{1}{C_k}p_k$ and $b \sim \mathcal{U}[0, 2\pi]$, $\mathcal{U}[0, 2\pi]$ represents the uniform distribution over $[0, 2\pi]$, p_k is the Fourier transformation of the function $k'(z-z')$, $C_k = \int_{\mathcal{Z}} p_k(\omega)d\omega$, and the expectation in (S.47) is evaluated with respect to the joint distribution of ω and b (Lopez-Paz et al. 2015).

With the Gaussian kernel function (S.1), which is shift-invariant, Lopez-Paz et al. (2015) showed that the Fourier transformation p_k and C_k associated with (S.47) are:

$$p_k(\omega) = (2\pi)^{-\frac{d}{2}}(2\gamma)^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}(2\gamma\omega^{\mathrm{T}}\omega)\Big\} \text{ and } C_k = 1,$$

where $\gamma > 0$ is an user-specified hyper parameter.

Often, the expectation in (S.47) has no analytical form and is approximated by a sample mean. Suppose that for a given positive integer r, for $t = 1, \cdots, r$, the (ω_t, b_t) are independently generated from $\omega_t \sim \frac{1}{C_k}p_k$ and $b_t \sim \mathcal{U}[0, 2\pi]$. Then the function $k(z, \cdot)$ in (S.47) is approximated by

$$\hat{g}_r^z(\cdot) = \frac{1}{r}\sum_{t=1}^{r} 2C_k cos(\langle \omega_t, z\rangle + b_t)cos(\langle \omega_t, \cdot\rangle + b_t), \tag{S.48}$$

where the dependence of $\hat{g}_r^z(\cdot)$ on the underlying kernel function is suppressed in the notation.

Consequently, we approximate $\mu_k(P_{\mathcal{S}_i})$ in (S.7) by:

$$\hat{\mu}(P_{\mathcal{S}_i}) = \frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot). \tag{S.49}$$

S2.1.2 Input Space Expansion

For any probability measure Q' on $\mathcal{Z}$, let

$$L^2(Q') \triangleq \Big\{h : \mathcal{Z} \to \mathbb{R} \;\Big|\; h \text{ is measurable and } \int_{z\in\mathcal{Z}} \big\{h(z)\big\}^2 dQ^{'}(z) < \infty\Big\}$$

denote the L^2 space associated with the measure Q'. For any $h \in L^2(Q')$, let $||h||_{L^2(Q')} \triangleq \Big[\int_{z\in\mathcal{Z}} \big\{h(z)\big\}^2 dQ^{'}(z)\Big]^{\frac{1}{2}}$ denote the $L^2(Q')$ norm of h (Brézis 2011, Section 4.2).

For any shift-invariant kernel function k satisfying $\sup_{z\in\mathcal{Z}} k(z,z) \le 1$, the proof of Lemma 3 in Lopez-Paz et al. (2015) shows that

$$\mathcal{H}_k \subseteq L^2(Q'). \tag{S.50}$$

Lemma 1 in Lopez-Paz et al. (2015) shows that for each i,

$$||\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot) - \mu_k(P_{\mathcal{S}_i})||_{L^2(Q')} \xrightarrow{p} 0 \quad \text{as } r \to \infty. \tag{S.51}$$

That is, by (S.48), $\big\{\mu_k(P_{\mathcal{S}_i}) \;\big|\; i = 1, \cdots, n\big\}$ can be reasonably approximated by some elements in the finite-dimensional subspace of $L^2(Q')$, denoted $\mathcal{S}_{k,r}(Q')$, spanned by $\big\{cos(\langle\omega_t, \cdot\rangle + b_t) \;\big|\; t = 1, \cdots, r\big\}$ for a large r, where ω_t and b_t are defined in the paragraph before (S.48) for $t = 1, \cdots, r$.

Specifically, for a given r, let

$$e_r(\cdot) \triangleq (cos(\langle\omega_1, \cdot\rangle + b_1), \cdots, cos(\langle\omega_r, \cdot\rangle + b_r))^{\mathrm{T}}, \tag{S.52}$$

then $\mathcal{S}_{k,r}(Q')$ is formulated as

$$\mathcal{S}_{k,r}(Q') = \left\{\alpha^{\mathrm{T}} e_r(\cdot) \;\middle|\; \alpha \in \mathbb{R}^r\right\}, \tag{S.53}$$

where subscript k shows the underlying kernel function k.

By (S.48), $\hat{g}_r^{Z_{ij}}(\cdot)$ is a linear combination of the bases $e_r(\cdot)$ for any i and j, and thus, (S.53) implies that $\hat{g}_r^{Z_{ij}}(\cdot)$ belongs to $\mathcal{S}_{k,r}(Q')$ for any i and j. On the other hand, (S.53) shows that $\mathcal{S}_{k,r}(Q')$ is a linear space, that is, any linear combination of the elements in $\mathcal{S}_{k,r}(Q')$ must belong to $\mathcal{S}_{k,r}(Q')$. Therefore, given that $\hat{g}_r^{Z_{ij}}(\cdot) \in \mathcal{S}_{k,r}(Q')$ for any i and j, the approximation $\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)$ of the input $\mu_k(P_{\mathcal{S}_i})$ in (S.49) belongs to the the finite-dimensional approximation space $\mathcal{S}_{k,r}(Q')$ for each i. Consequently, the causal learning procedure in Section 2.2 in the main text can be modified as classification in $\mathcal{S}_{k,r}(Q')$, where we use $\left\{\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot) \;\middle|\; i = 1, \cdots, n\right\}$ as inputs, together with labels $\left\{l_i \middle| i = 1, \cdots, n\right\}$ to train a classifier.

S2.1.3 Modified Classifier with Clean Data

By (S.50), the RKHS $\mathcal{H}_k$ is part of $L^2(Q')$ to which the approximation $\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)$ of the input $\mu_k(P_{\mathcal{S}_i})$ belongs, indicating that the approximated inputs $\left\{\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot) \;\middle|\; i = 1, \cdots, n\right\}$ may not belong to RKHS

$\mathcal{H}_k$. Consequently, to facilitate the rigorous exposition, here we consider the discriminant functional from $L^2(Q')$ rather than $\mathcal{H}_k$ to $\mathbb{R}$. To be specific, let $\mathcal{F}_{Q'}$ denote a set of the functionals mapping from $L^2(Q')$ to $\mathbb{R}$ that are of interest, then the class of candidate classifiers is formulated as $\{\text{sign}(f) \mid f \in \mathcal{F}_{Q'}\}$. Similar to (3.4) in Section 3 of the main text, we aim to find the optimal discriminant functional $\tilde{f}_0 \in \mathcal{F}_{Q'}$ that minimizes the φ-risk:

$$\tilde{f}_0 = \text{argmin}_{f \in \mathcal{F}_{Q'}} R_\varphi(f), \tag{S.54}$$

where $R_\varphi(\cdot)$ is defined at the end of the paragraph before (3.4) of the main text. By the definition of the kernel mean embedding of probability distributions, we have that $\mu_k(P) \in \mathcal{H}_k$ for any probability distribution P, then (S.50) implies $\mu_k(P) \in L^2(Q')$. Therefore, using $\mu_k(P)$ as the input of f in the definition of $R_\varphi(\cdot)$ is well-defined for any $f \in \mathcal{F}_{Q'}$.

We note that the approximation $\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)$ of the input $\mu_k(P_{\mathcal{S}_i})$ falls in the r-dimensional subspace $\mathcal{S}_{k,r}(Q')$ of $L^2(Q')$, which is a space of functions. Thus, directly taking the function $\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)$ as an input to train a classifier may be practically infeasible. To get around this, we equivalently express any element in $\mathcal{S}_{k,r}(Q')$ using its coordinates over the bases $e_r(\cdot)$ defined in (S.52), and then use those coordinates as an input to train a classifier.

Specifically, as opposed to (S.48), we define

$$\mu_{k,r}(P_{\mathcal{S}_i}) = \frac{2C_k}{|\mathcal{S}_i|}\Big(\sum_{Z\in\mathcal{S}_i} cos(\langle\omega_1, Z\rangle + b_1), \sum_{Z\in\mathcal{S}_i} cos(\langle\omega_2, Z\rangle + b_2), \\ \cdots, \sum_{Z\in\mathcal{S}_i} cos(\langle\omega_r, Z\rangle + b_r)\Big)^{\mathrm{T}}, \tag{S.55}$$

for $i = 1, \cdots, n$, where $P_{\mathcal{S}_i}$ is the empirical distribution derived from the samples $\mathcal{S}_i$, defined in (2.1) of the main text, C_k is a constant related to the kernel function k, introduced in (S.47), $|\mathcal{S}_i|$ represents the sample size of $\mathcal{S}_i$, and $\big\{\omega_t \bigm| t = 1, \cdots, r\big\}$ and $\big\{b_t \bigm| t = 1, \cdots, r\big\}$ are realizations generated from two probability distributions introduced before (S.48).

Consequently, (S.48) yields that

$$\begin{aligned}\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot) &= \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{r}\sum_{l=1}^{r} 2C_k cos(\langle\omega_l, Z_{ij}\rangle + b_l) cos(\langle\omega_l, \cdot\rangle + b_l) \\ &= \frac{1}{r}\sum_{l=1}^{r}\Big[\Big\{\frac{2C_k}{m_i}\sum_{j=1}^{m_i} cos(\langle\omega_l, Z_{ij}\rangle + b_l)\Big\}\cdot cos(\langle\omega_l, \cdot\rangle + b_l)\Big] \\ &= \frac{1}{r}\big\{\mu_{k,r}(P_{\mathcal{S}_i})\big\}^{\mathrm{T}} e_r(\cdot), \end{aligned} \tag{S.56}$$

showing that the vector of the coordinates of $\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)$ over the bases $e_r(\cdot)$ is $\frac{1}{r}\mu_{k,r}(P_{\mathcal{S}_i})$.

As considered by Lopez-Paz et al. (2015), we take r times the coordinate of $\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)$ over the bases $e_r(\cdot)$, i.e., $\mu_{k,r}(P_{\mathcal{S}_i})$ defined by (S.55), as the input in learning a classifier. That is, in our data analysis in Section 6 of the main text and Section S4, we train the classifier using the

r-dimensional approximated $\big\{(\mu_{k,r}(P_{\mathcal{S}_i}), l_i) \;\big|\; i = 1, \cdots, n\big\}$ vector rather than the infinite-dimensional kernel mean embeddings.

These r-dimensional approximate vectors can be computed in $O(r)$ time and stored in $O(1)$ memory, leading to much less computation and memory costs compared with the classification in the infinite-dimensional space $\mathcal{H}_k$ when the size n is large. In addition, they can be used in conjunction with any available classification algorithm.

S2.2 Modified Classifiers with Label Noise

In the presence of label noise, the naive and correction classifiers described in Sections 3 and 4 of the main text can be modified in a manner similar to (S.54), with the approximation (S.49) replacing $\mu_k(P_{\mathcal{S}_i})$, where the kernel function k is taken as shift-invariant.

Specifically, with $\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)$ in (S.49) approximating $\mu_k(P_{\mathcal{S}_i})$ for each i, finding the naive classifier (4.11) is modified to finding the optimal $\tilde{f}^* \in \mathcal{F}_{Q'}$ such that

$$\tilde{f}^* = \operatorname{argmin}_{f\in\mathcal{F}_{Q'}} \tilde{R}^*_\varphi(f), \tag{S.57}$$

where $\tilde{R}^*_\varphi(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \varphi\bigg(-f\Big\{\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\Big\} l_i^*\bigg)$.

Similarly, the *corrected* classifier (5.18) is modified as

$$\tilde{f}^{correct} = \operatorname{argmin}_{f\in\mathcal{F}_{Q'}} \tilde{R}_{\varphi^*}(f), \tag{S.58}$$

where $\tilde{R}_{\varphi^*}(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \varphi^*\Big\{f\Big(\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\Big), l_i^*\Big\}$.

Here, we investigate the performance of the naive classifier $\tilde{f}^*$ in (S.57) and the correction classifier $\tilde{f}^{correct}$ in (S.58) by examining upper bounds of $\mathbb{E}\{R_\varphi(\tilde{f}^*)\} - \mathbb{E}\{R_\varphi(\tilde{f}_0)\}$ and $\mathbb{E}\{R_\varphi(\tilde{f}^{correct})\} - \mathbb{E}\{R_\varphi(\tilde{f}_0)\}$. As in Section 3 of the main text, $m = \min\limits_{1\le i\le n} m_i$.

Theorem S2. *Let $\mathcal{Z} = \mathbb{R}^d$ and let Q' be any probability measure on $\mathcal{Z}$. Assume conditions (R3)-(R4) in Theorem 1 in the main text. Furthermore, assume the following two conditions that modify conditions (R1) and (R2) in Theorem 1 in the main text:*

(R1)$'$. *All elements in $\mathcal{F}_{Q'}$ are Lipschitz continuous with respect to the norm in $L^2(Q')$, and there exists a common Lipschitz constant, denoted $L_\mathcal{F}$, for all elements in $\mathcal{F}_{Q'}$. That is, for any $f \in \mathcal{F}_{Q'}$ and $h, h' \in L^2(Q')$, $|f(h) - f(h')| \le L_\mathcal{F}||h - h'||_{L^2(Q')}$;*

(R2)$'$. *There exists a positive constant B such that $\varphi(-f(h)l) \le B$ for any $f \in \mathcal{F}_{Q'}$, $h \in L^2(Q')$, and $l \in \mathcal{L} \triangleq \{-1, 1\}$.*

In addition, assume the following conditions that modify conditions (R5) and (R6) in Theorem 3 in the main text:

(R5)$'$. *all elements in $\mathcal{F}_{Q'}$ are uniformly bounded, that is, there exists a positive constant M' such that for any $f \in \mathcal{F}_{Q'}$ and $h \in L^2(Q')$,*

$|f(h)| \leq M^{'}||h||_{L^2(Q')}$.

$(R6)'$. *the kernel function k is shift-invariant, that is, there exists a positive definite function $k'(\cdot)$ such that $k'(z-z') = k(z, z')$ for any $z, z' \in \mathcal{Z}$.*

Then for $\tilde{f}_0$ in (S.54) and $\tilde{f}^*$ in (S.57),

$$\limsup_{n\to\infty} \limsup_{m\to\infty} \limsup_{r\to\infty} \mathbb{E}\{R_\varphi(\tilde{f}^*) - R_\varphi(\tilde{f}_0)\} \leq 4M^{'}L_\varphi AD.$$

Theorem S2 shows that as r approaches infinity, the asymptotic bias of the naive classification induced from the r-dimensional approximation is upper bounded by that derived from using the infinite-dimensional projections $\{\mu_k(P_{\mathcal{S}_i}) : i = 1, \cdots, n\}$ as stated in Theorem 3 of the main text.

Theorem S3. *Let $\mathcal{Z} = \mathbb{R}^d$ and let Q' be any probability measure on $\mathcal{Z}$. Assume conditions (R3)-(R4) in Theorem 1 in the main text and conditions $(R1)'$, $(R2)'$, and $(R6)'$ in Theorem S2. Then for $\tilde{f}_0$ in (S.54) and $\tilde{f}^{correct}$ in (S.58),*

(a).

$$\begin{aligned}
&\mathbb{E}\{R_\varphi(\tilde{f}^{correct}) - R_\varphi(\tilde{f}_0)\} \\
&\leq C\Big(n, \frac{1}{n}, L_\varphi, L_\mathcal{F}, B^*\Big) + \frac{4B^*}{n} \\
&\quad + \frac{2L^*_\varphi L_\mathcal{F}}{n} \sum_{i=1}^{n} \left\{ \frac{2C_k}{\sqrt{r}} \Big(1 + \sqrt{2log(rm_i)}\Big) + \frac{4|C_k|}{r} \right\},
\end{aligned}$$

where $C(\cdot,\cdot,\cdot,\cdot,\cdot)$ is defined in (3.6) in the main text, and B^* is defined in (5.20) in the main text.

(b).

$$\lim_{n\to\infty}\lim_{m\to\infty}\lim_{r\to\infty}\mathbb{E}\{R_\varphi(\tilde{f}^{correct})-R_\varphi(\tilde{f}_0)\}=0,$$

where $m=\min\limits_{1\le i\le n} m_i$, defined in Section 3 of the main text.

The proofs of Theorems S2 and S3 are deferred to Section S2.3. Theorem S3 (a) characterizes an upper bound for $\mathbb{E}\{R_\varphi(\tilde{f}^{correct})-R_\varphi(\tilde{f}_0)\}$ valid under finite settings. Theorem S3 (b) shows that $\mathbb{E}\{R_\varphi(\tilde{f}^{correct})\}$ converges to $\mathbb{E}\{R_\varphi(\tilde{f}_0)\}$ as $r\to\infty$, $m\to\infty$, and $n\to\infty$, suggesting that the *corrected* classifier induced from the r-dimensional approximation is *consistent* in terms of φ-risk as r, m, and n approach infinity.

The applicability of Theorem S3 depends on the validity of its associated conditions. Condition $(R1)'$ on Lipschitz continuity is widely used in the machine learning community, as discussed in Section 3 in the main text. A straightforward example of $\mathcal{F}$ satisfying condition $(R1)'$ is a class of bounded linear functionals with a common bound. Conditions $(R2)'$ and (R3) are mild and commonly hold in applications. For example, if the convex surrogate φ is continuous and for all $f\in\mathcal{F}$, $|f|$ is uniformly upper bounded, then Condition $(R2)'$ holds; and if φ is further continuously differentiable, Condition (R3) is also satisfied. Conditions (R4) and $(R6)'$ are related to the choice of kernel functions. Many commonly used kernel functions, such as Gaussian kernel defined in (S.1), satisfy these conditions.

In Section S2.4, we examine these conditions when φ is chosen for logistic or hinge loss and when $\mathcal{F}$ is specified as $\mathcal{F}_r$, as defined in Section 6.1 of the main text.

S2.3 Proofs of Theorems S2 and S3

S2.3.1 Proof of Theorem S2

First, we examine the difference $R_\varphi(\tilde{f}^*) - R_\varphi(\tilde{f}_0)$ by connecting it with $\hat{R}^*_\varphi(\cdot)$, $\tilde{R}^*_\varphi(\cdot)$, and $\mathcal{F}$:

$$\begin{aligned}
& R_\varphi(\tilde{f}^*) - R_\varphi(\tilde{f}_0) \\
&= R_\varphi(\tilde{f}^*) - \tilde{R}^*_\varphi(\tilde{f}^*) + \tilde{R}^*_\varphi(\tilde{f}^*) - \tilde{R}^*_\varphi(\tilde{f}_0) + \tilde{R}^*_\varphi(\tilde{f}_0) - R_\varphi(\tilde{f}_0) \\
&\le R_\varphi(\tilde{f}^*) - \tilde{R}^*_\varphi(\tilde{f}^*) + \tilde{R}^*_\varphi(\tilde{f}_0) - R_\varphi(\tilde{f}_0) \\
&\le 2 \sup_{f \in \mathcal{F}_{Q'}} |R_\varphi(f) - \tilde{R}^*_\varphi(f)| \\
&= 2 \sup_{f \in \mathcal{F}_{Q'}} |R_\varphi(f) - \hat{R}^*_\varphi(f) + \hat{R}^*_\varphi(f) - \tilde{R}^*_\varphi(f)| \\
&\le 2 \sup_{f \in \mathcal{F}_{Q'}} |R_\varphi(f) - \hat{R}^*_\varphi(f)| + 2 \sup_{f \in \mathcal{F}_{Q'}} |\hat{R}^*_\varphi(f) - \tilde{R}^*_\varphi(f)| \qquad \text{(S.59)}
\end{aligned}$$

where the first inequality holds since $\tilde{R}^*_\varphi(\tilde{f}^*) - \tilde{R}^*_\varphi(\tilde{f}_0) \leq 0$ by that $\tilde{f}^*$ is the minimum point of the functional $\tilde{R}^*_\varphi(f)$, and the last inequality comes from the triangle inequality and the definition of supremum.

Now we examine the two terms in (S.59) individually in the following two steps.

Step 1: Examining the first term of (S.59).

Noting that the first term in (S.59) is the same as (S.21) except that $\mathcal{F}$ in (S.21) is replaced by $\mathcal{F}_{Q'}$ here, we adapt the proof of Theorem 3 of the main text in Section S1.5 to obtain an upper bound for $\limsup_{n\to\infty} \limsup_{m\to\infty} \mathbb{E}\Big\{2 \sup_{f\in\mathcal{F}_{Q'}} |R_\varphi(f) - \hat{R}^*_\varphi(f)|\Big\}$ by first verifying that the conditions in Theorem 3 of the main text are satisfied if $\mathcal{F}$ is replaced by $\mathcal{F}_{Q'}$. Except for conditions (R1) and (R2) in Theorem 1 in the main text and condition (R5) in Theorem 3 of the main text, all other conditions in Theorem 3 of the main text are obviously satisfied if $\mathcal{F}$ in Theorem 3 of the main text is replaced by $\mathcal{F}_{Q'}$.

Now we examine these three conditions with $\mathcal{F}$ in Theorem 3 of the main text replaced by $\mathcal{F}_{Q'}$. We first examine condition (R1) in Theorem 1 in the main text with $\mathcal{F}$ replaced by $\mathcal{F}_{Q'}$. As stated in the second paragraph after (S.48) in Section S2, the proof of Lemma 3 in Lopez-Paz et al. (2015) shows that $\mathcal{H}_k \subseteq L^2(Q')$. Thus, $||h||_{L^2(Q')} \leq ||h||_{\mathcal{H}_k}$ for any $h \in \mathcal{H}_k$. Therefore, by the first condition of Theorem S2, for any $h, h' \in \mathcal{H}_k$ and $f \in \mathcal{F}_{Q'}$,

$$|f(h) - f(h')| \leq L_{\mathcal{F}}||h - h'||_{L^2(Q')} \leq L_{\mathcal{F}}||h - h'||_{\mathcal{H}_k},$$

showing condition (R1) in Theorem 1 in the main text when $\mathcal{F}$ is replaced by $\mathcal{F}_{Q'}$. By $\mathcal{H}_k \subseteq L^2(Q')$ and condition $(R2)'$ in Theorem S2, condition (R2) in Theorem 1 in the main text holds when $\mathcal{F}$ is replaced by $\mathcal{F}_{Q'}$. Finally, by $\mathcal{H}_k \subseteq L^2(Q')$, $||h||_{L^2(Q')} \leq ||h||_{\mathcal{H}_k}$ for any $h \in \mathcal{H}_k$. Then by condition $(R5)'$ in Theorem S2, there exists a constant $M^{'}$ such that for any $f \in \mathcal{F}_{Q'}$ and $h \in \mathcal{H}_k$,

$$|f(h)| \leq M^{'}||h||_{L^2(Q')} \leq M^{'}||h||_{\mathcal{H}_k},$$

suggesting that condition (R5) in Theorem 3 in the main text holds with $\mathcal{F}$ replaced by $\mathcal{F}_{Q'}$.

Consequently, repeating the proof of Theorem 3 of the main text in Section S1.5 with replacing $\mathcal{F}$ by $\mathcal{F}_{Q'}$, we obtain that

$$\begin{aligned}
&\limsup_{n\to\infty}\limsup_{m\to\infty}\limsup_{r\to\infty}\mathbb{E}\Big\{2\sup_{f\in\mathcal{F}_{Q'}}|R_\varphi(f) - \hat{R}^*_\varphi(f)|\Big\} \\
&= \limsup_{n\to\infty}\limsup_{m\to\infty}\mathbb{E}\Big\{2\sup_{f\in\mathcal{F}_{Q'}}|R_\varphi(f) - \hat{R}^*_\varphi(f)|\Big\} \\
&\leq 4M^{'}L_\varphi AD, \qquad \text{(S.60)}
\end{aligned}$$

where the first equation holds by that $2\sup_{f\in\mathcal{F}_{Q'}}|R_\varphi(f) - \hat{R}^*_\varphi(f)|$ does not involve r and the inequality is due to the results of Theorem 3 in the main text.

Step 2: Examining the second term in (S.59):

$$
\begin{aligned}
&\sup_{f\in\mathcal{F}_{Q'}}\Big|\hat{R}^*_\varphi(f)-\tilde{R}^*_\varphi(f)\Big|\\
&=\sup_{f\in\mathcal{F}_{Q'}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i^*)-\frac{1}{n}\sum_{i=1}^{n}\varphi\Big\{-f\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big)l_i^*\Big\}\Big|\\
&\le\frac{1}{n}\sup_{f\in\mathcal{F}_{Q'}}\sum_{i=1}^{n}\Big|\varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i^*)-\varphi\Big\{-f\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big)l_i^*\Big\}\Big|\\
&\le\frac{1}{n}\sum_{i=1}^{n}\sup_{f\in\mathcal{F}_{Q'}}\Big|\varphi(-f(\mu_k(P_{\mathcal{S}_i}))l_i^*)-\varphi\Big\{-f\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big)l_i^*\Big\}\Big|\\
&\le\frac{L_\varphi}{n}\sum_{i=1}^{n}\sup_{f\in\mathcal{F}_{Q'}}\Big|f(\mu_k(P_{\mathcal{S}_i}))l_i^*-f\Big\{\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\}l_i^*\Big|\\
&=\frac{L_\varphi}{n}\sum_{i=1}^{n}\sup_{f\in\mathcal{F}_{Q'}}\Big|f(\mu_k(P_{\mathcal{S}_i}))-f\Big\{\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\}\Big|\\
&\le\frac{L_\varphi L_{\mathcal{F}}}{n}\sum_{i=1}^{n}\sup_{f\in\mathcal{F}_{Q'}}\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\\
&=\frac{L_\varphi L_{\mathcal{F}}}{n}\sum_{i=1}^{n}\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')},
\end{aligned}
\tag{S.61}
$$

where the first inequality is due to the Jensen's inequality, the second inequality comes from that $\sup_x\{|g_1(x)|+|g_2(x)|\}\le\sup_x|g_1(x)|+\sup_x|g_2(x)|$ for any functions g_1 and g_2, the third inequality is due to Lipschitzness of φ, the fourth inequality is due to Lipschitzness of f, and the last equality holds because the expression does not depend on f.

Now we examine an upper bound for the expectation of the summands in (S.61), $\mathbb{E}\Big\{\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\Big\}$, which is derived in the following three parts.

Part 1: We show that

$$\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} \leq 4|C_k| \tag{S.62}$$

for Z_{ij} evaluated at every $\omega \in \Omega$. Here and below, for ease of exposition, when presenting an inequality involving a random variable, it is meant to be evaluated for every $\omega \in \Omega$. For instance, the inequality "$\big|k(Z_{ij},\cdot)\big| \leq a$" should be understood as "$\big|k(Z_{ij}(\omega),\cdot)\big| \leq a$ for any $\omega \in \Omega$".

By (S.47), we have that for any $z, z' \in \mathcal{Z}$,

$$\begin{aligned}|k(z,z')| &= \big|2C_k\mathbb{E}\big\{cos(\langle\omega,z\rangle + b)cos(\langle\omega,z'\rangle + b)\big\}\big| \\ &\leq 2\big|C_k\big|\mathbb{E}\Big\{\big|cos(\langle\omega,z\rangle + b)cos(\langle\omega,z'\rangle + b)\big|\Big\} \\ &\leq 2|C_k|, \end{aligned} \tag{S.63}$$

where the first inequality is due to Jensen's inequality and the second inequality is due to the fact that $|cos(x)| \leq 1$ for any $x \in \mathbb{R}$. Then by (S.7), we have that

$$\begin{aligned}\big|\mu_k(P_{\mathcal{S}_i})\big| &\leq \frac{1}{m_i}\sum_{j=1}^{m_i}\big|k(Z_{ij},\cdot)\big| \\ &\leq \frac{1}{m_i}\sum_{j=1}^{m_i} 2|C_k| \\ &= 2|C_k|, \end{aligned} \tag{S.64}$$

where the first inequality is due to Jensen's inequality and the second inequality comes from (S.63) with the inequality understood to hold for all

$\omega \in \Omega$.

By (S.48), we have that

$$
\begin{aligned}
\Big|\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big| &\le \frac{1}{m_i}\sum_{j=1}^{m_i}\Big|\hat{g}_r^{Z_{ij}}(\cdot)\Big| \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i}\Big|\frac{1}{r}\sum_{i=1}^{r}\big\{2C_k cos(\langle\omega_i, Z_{ij}\rangle + b_i)cos(\langle\omega_i, \cdot\rangle + b_i)\big\}\Big| \\
&\le \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{r}\sum_{i=1}^{r}\big\{|2C_k cos(\langle\omega_i, Z_{ij}\rangle + b_i)cos(\langle\omega_i, \cdot\rangle + b_i)|\big\} \\
&\le \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{r}\sum_{i=1}^{r}2|C_k| \\
&= 2|C_k|, \qquad \text{(S.65)}
\end{aligned}
$$

where the first and second inequalities are due to Jensen's inequality and the third inequality is due to the fact that $|cos(x)| \le 1$ for any $x \in \mathbb{R}$.

Then by the definition of the L^2 norm in $L^2(Q')$ (Section 4.2, Brézis 2011) and (S.64) (S.65), we further have that

$$||\mu_k(P_{\mathcal{S}_i})||_{L^2(Q')} \le 2|C_k|, \qquad \text{(S.66)}$$

and

$$\Big\|\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} \le 2|C_k|. \qquad \text{(S.67)}$$

Therefore, by the triangle inequality and (S.66) (S.67),

$$\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} \le ||\mu_k(P_{\mathcal{S}_i})||_{L^2(Q')} + \Big\|\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}$$

$$\leq 4|C_k|.$$

Part 2: Next, we show that $\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}$ is upper bounded with a probability greater than $1-\frac{1}{r}$:

For $i = 1, \cdots, n$ and any $\delta > 0$, we obtain that by Lemma 1 of Lopez-Paz et al. (2015),

$$\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} \leq \frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\log(\frac{m_i}{\delta})}\Big)$$

with probability larger than $1-\delta$. Then taking $\delta = \frac{1}{r}$ gives that

$$\mathbb{P}\Big\{\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} \leq \frac{2C_k}{\sqrt{r}}\big(1+\sqrt{2\log(rm_i)}\big)\Big\} \geq 1-\frac{1}{r}. \tag{S.68}$$

Part 3: Finally, we examine the expectation of $\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}$:

$$\begin{aligned}
&\mathbb{E}\Big\{\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\Big\}\\
&= \mathbb{E}\Bigg\{\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\\
&\quad\times\Bigg[I\Big\{\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} \leq \frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\log(rm_i)}\Big)\Big\}\\
&\qquad+I\Big\{\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} > \frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\log(rm_i)}\Big)\Big\}\Bigg]\Bigg\}\\
&= \mathbb{E}\Bigg[\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\\
&\quad\times I\Big\{\Big\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} \leq \frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\log(rm_i)}\Big)\Big\}\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}\Bigg[\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')} \\
&\quad\times I\Big\{\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}>\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big)\Big\}\Bigg] \\
&\leq \mathbb{E}\Bigg[\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big) \\
&\quad\times I\Big\{\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\leq\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big)\Big\}\Bigg] \\
&+\mathbb{E}\Bigg[4|C_k|I\Big\{\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}>\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big)\Big\}\Bigg] \\
&=\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big)\mathbb{P}\Bigg\{\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\leq\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big)\Bigg\} \\
&+4|C_k|\mathbb{P}\Bigg\{\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}>\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big)\Bigg\} \\
&\leq\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big)+\frac{4|C_k|}{r}, \qquad\qquad (S.69)
\end{aligned}
$$

where the first inequality is due to (S.62) and the second inequality is due to (S.68).

Therefore,

$$
\begin{aligned}
0&\leq\lim_{r\to\infty}\mathbb{E}\Big\{\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\Big\} \\
&\leq\lim_{r\to\infty}\Big\{\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\text{log}(rm_i)}\Big)+\frac{4|C_k|}{r}\Big\} \\
&=0, \qquad\qquad (S.70)
\end{aligned}
$$

and thus,

$$\lim_{n\to\infty}\lim_{m\to\infty}\lim_{r\to\infty}\mathbb{E}\Big\{\Big\|\mu_k(P_{\mathcal{S}_i})-\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)\Big\|_{L^2(Q')}\Big\}=0. \tag{S.71}$$

Combining (S.61) and (S.71) gives that

$$\lim_{n\to\infty}\lim_{m\to\infty}\lim_{r\to\infty}\mathbb{E}\Big\{2\sup_{f\in\mathcal{F}_{Q'}}\Big|\hat{R}^*_{\varphi}(f)-\tilde{R}^*_{\varphi}(f)\Big|\Big\}=0. \tag{S.72}$$

Applying (S.60) and (S.72) to (S.59) proves Theorem S2.

S2.3.2 Proof of Theorem S3

Part (b) is immediate by taking the limit with respect to r, m, and n on both sides of part (a). It now remains to show Part (a).

First, we examine the difference $R_{\varphi}(\tilde{f}^{correct})-R_{\varphi}(\tilde{f}_0)$ by connecting it with $R_{\varphi^*}(\cdot)$, $\tilde{R}_{\varphi^*}(\cdot)$, $\hat{R}_{\varphi^*}(\cdot)$, and $\mathcal{F}$:

$$\begin{aligned}
&R_{\varphi}(\tilde{f}^{correct})-R_{\varphi}(\tilde{f}_0)\\
&=R_{\varphi^*}(\tilde{f}^{correct})-R_{\varphi^*}(\tilde{f}_0)\\
&=R_{\varphi^*}(\tilde{f}^{correct})-\tilde{R}_{\varphi^*}(\tilde{f}^{correct})+\tilde{R}_{\varphi^*}(\tilde{f}^{correct})-\tilde{R}_{\varphi^*}(\tilde{f}_0)+\tilde{R}_{\varphi^*}(\tilde{f}_0)-R_{\varphi^*}(\tilde{f}_0)\\
&\le R_{\varphi^*}(\tilde{f}^{correct})-\tilde{R}_{\varphi^*}(\tilde{f}^{correct})+\tilde{R}_{\varphi^*}(\tilde{f}_0)-R_{\varphi^*}(\tilde{f}_0)\\
&\le 2\sup_{f\in\mathcal{F}_{Q'}}\big|R_{\varphi^*}(f)-\tilde{R}_{\varphi^*}(f)\big|\\
&=2\sup_{f\in\mathcal{F}_{Q'}}\big|R_{\varphi^*}(f)-\hat{R}_{\varphi^*}(f)+\hat{R}_{\varphi^*}(f)-\tilde{R}_{\varphi^*}(f)\big|\\
&\le 2\sup_{f\in\mathcal{F}_{Q'}}\big|R_{\varphi^*}(f)-\hat{R}_{\varphi^*}(f)\big|+2\sup_{f\in\mathcal{F}_{Q'}}\big|\hat{R}_{\varphi^*}(f)-\tilde{R}_{\varphi^*}(f)\big|, \qquad \text{(S.73)}
\end{aligned}$$

where the first equality is due to Theorem 4 in the main text, the first inequality holds since $\tilde{R}_{\varphi^*}(\tilde{f}^{correct}) - \tilde{R}_{\varphi^*}(\tilde{f}_0) \leq 0$ by that $\tilde{f}^{correct}$ is the minimum point of the functional $\tilde{R}_{\varphi^*}(\cdot)$, the second inequality is due to the property of the supremum, and the last inequality comes from the triangle inequality and the definition of supremum.

Now we examine the two terms in (S.73) individually in the following two steps.

Step 1: Examining the first term of (S.73).

By applying Lemmas 3 and 4 in Section S1.7 to the proof of Theorem 3 of Lopez-Paz et al. (2015) by letting their $\delta = \frac{1}{n}$, we have that

$$\mathbb{P}\Big\{2 \sup_{f \in \mathcal{F}_{Q'}} \big|R_{\varphi^*}(f) - \hat{R}_{\varphi^*}(f)\big| \leq C(n, m, L_{\varphi}^*, L_{\mathcal{F}}, B^*)\Big\} \geq 1 - \frac{1}{n},$$

where L_{φ}^*, B^*, and $C(\cdot,\cdot,\cdot,\cdot,\cdot)$ are defined in (5.19), (5.20), and (3.6) in the main text, respectively.

By the proof of Theorem S2, the conditions in Theorem S3 implies the conditions in Theorem 1 in the main text. Then repeating the calculation of $\mathbb{E}\Big\{2\sup_{f\in\mathcal{F}}\big|R_{\varphi}(f) - \hat{R}_{\varphi}(f)\big|\Big\}$ in Section S1.4 by replacing B, φ and $\mathcal{F}$ with B^*, φ^* and $\mathcal{F}_{Q'}$, respectively, we obtain that

$$\mathbb{E}\Big\{2 \sup_{f \in \mathcal{F}_{Q'}} \big|R_{\varphi^*}(f) - \hat{R}_{\varphi^*}(f)\big|\Big\} = C\Big(n, m, L_{\varphi}^*, L_{\mathcal{F}}, B^*\Big) + \frac{4B^*}{n}. \tag{S.74}$$

Step 2: Examining the second term of (S.73).

By (5.17) in the main text and the definition of $\tilde{R}_{\varphi^*}(f)$ after (S.58), we obtain that

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_{Q'}} \left|\hat{R}_{\varphi^*}(f) - \tilde{R}_{\varphi^*}(f)\right| \\
&= \sup_{f\in\mathcal{F}_{Q'}} \left|\frac{1}{n}\sum_{i=1}^{n} \varphi^*(f(\mu_k(P_{\mathcal{S}_i})), l_i^*) - \frac{1}{n}\sum_{i=1}^{n} \varphi^*\Big\{f\Big(\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\Big), l_i^*\Big\}\right| \\
&\leq \frac{1}{n}\sup_{f\in\mathcal{F}_{Q'}} \sum_{i=1}^{n} \left|\varphi^*(f(\mu_k(P_{\mathcal{S}_i})), l_i^*) - \varphi^*\Big\{f\Big(\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\Big), l_i^*\Big\}\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \sup_{f\in\mathcal{F}_{Q'}} \left|\varphi^*(f(\mu_k(P_{\mathcal{S}_i})), l_i^*) - \varphi^*\Big\{f\Big(\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\Big), l_i^*\Big\}\right| \\
&\leq \frac{L_\varphi^*}{n}\sum_{i=1}^{n} \sup_{f\in\mathcal{F}_{Q'}} \left|f(\mu_k(P_{\mathcal{S}_i})) - f\Big\{\frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\Big\}\right| \\
&\leq \frac{L_\varphi^* L_{\mathcal{F}}}{n}\sum_{i=1}^{n} \sup_{f\in\mathcal{F}_{Q'}} \left\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\right\|_{L^2(Q')} \\
&= \frac{L_\varphi^* L_{\mathcal{F}}}{n}\sum_{i=1}^{n} \left\|\mu_k(P_{\mathcal{S}_i}) - \frac{1}{m_i}\sum_{j=1}^{m_i} \hat{g}_r^{Z_{ij}}(\cdot)\right\|_{L^2(Q')}, \qquad \text{(S.75)}
\end{aligned}$$

where the first inequality is due to the triangle inequality, the second inequality comes from that $\sup_x\{|g_1(x)|+|g_2(x)|\} \leq \sup_x|g_1(x)|+\sup_x|g_2(x)|$ for any functions g_1 and g_2, the third inequality is due to Lemma 3 in Section S1.7, the fourth inequality is due to Lipschitzness of f, and the last equality holds because the expression does not depend on f.

Then combining (S.69) and (S.75) yields that

$$\mathbb{E}\Big\{2\sup_{f\in\mathcal{F}_{Q'}}\left|\hat{R}_{\varphi^*}(f)-\tilde{R}_{\varphi^*}(f)\right|\Big\} \leq \frac{2L_\varphi^* L_{\mathcal{F}}}{n}\sum_{i=1}^{n}\Big\{\frac{2C_k}{\sqrt{r}}\Big(1+\sqrt{2\log(rm_i)}\Big)+\frac{4|C_k|}{r}\Big\}. \qquad \text{(S.76)}$$

Applying (S.74) and (S.76) to (S.73) proves part (a) of Theorem S3.

S2.4 Examination of the Conditions for Theorem S3

As discussed in Section S2, we train the true, naive and corrected classifiers in an r-dimensional approximated space rather than the original infinite-dimensional space $\mathcal{H}_k$, whose theoretical validity of the true and corrected classifiers requires the conditions in Theorem S3. In particular, condition $(R1)^{'}$ about the Lipschitz continuity modifies condition (R1) in Theorem 1 of the main text to accommodate the approximation of the inputs. As discussed in Section 3 of the main text, the Lipschitz continuity is widely used in machine learning community to ensure stability and convergence of algorithms.

Now we discuss the feasibility of other conditions in Theorem S3 in conducting sensitivity analyses presented in Section 6 of the main text and the simulation studies presented in Section S4, where logistic regression and SVM classifiers are considered.

As discussed in Section S2, we take $\left\{\mu_{k,r}(P_{\mathcal{S}_i}) \;\middle|\; i = 1, \cdots, n\right\}$ as the input, which for each i, by (S.55), satisfies

$$||\mu_{k,r}(P_{\mathcal{S}_i})||_2^2 = \frac{4C_k^2}{|\mathcal{S}_i|^2} \sum_{j=1}^{r} \Big\{ \sum_{Z \in \mathcal{S}_i} cos(\langle \omega_j, Z \rangle + b_j) \Big\}^2$$

$$\begin{aligned}&\leq \frac{4rC_k^2|\mathcal{S}_i|^2}{|\mathcal{S}_i|^2}\\&= 4rC_k^2. \end{aligned} \tag{S.77}$$

where C_k is a constant dependent of the kernel function k, introduced in (S.47) in Section S2.

Let

$$\mathcal{S}_{k,0}(Q') \triangleq \{h \mid h = a^{\mathrm{T}}e_r(\cdot) \text{ with } a \text{ satisfying } ||r \cdot a||_2^2 \leq 4rC_k^2\}, \tag{S.78}$$

where $||\cdot||_2$ is the L_2-norm in $\mathbb{R}^d$ with $||b||_2 = \sqrt{b^{\mathrm{T}}b}$ for $b \in \mathbb{R}^r$. Then by (S.53), it is obvious that $\mathcal{S}_{k,0}(Q') \subseteq \mathcal{S}_{k,r}(Q')$. Moreover, by (S.56) and (S.77), the approximation $\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{g}_r^{Z_{ij}}(\cdot)$ of the input $\mu_k(P_{\mathcal{S}_i})$ belongs to $\mathcal{S}_{k,0}(Q')$ for each i. Then, in our sensitivity analyses and simulation studies, we actually implement the classification in $\mathcal{S}_{k,0}(Q')$. Therefore, each element in $\mathcal{F}_{Q'}$ is restricted to $\mathcal{S}_{k,0}(Q')$. That is, for any $f \in \mathcal{F}_{Q'}$, $f(h) = 0$ if $h \notin \mathcal{S}_{k,0}(Q')$.

As discussed in Section S2, for any element in $\mathcal{S}_{k,r}(Q')$, we use r times its coordinates over the bases $e_r(\cdot)$ defined in (S.52) as the input in our numerical studies, so the class $\mathcal{F}_{Q'}$ of the discriminant functionals corresponding to $\mathcal{F}_r$ defined in Section 6.1 of the main text is given by

$$\begin{aligned}\mathcal{F}_{Q'} \triangleq &\Big\{f \;\Big|\; f(h) = \{w^{\mathrm{T}}(r\cdot a) + c\} \cdot I\{h \in \mathcal{S}_{k,0}(Q')\}\\ &\qquad \text{with } h = a^{\mathrm{T}}e_r(\cdot) \text{ for } h \in \mathcal{S}_{k,0}(Q')),\end{aligned}$$

$$\text{satisfying } ||r \cdot a||_2^2 \leq 4rC_k^2,\ ||w||_2^2 \leq C_r, \text{ and } |c| \leq C_r \Big\} \qquad \text{(S.79)}$$

and

$$\begin{aligned}\mathcal{F}_{Q'} = \Bigg\{ f \ \Bigg| \ & f(h) = \Big\{ \sum_{i=1}^{n} \alpha_i l_i \exp(-||\mu_{k,r}(P_{\mathcal{S}_i}) - (r \cdot a)||_2^2) + b \Big\} \\ & \times I\big\{h \in \mathcal{S}_{k,0}(Q')\big\}, \\ & \text{with } h = a^{\mathrm{T}} e_r(\cdot) \text{ for } h \in \mathcal{S}_{k,0}(Q'), \text{ satisfying } ||r \cdot a||_2^2 \leq 4rC_k^2, \\ & |\alpha_i| \leq C_r \text{ for } i = 1, \cdots, n, \text{ and } |b| \leq C_r \ \Bigg\}, \end{aligned} \qquad \text{(S.80)}$$

respectively, for logistic regression and Gaussian kernel-based SVM classifiers.

With the Gaussian kernel (S.1), it is easy to verify that it satisfies condition (R4) in Theorem 1 in the main text and condition $(R6)'$ in Theorem S2. In what follows, we verify conditions $(R2)'$ and (R3) required by Theorem S3 when logistic regression and Gaussian kernel-based SVM classifiers are used.

S2.4.1 Verification of Condition $(R2)'$

First, we examine logistic regression. By (S.79), for any $f \in \mathcal{F}_{Q'}$, when $h \notin \mathcal{S}_{k,0}(Q')$, we have that

$$f(h) = 0. \qquad \text{(S.81)}$$

Furthermore, by the Cauchy–Schwarz inequality, for any $h = a^{\mathrm{T}} e_r(\cdot) \in \mathcal{S}_{k,0}(Q')$ and $f \in \mathcal{F}_{Q'}$, we have that

$$\begin{aligned} \big|f(h)\big| &= \big|w^{\mathrm{T}}(r \cdot a) + c\big| \\ &\le \big|w^{\mathrm{T}}(r \cdot a)\big| + \big|c\big| \\ &\le ||w||_2 \cdot ||r \cdot a||_2 + \big|c\big| \\ &\le \sqrt{4rC_k^2 C_r} + C_r \\ &= 2C_k\sqrt{rC_r} + C_r. \end{aligned} \tag{S.82}$$

where the last inequality holds because $||r \cdot a||_2^2 \le 4rC_k^2$ by the definition of $\mathcal{S}_{k,0}(Q')$ in (S.78) and $||w||_2^2 \le C_r$, as well as $|c| \le C_r$ by (S.79).

For any $h \in \mathcal{S}_{k,0}(Q')$ and $f \in \mathcal{F}_{Q'}$, combining (S.81) and (S.82) yields that $f(h) \in [-2C_k\sqrt{rC_r} - C_r, 2C_k\sqrt{rC_r} + C_r]$. Then since the logistic loss used in logistic regression is continuous and any continuous function is bounded over a bounded closed set in $\mathbb{R}$, condition $(R2)'$ presented in Theorem S2 holds.

Next, we examine SVM. By (S.80), for any $f \in \mathcal{F}_{Q'}$, when $h \notin \mathcal{S}_{k,0}(Q')$, we have that

$$f(h) = 0. \tag{S.83}$$

Furthermore, for any $h = a^{\mathrm{T}} e_r(\cdot) \in \mathcal{S}_{k,0}(Q')$ and $f \in \mathcal{F}_{Q'}$, we have that

$$|f(h)| = \Big|\sum_{i=1}^{n} \alpha_i l_i \exp\big(-||\mu_{k,r}(P_{\mathcal{S}_i}) - r \cdot a||_2^2\big) + b\Big|$$

$$\begin{aligned}
&\leq \sum_{i=1}^{n} \big|\alpha_i\big| \cdot \big|l_i\big| \cdot \exp\big(- ||\mu_{k,r}(P_{\mathcal{S}_i}) - r \cdot a||_2^2\big) + \big|b\big| \\
&\leq \sum_{i=1}^{n} \big|\alpha_i\big| + \big|b\big| \\
&\leq (n+1)C_r, \qquad \text{(S.84)}
\end{aligned}$$

where the first inequality is due to the triangle inequality of the absolute value, the second inequality comes from the property of exponential function and $\big|l_i\big| = 1$ for all i, and the last inequality is due to $|\alpha_i| \leq C_r$ for $i = 1, \cdots, n$ and $|b| \leq C_r$ by (S.80).

For any $h \in \mathcal{S}_{k,0}(Q')$ and $f \in \mathcal{F}_{Q'}$, combining (S.83) and (S.84) yields that $f(h) \in [-(n+1)C_r, (n+1)C_r]$. Then since the hinge loss used in SVM is continuous and any continuous function is bounded over a bounded closed set in $\mathbb{R}$, condition $(R2)'$ presented in Theorem S2 holds.

S2.4.2 Verification of Condition (R3)

First, we examine logistic regression, where we use the logistic loss $\varphi(\alpha) = \log_2\left(1 + \exp(\alpha)\right)$. By the mean value theorem (Thomas 2014, p.194), for any $\alpha_1, \alpha_2 \in \mathbb{R}$, we have that

$$\varphi(\alpha_1) - \varphi(\alpha_2) = \varphi^{'}(\alpha_0) \cdot (\alpha_1 - \alpha_2) \qquad \text{(S.85)}$$

where α_0 is a constant between α_1 and α_2. Noting that

$$\varphi^{'}(\alpha_0) = \frac{\exp(\alpha_0)}{(\exp(\alpha_0)+1)\log 2} \leq \frac{1}{\log 2},$$

we apply (S.85) and obtain that

$$\big|\varphi(\alpha_1) - \varphi(\alpha_2)\big| = \varphi^{'}(\alpha_0)\big|\alpha_1 - \alpha_2\big| \leq \frac{1}{\log 2}\big|\alpha_1 - \alpha_2\big|.$$

That is, $\varphi(\cdot)$ is a Lipschitz continuous function with a Lipschitz constant being $\frac{1}{\log 2}$. Moreover, for any $\alpha \in \mathbb{R}$, it is easy to verify that $\log_2\,(1 + \exp(\alpha)) \geq \ell(\alpha)$. Therefore, condition (R3) of Theorem 1 in the main text holds for logistic regression.

Next, we examine SVM, where we use the hinge loss $\varphi(\alpha) = \max\{0, 1+\alpha\}$. For any $\alpha_1 \geq -1$ and $\alpha_2 \geq -1$, we have that

$$\varphi(\alpha_1) - \varphi(\alpha_2) = \alpha_1 - \alpha_2,$$

yielding that

$$\big|\varphi(\alpha_1) - \varphi(\alpha_2)\big| = \big|\alpha_1 - \alpha_2\big|. \tag{S.86}$$

For $\alpha_1 \geq -1$ and $\alpha_2 \leq -1$, we have that

$$\varphi(\alpha_1) - \varphi(\alpha_2) = 1 + \alpha_1 = \alpha_1 - (-1) \leq \alpha_1 - \alpha_2,$$

yielding that

$$\big|\varphi(\alpha_1) - \varphi(\alpha_2)\big| \leq \big|\alpha_1 - \alpha_2\big|. \tag{S.87}$$

For $\alpha_1 \leq -1$ and $\alpha_2 \geq -1$, we have that

$$\varphi(\alpha_1) - \varphi(\alpha_2) = 0 - (1 + \alpha_2) = -1 - \alpha_2 \geq \alpha_1 - \alpha_2,$$

yielding that

$$\left|\varphi(\alpha_1) - \varphi(\alpha_2)\right| \leq \left|\alpha_1 - \alpha_2\right|. \tag{S.88}$$

For $\alpha_1 \leq -1$ and $\alpha_2 \leq -1$, we have that

$$\left|\varphi(\alpha_1) - \varphi(\alpha_2)\right| = 0 \leq \left|\alpha_1 - \alpha_2\right|. \tag{S.89}$$

Combining (S.86), (S.87), (S.88), and (S.89) yields that for any $u_1, u_2 \in \mathbb{R}$,

$$\left|\varphi(\alpha_1) - \varphi(\alpha_2)\right| \leq \left|\alpha_1 - \alpha_2\right|.$$

That is, the hinge loss is a Lipschitz continuous function with a Lipschitz constant being 1. Moreover, for any $\alpha \in \mathbb{R}$, it is easy to verify that $\max\{0, 1 + \alpha\} \geq \ell(\alpha)$. Therefore, condition (R3) of Theorem 1 in the main text holds for SVM.

S3 SUP3 Dataset and Additional Analysis Results

S3.1 SUP3 Data

The SUP3 dataset is available at *https://www.kaggle.com/c/cause-effect-pairs/data*. This dataset does not contain any personally identifiable information or offensive content. The dataset includes 162 pairs of variables

$\big\{(X_i, W_i) \bigm| i = 1, \cdots, 162\big\}$ from diverse domains including chemistry, climatology, ecology, economy, engineering, epidemiology, genomics, medicine, physics, and sociology. While the dataset does not include the information about the meaning of X_i and W_i for each i, the information whether X_i is the cause of W_i is provided for $i = 1, \cdots, n$ with $n = 162$. That is, the reported value l_i^*, either 1 or -1, of l_i is included in the dataset, where l_i represents the *true* label for reflecting the causal relationship of $\{X_i, W_i\}$, with $l_i = 1$ indicating that X_i is the cause of W_i and $l_i = -1$ otherwise. As those labels $\big\{l_i^* \bigm| i = 1, \cdots, n\big\}$ are identified based on applying the subjective views to determine the causation of each pair of variables, there is basically a discrepancy between the reported l_i^* and the true label l_i for some pairs. Among those 162 reported l_i^*, 42 of them take 1 and the rest assume -1. For each i, there are realizations of a sequence of i.i.d samples $\mathcal{S}_i = \big\{(x_{ij}, w_{ij})^{\mathrm{T}} \bigm| j = 1, \cdots, m_i\big\}$ of $(X_i, W_i)^{\mathrm{T}}$. Table S.1 reports the values of l_i^* and m_i for $i = 1, \cdots, 162$.

S3.2 Additional Sensitivity Analyses for Section 6.2 of the Main Text

In Section 6.2 of the main text, we investigate the mislabeling effects and the performance of the proposed correction method for several values of r

Table S.1: *Values of l_i^* and m_i for $i = 1, \cdots, 162$.*

i	l_i^*	m_i	i	l_i^*	m_i	i	l_i^*	m_i	i	l_i^*	m_i	i	l_i^*	m_i	i	l_i^*	m_i
1	-1	349	28	-1	1331	55	-1	5559	82	1	254	109	-1	365	136	-1	2027
2	-1	2135	29	-1	392	56	1	365	83	-1	520	110	-1	192	137	-1	768
3	1	347	30	-1	194	57	-1	1030	84	-1	451	111	-1	1030	138	-1	365
4	-1	802	31	-1	349	58	-1	254	85	-1	192	112	-1	314	139	-1	192
5	1	1414	32	-1	1030	59	-1	1030	86	-1	4177	113	1	1030	140	-1	192
6	1	2782	33	-1	392	60	-1	2633	87	-1	1674	114	-1	365	141	1	349
7	1	192	34	1	345	61	-1	345	88	1	564	115	1	192	142	-1	1991
8	1	349	35	-1	194	62	-1	1067	89	-1	452	116	1	1030	143	-1	4177
9	-1	1934	36	-1	3363	63	-1	533	90	-1	538	117	-1	392	144	-1	1030
10	1	205	37	-1	721	64	-1	349	91	-1	1585	118	1	345	145	-1	768
11	-1	452	38	1	345	65	-1	345	92	-1	1030	119	1	349	146	-1	345
12	1	392	39	1	392	66	1	452	93	-1	349	120	-1	2425	147	1	194
13	1	349	40	-1	365	67	-1	349	94	-1	1331	121	-1	345	148	-1	716
14	-1	192	41	1	392	68	1	162	95	-1	205	122	-1	1785	149	-1	1575
15	-1	1030	42	-1	452	69	-1	721	96	-1	765	123	-1	1546	150	1	192
16	-1	2795	43	-1	730	70	-1	192	97	-1	192	124	1	721	151	-1	1030
17	-1	192	44	1	1632	71	1	653	98	-1	451	125	-1	975	152	1	1030
18	-1	768	45	-1	596	72	-1	1632	99	-1	606	126	-1	721	153	1	3063
19	-1	349	46	-1	168	73	-1	1045	100	-1	892	127	1	4499	154	-1	707
20	-1	3034	47	-1	994	74	1	966	101	-1	526	128	-1	2186	155	-1	345
21	-1	347	48	-1	800	75	-1	1030	102	-1	192	129	1	623	156	-1	3102
22	-1	192	49	-1	722	76	-1	994	103	-1	392	130	-1	365	157	-1	392
23	-1	536	50	1	365	77	-1	168	104	-1	192	131	-1	349	158	-1	345
24	-1	194	51	-1	345	78	-1	1263	105	-1	1030	132	1	721	159	1	1404
25	-1	162	52	-1	850	79	1	365	106	1	192	133	-1	768	160	-1	1245
26	1	668	53	-1	782	80	-1	349	107	-1	365	134	-1	365	161	-1	721
27	-1	1727	54	1	365	81	1	314	108	-1	192	135	1	1331	162	-1	365

and γ via sensitivity analyses.

Here, we explore more intensively how the mislabeling effects and the performance of the proposed correction method may change with r with a given γ, or vice versa. In particular, given $\gamma = 3$, we consider different values of r, given by $r = 100 \times (1 + j)$; and given $r = 500$, we examine different values of γ, given by $\gamma = 10^{-2+\frac{j}{3}}$, where $j = 0, 1, \cdots, 9$.

In Figure S.1, we plot $T_X(50, r, 3)$ and $T_X^{correct}(50, r, 3)$ against r and $T_X(50, 500, \gamma)$ and $T_X^{correct}(50, 500, \gamma)$ against γ, with X representing A or R, where both the naive and correction methods are applied to the logistic regression and SVM classifiers.

S4 Simulation Studies

To further demonstrate the mismeasurement effects and the performance of the proposed correction method described in in Section 5 of the main text, here we conduct simulation studies, with one hundred simulations run for each configuration described below.

S4.1 Simulation Design

We let the set $\mathcal{P}$ of distributions be the class of bivariate normal distributions with mean $\mu = (\mu_1, \mu_2)^{\mathrm{T}}$ and covariance matrix Σ, where $\mu_1, \mu_2 \in \mathbb{R}$

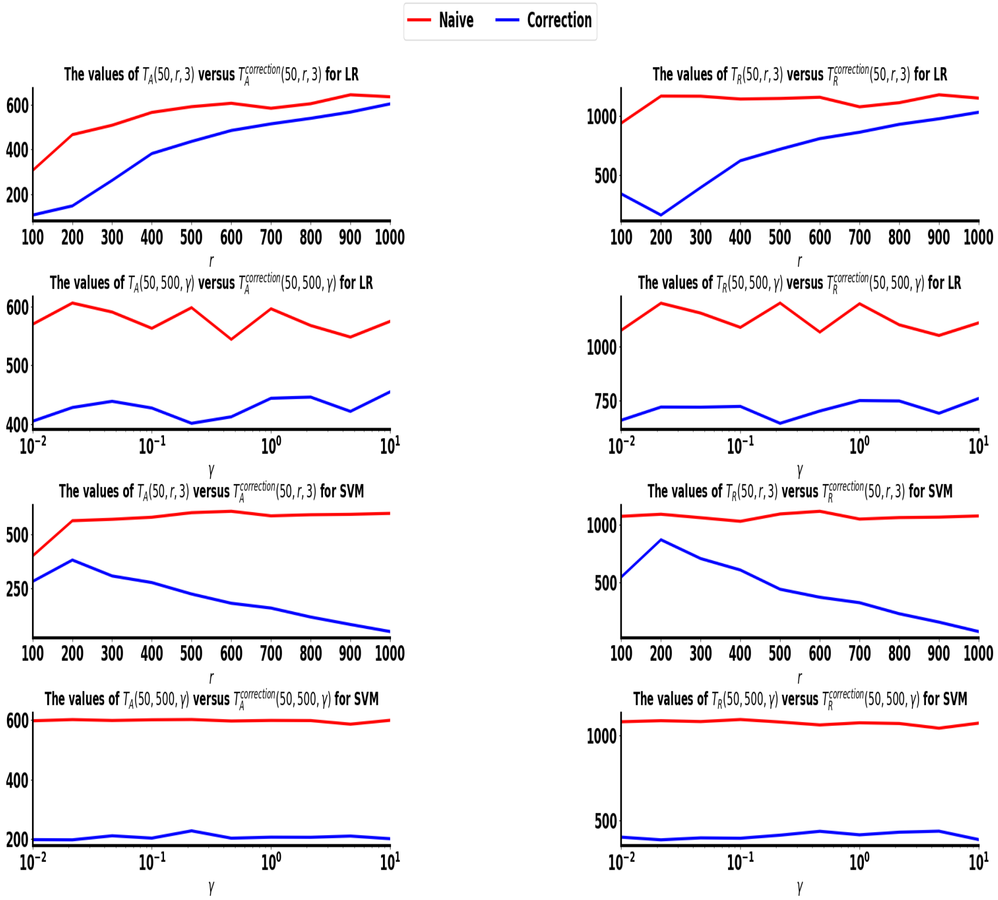


Figure S.1: *Plots of* $T_X(50, r, 3)$ *(in Red) and* $T_X^{correct}(50, r, 3)$ *(in Blue), against* r*, with* r *taking 10 equally spaced values in* $[100, 1000]$*; and plots of* $T_X(50, 500, \gamma)$ *(in Red) and* $T_X^{correct}(50, 500, \gamma)$ *(in Blue) against* γ*, with log* γ *taking 10 equally spaced values in* $[log\, 0.01, log\, 10]$*, where* X *represents* A *or* R*, and the results are obtained from the naive and correction methods with logistic regression and SVM.*

and Σ is a 2×2 positive definite matrix. To generate a sequence of n probability distributions $\{P_1, \cdots, P_n\}$ from $\mathcal{P}$, we independently generate μ_{i1} from the uniform distribution UNIF $[a_1, a_2]$ and μ_{i2} from the uniform distribution UNIF $[a_3, a_4]$ for $i = 1, \cdots, n$, where a_j are constants for $j = 1, \cdots, 4$; for $i = 1, \cdots, n$, we independently generate Σ_i from the Wishart distribution with mean Σ_0 and the degree of freedom n_w. Then we specify $P_i = \mathcal{N}(\mu_i, \Sigma_i)$ for $i = 1, \cdots, n$. For each i, to generate the label associated with the probability distribution P_i, we generate a sample l_i from the Bernoulli distribution, taking value 1 with probability $p = \frac{1}{2}\left\{\frac{\mu_{i1}-a_1}{a_2-a_1} + \frac{\mu_{i2}-a_3}{a_4-a_3}\right\}$ and -1 with the probability $1 - p$.

With (P_i, l_i) for $i = 1, 2, \cdots, n$, we independently generate m_i samples $\mathcal{S}_i = \left\{(X_{ij}, W_{ij})^{\mathrm{T}} \mid j = 1, \cdots, m_i\right\}$ from the probability distribution P_i for a given positive integer m_i.

With $(\mathcal{S}_i, l_i)$, for $i = 1, 2, \cdots, n$, we independently generate the mismeasured output l_i^* of l_i using the probabilities in (4.9) in the main text. We comment that here we use p_1^* and p_{-1}^* to facilitate the mismeasurement in outputs, whereas in sensitivity analyses in Section 6 in the main text, we use p_1 and p_{-1} to describe mismeasurement in output. While either probabilities can be used to characterize mismeasurement degrees, the convenience level is different, and the choice of a particular form is driven by individual contexts.

Here we set $n = 2000$, $m_i = 1000$ for each i, $a_1 = -200$, $a_2 = 100$, $a_3 = -100$, $a_4 = 300$, $n_w = 20$, and $\Sigma_0 = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix}$. We set different values for p_1^* and p_{-1}^* to reflect different magnitudes of mismeasurement in outputs, as detailed in the following sections.

S4.2 Simulation Results

With the generated precise dataset $\big\{(\mathcal{S}_i, l_i) \bigm| i = 1, 2, \cdots, n\big\}$ and mismeasured dataset $\big\{(\mathcal{S}_i, l_i^*) \bigm| i = 1, 2, \cdots, n\big\}$, we employ Steps 2 and 3 in Section 6.1 in the main text respectively using logistic regression and Gaussian kernel-based SVM to train the *true*, *naive*, and *correction* classifiers, denoted $\text{sign}(f_\theta)$, $\text{sign}(f_\theta^*)$, and $\text{sign}(f_\theta^{correct})$, respectively. As noted in Section 5 of the main text, the optimization problem (5.18) in the main text for *correction* classifiers may be nonconvex when (4.9) in the main text is taken. In this case, we employ Adam (Kingma and Bac2015), a widely used stochastic optimization algorithm in machine learning, to derive the *correction* classifiers. Then as described in Section 6.2 of the main text, we calculate the average values of $D_A(\theta)$, $D_R(\theta)$, $D_A^{correct}(\theta)$, and $D_R^{correct}(\theta)$ over one hundred simulations.

First, we examine the mislabeling effects under different scenarios with varying magnitudes of p_1^* and p_{-1}^* for label noise. To see how the average values of $D_A(\theta)$ and $D_R(\theta)$ over 100 hundred simulations vary as p_1^* and p_{-1}^* change, we first set $r = 500$, $\gamma = 3$ and consider four settings: $p_1^* = p_{-1}^* = 0.70, 0.80, 0.90$, or 0.99. We apply logistic regression and SVM respectively to the simulated data, with the results reported in Table S.2.

Table S.2: *Simulation studies obtained from logistic regression (LR) and SVM classifiers - assessing the impact of the mislabeling degrees: Average values of $D_A(\theta)$ and $D_R(\theta)$ for $p_1^* = p_{-1}^* = 0.7, 0.8, 0.9$, or 0.99 over 100 simulations.*

Mislabeling	$D_A(\theta)$		$D_R(\theta)$	
Degrees	LR	SVM	LR	SVM
$p_1^* = p_{-1}^* = 0.7$	0.16	0.28	0.12	0.27
$p_1^* = p_{-1}^* = 0.8$	0.10	0.18	0.11	0.18
$p_1^* = p_{-1}^* = 0.9$	0.06	0.11	0.06	0.11
$p_1^* = p_{-1}^* = 0.99$	0.01	0.03	0.03	0.03

More comprehensively, we consider more values of p_1^* and p_{-1}^* which are the cutpoints dividing $[0.5, 1]$ into N equal-length subintervals with $N = 20$ except $(p_1^*, p_{-1}^*) = (0.5, 0.5)$ or $(1, 1)$. We construct heatmaps for $D_A(p_1^*, p_{-1}^*, 500, 3)$ and $D_R(p_1^*, p_{-1}^*, 500, 3)$, and display them in the first two columns in Figure S.2. Figure S.2 conveys similar patterns shown by Figure 1 in the main text.

Next, we assess how the proposed correction method may perform under settings with different magnitudes of label noise. In particular, we produce heatmaps for $D_A^{correct}(p_1^*, p_{-1}^*, 500, 3)$ and $D_R^{correct}(p_1^*, p_{-1}^*, 500, 3)$ by applying the proposed method described in Section 5 of the main text to the LR and SVM classifiers, and display the results in the last two columns in Figure S.2. Comparing the last two columns to the first two columns in Figure S.2 shows that the proposed correction method generally outperforms the naive method for both classifiers.

Finally, similar to the consideration in Section 6.2 in the main text, we evaluate how the mismeasurement effects and the performance of the proposed correction method may be affected by the choice of r and γ. First, we consider $r = 100, 500$, or 1000 and $\gamma = 0.01, 0.1, 1, 3$, or 10, and report the average values of $T_X(20, r, \gamma)$ and $T_X^{correct}(20, r, \gamma)$ over 100 simulations obtained from the logistic regression and SVM classifiers in Tables S.3 and S.4, where "X" represents "A" or "R".

Furthermore, we assess the mislabeling effects and the performance of the proposed correction method by considering more refined values of r and γ. In particular, given $\gamma = 3$, we consider different values of r, given by $r = 100 \times (1 + j)$; and given $r = 500$, we examine different values of γ, given by $\gamma = 10^{-2+\frac{j}{3}}$, where $j = 0, 1, \cdots, 9$. We report in Figure S.3 the results of $T_X(20, r, 3)$ and $T_X^{correct}(20, r, 3)$ against r and $T_X(20, 500, \gamma)$ and

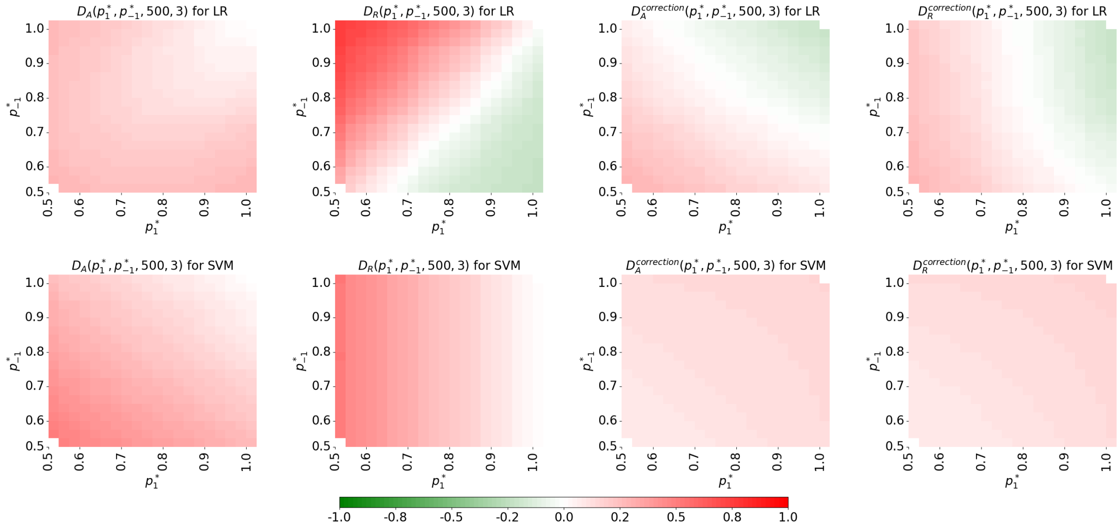
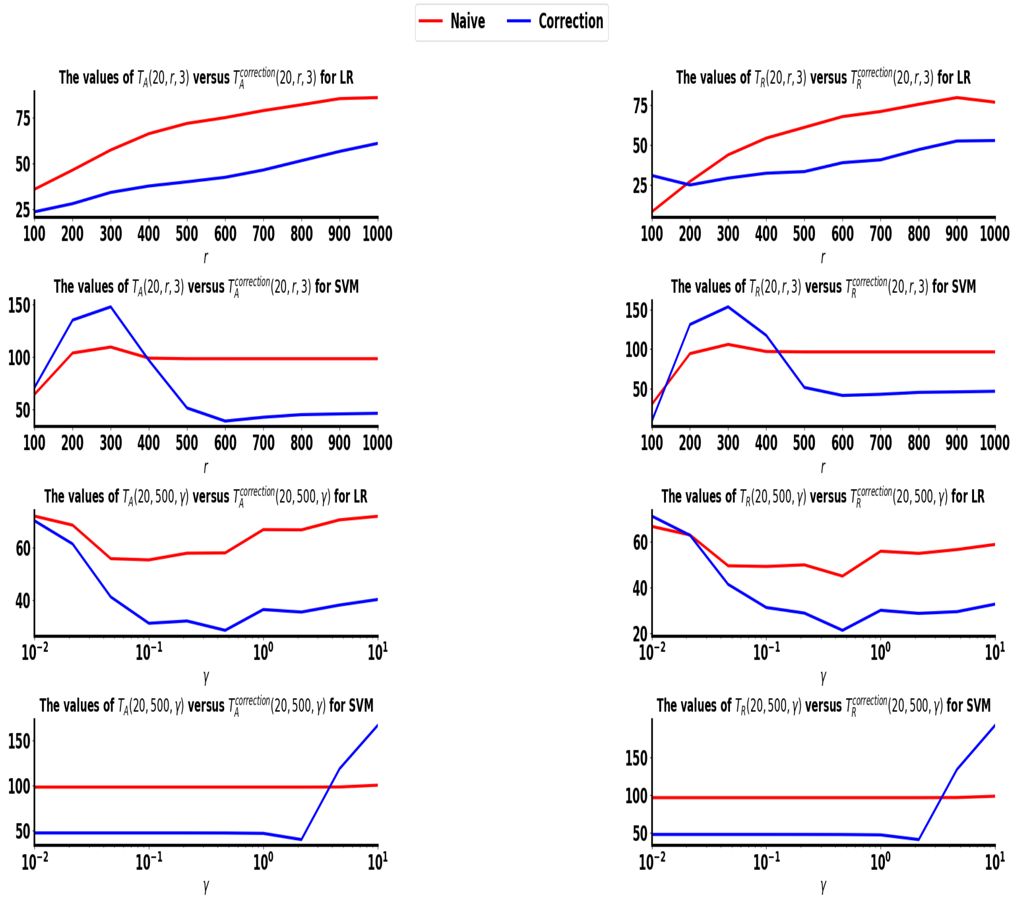


Figure S.2: *Heatmaps in simulation studies, generated from the naive method for* $D_A(p_1^*, p_{-1}^*, 500, 3)$ *and* $D_R(p_1^*, p_{-1}^*, 500, 3)$ *(displayed in the left two columns) and the proposed correction method for* $D_A^{correct}(p_1^*, p_{-1}^*, 500, 3)$ *and* $D_R^{correct}(p_1^*, p_{-1}^*, 500, 3)$ *(displayed in the last two columns). Two classifiers are considered, with the results for Logistic Regression (LR) and SVM displayed at top and bottom rows, respectively. Here,* $r = 500$*,* $\gamma = 3$*, and* p_1 *and* p_{-1} *take the values of the cutpoints dividing* $[0.5, 1]$ *into 20 equal lengthed subintervals, with* $(p_1, p_{-1}) = (0.5, 0.5)$ *and* $(1, 1)$ *excluded.*

Table S.3: *Simulation studies obtained from the naive and proposed methods with logistic regression (LR) and SVM classifiers - assessing the impact of different choices of* r *and* γ *on accuracy: The values of* $T_A(20, r, \gamma)$ *and* $T_A^{correct}(20, r, \gamma)$ *for* $r = 100, 500,$ *or* 1000 *and* $\gamma = 0.01, 0.1, 1, 3,$ *or* 10.

γ	$T_A(20, 100, \gamma)$		$T_A(20, 500, \gamma)$		$T_A(20, 1000, \gamma)$	
	LR	SVM	LR	SVM	LR	SVM
0.01	33.34	98.56	71.86	98.39	90.19	98.39
0.1	27.82	98.47	55.20	98.39	76.52	98.39
1	32.31	90.81	66.71	98.39	80.45	98.39
3	35.96	64.46	72.03	98.40	86.17	98.39
10	36.13	48.10	71.79	100.51	89.04	98.39
γ	$T_A^{correct}(20, 100, \gamma)$		$T_A^{correct}(20, 500, \gamma)$		$T_A^{correct}(20, 1000, \gamma)$	
	LR	SVM	LR	SVM	LR	SVM
0.01	30.21	47.99	57.76	47.68	93.20	47.66
0.1	18.67	47.81	10.88	47.67	61.60	47.66
1	23.79	102.98	15.55	47.22	56.03	47.64
3	23.63	70.96	22.54	51.54	61.10	46.49
10	23.60	51.59	19.75	166.72	64.07	37.26

$T_X^{correct}(20, 500, \gamma)$ against γ, with X representing A or R, where both the naive and correction methods are applied to the logistic regression and SVM classifiers. The results in Tables S.3 - S.4 and Figure S.3 reveal patterns similar to those displayed by Tables 1 in the main text and Figure S.1.



The values of $T_A(20, r, 3)$ versus $T_A^{correction}(20, r, 3)$ for LR

The values of $T_R(20, r, 3)$ versus $T_R^{correction}(20, r, 3)$ for LR

The values of $T_A(20, r, 3)$ versus $T_A^{correction}(20, r, 3)$ for SVM

The values of $T_R(20, r, 3)$ versus $T_R^{correction}(20, r, 3)$ for SVM

The values of $T_A(20, 500, \gamma)$ versus $T_A^{correction}(20, 500, \gamma)$ for LR

The values of $T_R(20, 500, \gamma)$ versus $T_R^{correction}(20, 500, \gamma)$ for LR

The values of $T_A(20, 500, \gamma)$ versus $T_A^{correction}(20, 500, \gamma)$ for SVM

The values of $T_R(20, 500, \gamma)$ versus $T_R^{correction}(20, 500, \gamma)$ for SVM

Figure S.3: *Plots of $T_X(20, r, 3)$ (in Red) and $T_X^{correct}(20, r, 3)$ (in Blue), against r, with r taking 10 equally spaced values in $[100, 1000]$; and plots of $T_X(20, 500, \gamma)$ (in Red) and $T_X^{correct}(20, 500, \gamma)$ (in Blue) against γ, with log γ taking 10 equally spaced values in $[log\, 0.01, log\, 10]$, where X represents A or R, and the results are obtained from applying the naive and correction methods to logistic regression and SVM classifiers.*

Table S.4: *Simulation studies obtained from the naive and proposed methods with logistic regression (LR) and SVM classifiers - assessing the impact of different choices of* r *and* γ *on recall: The values of* $T_R(20, r, \gamma)$ *and* $T_R^{correct}(20, r, \gamma)$ *for* $r = 100, 500,$ *or* 1000 *and* $\gamma = 0.01, 0.1, 1, 3,$ *or* 10.

γ	$T_R(20,100,\gamma)$		$T_R(20,500,\gamma)$		$T_R(20,1000,\gamma)$	
	LR	SVM	LR	SVM	LR	SVM
0.01	23.70	96.63	66.73	96.47	87.42	96.46
0.1	10.61	96.58	49.25	96.47	70.88	96.47
1	6.38	84.89	55.90	96.47	73.21	96.47
3	8.08	30.09	61.07	96.47	76.86	96.47
10	8.71	12.03	58.87	98.40	82.52	96.47
γ	$T_R^{correct}(20,100,\gamma)$		$T_R^{correct}(20,500,\gamma)$		$T_R^{correct}(20,1000,\gamma)$	
	LR	SVM	LR	SVM	LR	SVM
0.01	56.64	48.00	54.63	47.76	92.60	47.76
0.1	38.24	47.58	6.16	47.76	59.54	47.76
1	35.71	120.90	4.72	47.18	50.02	47.73
3	30.85	8.89	11.53	51.46	52.79	46.61
10	32.56	2.44	6.48	192.18	58.79	38.65

S4.3 Sensitivity Study of the Proposed Method

The validity of the proposed method relies on the knowledge of the misclassification probabilities. It is useful to assess how the proposed method may perform if the misclassification probabilities are misspecified. To this end, here we conduct simulation studies.

We generate the precise dataset $\left\{(\mathcal{S}_i, l_i) \mid i = 1, 2, \cdots, n\right\}$ and the mismeasured dataset $\left\{(\mathcal{S}_i, l_i^*) \mid i = 1, 2, \cdots, n\right\}$ by repeating the procedure in Section S4.1, where we set $p_1^* = s_1$ and $p_{-1}^* = s_{-1}$, with $s_1 = s_{-1} = 0.65, 0.7, 0.75,$ or 0.8, respectively called Setting 1, 2, 3, or, 4. Given the generated data, we employ Steps 2 and 3 in Section 6.1 in the main text using logistic regression or SVM classifier, where we set $r = 500$ and $\gamma = 3$. To implement the proposed correction method, we purposefully misspecify p_1^* and p_{-1}^* as $s_1 + a_1$ and $s_{-1} + a_{-1}$, respectively, where we consider $(a_1, a_{-1}) = (0, 0), (-0.05, -0.05), (-0.05, 0.05), (0.05, -0.05), (0.05, 0.05), (-0.1, -0.1),$ $(-0.1, 0.1), (0.1, -0.1), (0.1, 0.1)$, called Situations 1-9, respectively; Situation 1 represents the scenario without misspecification and other situations reflect different misspecification scenarios.

Tables S.5 and S.6 show the average values of $D_A^{correct}(\theta)$ and $D_R^{correct}(\theta)$ over 100 simulations in Situations 1-9 obtained from logistic regression and SVM, respectively. As expected, average values of $D_A^{correct}(\theta)$ and $D_R^{correct}(\theta)$ may be differently affected by varying degrees of misspecifying p_1^* and p_{-1}^*. However, the proposed correction method with the SVM classifier tends to be less sensitive than the corrected logistic regression, showing better robustness to misspecification of p_1^* and p_{-1}^* than logistic regression.

Table S.5: *Simulation results with misspecified* (p_1^*, p_{-1}^*)*: Average values of* $D_A^{correct}(\theta)$ *(DA) and* $D_R^{correct}(\theta)$ *(DR) obtained from the proposed method applied to logistic regression under four settings of* $s_1 = s_{-1}$ *and nine situations of* (a_1, a_{-1})*.*

Situation	1		2		3		4		5		6		7		8		9	
Setting	DA	DR	DA	DR	DA	DR	DA	DR	DA	DR	DA	DR	DA	DR	DA	DR	DA	DR
1	0.19	0.20	0.20	0.20	0.19	0.15	0.19	0.25	0.18	0.19	0.20	0.21	0.19	0.13	0.19	0.29	0.17	0.19
2	0.14	0.15	0.15	0.16	0.14	0.11	0.15	0.20	0.13	0.14	0.16	0.17	0.15	0.08	0.15	0.25	0.12	0.13
3	0.09	0.11	0.10	0.12	0.09	0.06	0.10	0.16	0.08	0.10	0.12	0.13	0.10	0.03	0.11	0.21	0.07	0.09
4	0.04	0.07	0.06	0.08	0.04	0.02	0.06	0.13	0.03	0.06	0.07	0.10	0.05	0.00	0.07	0.19	0.03	0.05

Bibliography

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

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006). Convexity, clas-

Table S.6: *Simulation results with misspecified* (p_1^*, p_{-1}^*)*: Average values of* $D_A^{correct}(\theta)$ *(DA) and* $D_R^{correct}(\theta)$ *(DR) obtained from the proposed method applied to SVM under four settings of* $s_1 = s_{-1}$ *and nine situations of* (a_1, a_{-1})*.*

Situation / Setting	1		2		3		4		5	
	DA	DR	DA	DR	DA	DR	DA	DR	DA	DR
1	0.11	0.09	0.10	0.08	0.11	0.09	0.11	0.09	0.12	0.10
2	0.12	0.10	0.11	0.09	0.12	0.10	0.12	0.10	0.13	0.11
3	0.13	0.11	0.12	0.10	0.13	0.11	0.13	0.11	0.13	0.12
4	0.13	0.12	0.13	0.11	0.13	0.12	0.13	0.12	0.14	0.13

Situation / Setting	6		7		8		9	
	DA	DR	DA	DR	DA	DR	DA	DR
1	0.10	0.08	0.11	0.09	0.11	0.09	0.13	0.11
2	0.10	0.08	0.12	0.10	0.12	0.10	0.13	0.12
3	0.11	0.09	0.13	0.11	0.13	0.11	0.14	0.13
4	0.12	0.10	0.13	0.12	0.13	0.12	0.15	0.13

sification, and risk bounds. *Journal of the American Statistical Association 101*(473), 138–156.

Brézis, H. (2011). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York: Springer.

Diestel, J. and J. J. Uhl (1977). *Vector Measures*. Providence: American Mathematical Society.

Geiss, C. and S. Geiss (2004). *An Introduction to Probability Theory*. Lecture Notes.

Hoffmann, H. (2015). On the continuity of the inverses of strictly monotonic functions. *Irish Mathematical Society Bulletin 75*(1), 45–57.

Kelley, J. L. (2017). *General Topology*. New York: Courier Dover Publications.

Kingma, D. P. and J. Ba (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

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.

Rudin, W. (1962). *Fourier Analysis on Groups.* New York: Wiley.

Song, L. (2008). Learning via Hilbert space embedding of distributions. *PhD thesis, The University of Sydney*.

Thomas, G. B., M. D. Weir, J. Hass, and C. Heil (2014). *Thomas' Calculus* (Thirteen ed.). Pearson.