Learning from Noisy Labels via Conditional Distributionally Robust Optimization

Hui Guo
Department of Computer Science
University of Western Ontario
hguo288@uwo.ca

Grace Y. Yi [*]
Department of Statistical and Actuarial Sciences
Department of Computer Science
University of Western Ontario
gyi5@uwo.ca

Boyu Wang
Department of Computer Science
University of Western Ontario
bwang@csd.uwo.ca

Abstract

While crowdsourcing has emerged as a practical solution for labeling large datasets, it presents a significant challenge in learning accurate models due to noisy labels from annotators with varying levels of expertise. Existing methods typically estimate the true label posterior, conditioned on the instance and noisy annotations, to infer true labels or adjust loss functions. These estimates, however, often overlook potential misspecification in the true label posterior, which can degrade model performances, especially in high-noise scenarios. To address this issue, we investigate learning from noisy annotations with an estimated true label posterior through the framework of *conditional distributionally robust optimization* (CDRO). We propose formulating the problem as minimizing the worst-case risk within a distance-based ambiguity set centered around a reference distribution. By examining the strong duality of the formulation, we derive upper bounds for the worst-case risk and develop an analytical solution for the dual robust risk for each data point. This leads to a novel robust pseudo-labeling algorithm that leverages the likelihood ratio test to construct a pseudo-empirical distribution, providing a robust reference probability distribution in CDRO. Moreover, to devise an efficient algorithm for CDRO, we derive a closed-form expression for the empirical robust risk and the optimal Lagrange multiplier of the dual problem, facilitating a principled balance between robustness and model fitting. Our experimental results on both synthetic and real-world datasets demonstrate the superiority of our method.

1 Introduction

Recent advancements in supervised learning have spurred a growing demand for large labeled datasets [1, 2]. However, acquiring accurately annotated datasets is typically costly and time-consuming, often requiring a pool of annotators with adequate domain expertise to manually label the data. Crowdsourcing has emerged as an efficient and cost-effective solution for annotating large datasets. On crowdsourcing platforms, multiple annotators with varying levels of labeling skills are employed to gather extensive labeled data. However, this approach introduces a significant challenge: the labels collected through crowdsourcing are often subject to unavoidable noise, especially in fields

[*] Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2627

requiring substantial domain knowledge, such as medical imaging. Consequently, models trained on noisy labels are prone to error, including overfitting, since deep models can memorize vast amounts of data [3]. In addition to statistical research on label noise (often termed response measurement error, e.g., [4–6]), a growing body of recent machine learning literature has focused on developing effective algorithms capable of training accurate classifiers using noisy data, e.g., [7–10]. Many of these methods seek to approximate the *posterior distribution of the underlying true labels* using the observed data.

Let $\mathbf{X}$ be an instance, Y denote the unobserved true label for $\mathbf{X}$, and $\widetilde{\mathbf{Y}}$ represent a vector of crowdsourced noisy labels for $\mathbf{X}$. The data-generating distribution, $P^*_{\mathbf{x},\mathrm{y},\widetilde{\mathbf{y}}}$, can be factorized in two ways: $P^*_{\mathbf{x}}P^*_{\mathrm{y}|\mathbf{x}}P^*_{\widetilde{\mathbf{y}}|\mathbf{x},\mathrm{y}}$ or $P^*_{\mathbf{x},\widetilde{\mathbf{y}}}P^*_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$, with P^* denoting the (conditional) distribution for the variables indicated by the corresponding subscripts. These factorizations have inspired research that trains models by estimating the posterior distribution of the true labels, $P^*_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$, in the latter factorization.

Previous work [9–12] introduced various algorithms for estimating the *annotator confusions*, also known as *noise transition probabilities*, which yield an approximated conditional distribution of $\widetilde{\mathbf{Y}}$, given Y and $\mathbf{X}$, denoted $P_{\widetilde{\mathbf{y}}|\mathrm{y},\mathbf{x}}$. For ease of reference, we use P^* and P to denote the true distribution and an approximate distribution for the variables indicated by the corresponding subscripts. Given the observed data $\mathbf{X}$ and $\widetilde{\mathbf{Y}}$, along with an approximated conditional distribution $P_{\widetilde{\mathbf{y}}|\mathrm{y},\mathbf{x}}$ and a prior for Y given $\mathbf{X}$, denoted $P_{\mathrm{y}|\mathbf{x}}$, the true label posterior is then computed as $P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}} \propto P_{\mathrm{y}|\mathbf{x}} \cdot P_{\widetilde{\mathbf{y}}|\mathrm{y},\mathbf{x}}$ by Bayes's theorem [10, 13]. This estimated true label posterior is often used to infer the underlying true labels or to weight the loss functions [7, 9, 10, 14].

However, accurately computing the posterior of the true label is challenging, and the estimated posterior $P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$ may deviate from the underlying true distribution $P^*_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$ due to potential misspecifications in the prior belief and the conditional noise transition probabilities [15]. To address this issue, we introduce a robust scheme for handling crowdsourced noisy labels through conditional distributionally robust optimization (CDRO), as discussed in [16]. Specifically, we frame the problem as minimizing the worst-case risk within a distance-based *ambiguity set*, which constrains the degree of conditional distributional uncertainty around a *reference distribution*. By leveraging the strong duality in linear programming, we derive the dual form of the robust risk and establish informative upper bounds for the worst-case risk. Additionally, for each data point, we develop an analytical solution to the robust risk minimization problem, which encompasses existing approaches as special cases [9]. This solution is presented in a likelihood ratio format and inspires a robust approach that assigns pseudo-labels only to instances with high confidence, with uncertain data filtered out. These pseudo-labels also enable us to construct a pseudo-empirical distribution that serves as a robust reference probability distribution in CDRO under potential model misspecifications. Moreover, we derive a closed-form expression for the empirical robust risk by identifying the optimal Lagrange multiplier in the dual form. Building on this, we ultimately develop an algorithm for learning from noisy labels via conditional distributionally robust true label posterior with an adaptive Lagrange multiplier (AdaptCDRP).

Our contributions are summarized as follows: (1) We formulate learning with noisy labels as a CDRO problem and develop its dual form to tackle the challenge of potential misspecification in estimating the true label posterior from noisy data. (2) We derive an analytical solution to the dual problem for each data point, and propose a novel algorithm that constructs a robust reference distribution for this problem. (3) By deriving the optimal Lagrange multiplier for the empirical robust risk, we develop an efficient one-step update method for the Lagrange multiplier, allowing for a principled balance between robustness and model fitting. Code is available at https://github.com/hguo1728/AdaptCDRP.

Notations. We use $[k]$ to denote $\{1, \ldots, k\}$ for any positive integer k, and $\mathbf{1}(\cdot)$ to denote the indicator function. For a vector $\boldsymbol{v}$, v_j stands for its jth element, and $\boldsymbol{v}^\top$ denotes its transpose. For $\boldsymbol{v} = (v_1, ..., v_p)^\top$ and $q \in [1, +\infty]$, the L^q norm is defined as $\|\boldsymbol{v}\|_q = (\sum_{j=1}^p |v_j|^q)^{1/q}$ if $1 \leq q < \infty$, and $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$ if $q = +\infty$. For a matrix $\boldsymbol{V}$, we use $V_{i,j}$ to represent its (i, j) element. Furthermore, let $(\Omega, \mathcal{G}, \mu)$ denote the measure space under consideration, where Ω is a set, $\mathcal{G}$ is the σ-field of subsets of Ω, and μ is the associated measure. For $q > 0$, let $L^q(\mu)$ represent the collection of Borel-measurable functions $f : \Omega \to \mathbb{R}$ such that $\int |f|^q d\mu < \infty$. Let $\mathsf{d}(\cdot, \cdot)$ denote a metric on Ω. We call f L-Lipschitz with respect to $\mathsf{d}(\cdot, \cdot)$ if $|f(u_1) - f(u_2)| \leq L \cdot \mathsf{d}(u_1, u_2)$ for all $u_1, u_2 \in \Omega$, where L is a positive constant.

https://doi.org/10.52202/079017-2627

2 Proposed Framework

2.1 Problem Formulation

Consider a classification task with feature space $\mathcal{X} \subset \mathbb{R}^d$ and label space $\mathcal{Y}$, where d is the feature dimension. Here $\mathcal{Y}$ is taken as $\{0, 1\}$ for binary classification and $[K]$ for multi-class classification with $K > 2$. Let $\mathbf{X} \in \mathcal{X}$ denote an instance and $\mathrm{Y} \in \mathcal{Y}$ denote its true label. Let Ψ denote the considered hypothesis class consisting of functions ψ defined over $\mathcal{X}$, which, for example, can be neural networks that output predicted label probabilities for each $\mathbf{x} \in \mathcal{X}$. Specifically, for binary classification, $\psi : \mathcal{X} \to [0, 1]$, with $\psi(\mathbf{x})$ representing $P(\mathrm{Y} = 1|\mathbf{X} = \mathbf{x})$, and the classified value is given by $\mathbf{1}(\psi(\mathbf{x}) > 0.5)$. For multi-class classification, $\psi : \mathcal{X} \to \Delta^{K-1}$ with $K > 2$ and Δ^{K-1} denoting the K-simplex, where the jth component of $\psi(\mathbf{x})$, denoted $\psi(\mathbf{x})_j$, represents the conditional probability $P(\mathrm{Y} = j|\mathbf{X} = \mathbf{x})$ for $j \in [K]$, with the classified value defined as $\arg\max_{j\in[K]} \psi(\mathbf{x})_j$.

In applications, the true label Y is often unobserved, and instead, a set of crowdsourced noisy labels $\widetilde{\mathbf{Y}} \triangleq \{\widetilde{\mathrm{Y}}^{(r)}\}_{r=1}^{R}$ is collected, where $\widetilde{\mathrm{Y}}^{(r)} \in \mathcal{Y}$, denoting the label provided by annotator r out of R annotators. Let $\mathcal{D} \triangleq \{\mathbf{X}_i, \widetilde{\mathbf{Y}}_i\}_{i=1}^{n}$ denote the observed data of size n, where $\widetilde{\mathbf{Y}}_i$ contains noisy labels provided by R annotators for instance $\mathbf{X}_i$, which may differ from the true label Y_i for each $i \in [n]$. Our goal is to train a classifier using $\mathcal{D}$ to accurately predict the true label for a future instance.

A common assumption in supervised learning is that the data points $\{\mathbf{X}_i, \mathrm{Y}_i, \widetilde{\mathbf{Y}}_i\}$ for $i \in [n]$ are independently drawn from a probability measure $P^*_{\mathbf{x},\mathrm{y},\widetilde{\mathbf{y}}}$ for $\{\mathbf{X}, \mathrm{Y}, \widetilde{\mathbf{Y}}\}$, defined over the space $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}^R$. Under this assumption, many existing methods aim to approximate the posterior distribution of the underlying true label Y, given the observed data $\mathbf{X}$ and $\widetilde{\mathbf{Y}}$ [7, 9–11]. The estimated true label posterior, denoted $P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$, is then applied to either infer the true labels or to weight the loss functions. For example, [9] utilized $P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$ as a weight in the loss functions, without considering potential misspecification of the associated model. However, such strategies typically ignore the variability induced in estimating the true label posterior.

To mitigate the effects of potential misspecifications, we propose a conditional distributionally robust risk optimization problem:

$$\inf_{\psi\in\Psi} \mathsf{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}), \text{ with } \mathsf{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \triangleq \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\Big[\sup_{Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}\in\Gamma_\epsilon(P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})} \mathbb{E}_{Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}\{\ell(\psi(\mathbf{X}), \mathrm{Y})\}\Big], \tag{1}$$

where $\ell(\cdot, \cdot)$ is a loss function, the expectation $\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}$ is taken with respect to the joint distribution of the observed data $\mathbf{X}$ and $\widetilde{\mathbf{Y}}$, and the expectation $\mathbb{E}_{Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}$ is evaluated under the conditional distribution model, denoted $Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$, of the true label Y, given $\mathbf{X}$ and $\widetilde{\mathbf{Y}}$. Here, $\Gamma_\epsilon(P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$ is an *ambiguity set* of probability measures centered around the *reference probability distribution* $P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$, indexed by $\epsilon > 0$ [16–18]. For instance, $\Gamma_\epsilon(P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$ can be conceptualized as a "ball" with $P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$ at its center and ϵ as the radius, where elements in the ball represent possible distribution models for $P^*_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$, and the distance between two points is measured using a standard metric for distributions. Specifically,

$$\Gamma_\epsilon(P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) = \left\{Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}} \in \mathcal{P}(\mathcal{Y}) : \mathcal{d}(Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}, P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \le \epsilon\right\}, \tag{2}$$

where $\mathcal{P}(\mathcal{Y})$ denotes all Borel probability measures on $\mathcal{Y}$, and $\mathcal{d}$ is a discrepancy metric of probability measures. In this paper, we employ the Wasserstein distance in Definition 2.1 to define the ambiguity set. By taking the supremum in (1) over the ambiguity set (2), we aim to minimize the worst-case risk around the reference distribution, thereby mitigating the impact of potential model misspecifications.

One main obstacle in solving (1) is constructing a reliable reference distribution $P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}$, which typically depends on an empirical distribution that requires true labels in conventional distributionally robust optimization (DRO). We address this issue by investigating the dual form of the robust risk presented in (1), which enables us to create a robust pseudo-empirical distribution using a likelihood ratio test, as detailed in Section 3.1. An additional advantage of our approach is that it provides informative upper bounds for the worst-case risk in Section 2.3 via the dual formulation.

Remark 2.1. For simplicity in theoretical presentation, we assume access to all annotations from all R annotators. However, the theoretical framework presented in this paper is applicable to both single-annotator ($R = 1$) and multiple-annotator ($R > 1$) scenarios. In our experiments in Section 4, we also consider the scenario of sparse labeling, where we generate a total of R annotators and then randomly select one annotation per instance from these R annotators. We also conduct experiments with varying numbers of annotators for a comprehensive analysis.

2.2 Duality Result and Relaxed Problem

To derive the pseudo-label generation algorithm and establish a reliable reference distribution, we analyze the dual form of (1). We first define the Wasserstein distance of order p for $p \in [1, +\infty)$.

Definition 2.1 (p-Wasserstein distance, [17])**.** For a Polish space $\mathcal{S}$ (i.e., a complete separable metric space) endowed with a metric $c : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$, also called a cost function, let $\mathcal{P}(\mathcal{S})$ represent the set of all Borel probability measures on $\mathcal{S}$, where $\mathbb{R}_{\geq 0}$ represents the set of all nonnegative real values. For $p \geq 1$, let $\mathcal{P}_p(\mathcal{S})$ stand for the subset of $\mathcal{P}(\mathcal{S})$ with finite pth moments. Then, for $P_1, P_2 \in \mathcal{P}_p(\mathcal{S})$, the Wasserstein distance of order p is defined as

$$W_p(P_1, P_2) \triangleq \inf_{\Pi \in \mathrm{Cpl}(P_1, P_2)} \left[\mathbb{E}_{(S_1, S_2) \sim \Pi} \left\{ c^p(S_1, S_2) \right\} \right]^{1/p},$$

where $\mathrm{Cpl}(P_1, P_2)$ comprises all probability measures on the product space $\mathcal{S} \times \mathcal{S}$ such that their marginal measures are $P_1(\cdot)$ and $P_2(\cdot)$. Here, $c^p(\cdot, \cdot)$ represents $\{c(\cdot, \cdot)\}^p$.

In (2), we set $\mathscr{d}(\cdot, \cdot)$ as the p-Wasserstein distance and incorporate the constraint $\mathscr{d}(Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}, P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \leq \epsilon$ using the Lagrange formulation, and then establish the strong duality result for (1) as follows.

Proposition 2.1 (dual problem)**.** *Assume that for every given* $\mathbf{x} \in \mathcal{X}$, $\widetilde{\mathbf{y}} \in \mathcal{Y}^R$ *and* $\psi \in \Psi$, $\ell(\psi(\mathbf{x}), \cdot) \in L^1(P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$*, where* $L^1(\cdot)$ *is defined in Section 1. Consider* $\mathscr{d}(\cdot, \cdot)$ *in (2) as the Wasserstein distance of order* p*. Then, for any* $\epsilon > 0$, $\mathsf{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$ *in (1) becomes:*

$$\mathsf{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) = \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}} \left\{ \inf_{\gamma \geq 0} \left(\gamma \epsilon^p + \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}} \left[\sup_{y' \in \mathcal{Y}} \left\{ \ell(\psi(\mathbf{X}), y') - \gamma c^p(y', \mathrm{Y}) \right\} \right] \right) \right\}. \tag{3}$$

To avoid solving nested optimization problems, we consider an alternative formulation by swapping the infimum and the first expectation operations:

$$\mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \triangleq \inf_{\gamma \geq 0} \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}} \left(\gamma \epsilon^p + \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}} \left[\sup_{y' \in \mathcal{Y}} \left\{ \ell(\psi(\mathbf{X}), y') - \gamma c^p(y', \mathrm{Y}) \right\} \right] \right), \tag{4}$$

which is an upper bound of $\mathsf{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$ according to Proposition 2.1, and hence, (4) can be regarded as an relaxation of (3). The empirical counterpart of $\mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$ is given by

$$\widehat{\mathfrak{R}}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \triangleq \inf_{\gamma \geq 0} \mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}} \left(\gamma \epsilon^p + \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}} \left[\sup_{y' \in \mathcal{Y}} \left\{ \ell(\psi(\mathbf{X}), y') - \gamma c^p(y', \mathrm{Y}) \right\} \right] \right), \tag{5}$$

where $P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}} \triangleq \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{X}_i, \widetilde{\mathbf{Y}}_i}$ is the empirical distribution of $(\mathbf{X}, \widetilde{\mathbf{Y}})$ based on the dataset $\mathcal{D}$ defined in Section 2.1. Here, for any $\mathbf{v} \in \mathcal{X} \times \mathcal{Y}^R$, $\delta_{\mathbf{v}}$ represents the Dirac measure on $\mathcal{X} \times \mathcal{Y}^R$, defined as $\delta_{\mathbf{v}}(A) \triangleq \mathbf{1}\{\mathbf{v} \in A\}$ for any $A \subset \mathcal{X} \times \mathcal{Y}^R$.

Remark 2.2. The Lagrange multiplier γ in (4) and (5) captures the trade-off between robustness and model fitting in the presence of label noise and potential model misspecifications. When the solution in γ is large, the inner supremum tends to favor $\mathrm{y}' = \mathrm{Y}$, thus encouraging the minimization of the natural risk using the reference distribution directly. In contrast, a small solution in γ introduces perturbations to the data, pushing the classifier away from the sample instances weighted by the reference distribution.

Remark 2.3. When $p = 1$, (5) represents the dual form of the following problem:

$$\sup_{Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}} \in \overline{\Gamma}_\epsilon(P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})} \mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}} \left[\mathbb{E}_{Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}} \left\{ \ell(\psi(\mathbf{X}), \mathrm{Y}) \right\} \right], \tag{6}$$

where $\overline{\Gamma}_\epsilon(P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) = \left\{ Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}} \in \mathcal{P}(\mathcal{Y}) : \mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}} \mathscr{d}(Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}, P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \leq \epsilon \right\}$. The proof of this statement is deferred to Appendix A.3. This result indicates that the empirical robust risk in the relaxed problem (5) corresponds to the worst-case risk within an ambiguity set that constrains the size of the average conditional distributional uncertainty.

2.3 Generalization Bounds

With the duality result in Proposition 2.1 and the derivations of the alternative formulations (4) and (5), we now characterize the difference between $\widehat{\mathfrak{R}}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$ and its population counterpart $\mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$.

Theorem 2.2. *Consider the loss function $\ell(\cdot,\cdot)$ in Proposition 2.1, and let the cost function $c(y,y') = \kappa\mathbf{1}(y \neq y')$ for $y, y' \in \mathcal{Y}$, where κ is a positive constant. Assume that there exists a positive constant M such that $\ell(\psi(\mathbf{x}), \mathrm{y}) \in [0, M]$ for all $\mathbf{x} \in \mathcal{X}$, $\mathrm{y} \in \mathcal{Y}$, and $\psi \in \Psi$, and that $\ell(\psi(\mathbf{x}), \mathrm{y})$ is L-Lipschitz in the second argument with respect to the cost function $c(\cdot,\cdot)$. Then, there exists a positive constant C_1 such that for any given $\epsilon > 0$, $\psi \in \Psi$, and $0 < \eta < 1$, with probability at least $1-\eta$:*

$$\left|\mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \widehat{\mathfrak{R}}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})\right| \leq \frac{C_1 L \kappa^p}{\epsilon^{p-1}\sqrt{n}} + M\sqrt{\frac{\log(1/\eta)}{2n}}.$$

Theorem 2.2 suggests that the empirical counterpart, $\widehat{\mathfrak{R}}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$, is a useful approximation for the risk function $\mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$, as their disparity is bounded and cannot grow indefinitely large. For a finite sample size n, this disparity is upper bounded by a finite value depending on the characteristics of the cost and loss functions, as reflected by κ, L, and M. As the sample size $n \to \infty$, the difference tends to zero with high probability, and specifically, the difference is of order $O(n^{-1/2})$.

Next, we establish an informative bound for the empirical robust risk minimizer. For $\psi_1, \psi_2 \in \Psi$ and for any given norm $\|\cdot\|$, let $\|\psi_1 - \psi_2\|_\infty \triangleq \sup_{\mathbf{x}\in\mathcal{X}} \|\psi_1(\mathbf{x}) - \psi_2(\mathbf{x})\|$. Here, $\|\cdot\|$ can be taken as any specific norms, including the L^q norm with $q \geq 1$ that is defined in Section 1.

Corollary 2.3 (Empirical Robust Minimizer). *Let $\widehat{\psi}_{\epsilon,n} \in \inf_{\psi\in\Psi} \widehat{\mathfrak{R}}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$. Under the assumptions in Theorem 2.2, if we further assume that the loss function $\ell(\cdot,\cdot)$ is L'-Lipschitz in terms of the first argument with respect to the supremum metric $\|\cdot\|_\infty$, then there exists a positive constant C_2 such that for any $\epsilon > 0$ and $0 < \eta < 1$, with probability at least $1-\eta$, the empirical robust risk minimizer $\widehat{\psi}_{\epsilon,n}$ satisfies:*

$$\begin{aligned}
&\mathsf{R}_\epsilon(\widehat{\psi}_{\epsilon,n}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \leq \mathfrak{R}_\epsilon(\widehat{\psi}_{\epsilon,n}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \\
&\leq \inf_{\psi\in\Psi} \mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) + C_2\left\{\frac{L\kappa^p}{\epsilon^{p-1}} + L'\int_0^\infty \sqrt{\log N(s; \Psi, \|\cdot\|_\infty)}ds\right\} \cdot \frac{1}{\sqrt{n}} + 2M\sqrt{\frac{\log(1/\eta)}{2n}},
\end{aligned}$$

where $N(s; \Psi, \|\cdot\|_\infty)$ denotes the s-covering number of Ψ with respect to the supremum metric.

3 Implementation Algorithm

In Section 3.1, we derive the analytical solution to the dual robust risk minimization problem in (5), which leads to the development of a novel approach for assigning pseudo-labels using the likelihood ratio test. These pseudo-labels facilitate the construction of a *pseudo-empirical distribution*, serving as a robust reference distribution in using (5). In Section 3.2, we derive the optimal value in γ for the empirical robust risk (5) and establish its closed-form expression. This analysis provides a principled framework for balancing the trade-off between robustness and model fitting and also motivates an efficient one-step update technique in solving the robust empirical risk minimization problem.

3.1 Optimal Solution for Single Data Point

In this subsection, we determine the optimal value of $\psi(\mathbf{x})$ in (5) for a single data point $(\mathbf{x}, \widetilde{\mathbf{y}})$. To simplify the analysis, we first focus on the binary classification problem with $\mathcal{Y} = \{0, 1\}$ and consider a broad family of loss functions of the form:

$$\ell(\psi(\mathbf{x}), \mathrm{y}) = (1-\mathrm{y})\mathcal{T}(1-\psi(\mathbf{x})) + \mathrm{y}\mathcal{T}(\psi(\mathbf{x})), \tag{7}$$

where $\psi(\mathbf{x})$ represents the conditional distribution $P(\mathrm{Y} = 1|\mathbf{X} = \mathbf{x})$ as described in Section 2.1, $\mathcal{T} : [0,1] \to \mathcal{I}$ is a bounded, decreasing, and twice differentiable function, and $\mathcal{I}$ is a compact subset of $\mathbb{R}$.

For any given $\mathbf{x} \in \mathcal{X}$ and $\widetilde{\mathbf{y}} \in \mathcal{Y}^R$, let $P_j(\mathbf{x}, \widetilde{\mathbf{y}}) \triangleq P(\mathrm{Y} = j|\mathbf{X} = \mathbf{x}, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}})$ for $j = 0, 1$. With the loss function in (7) and the metric $c(\cdot,\cdot)$ considered in Theorem 2.2, minimizing (5) with respect to $\psi \in \Psi$ becomes:

$$\begin{aligned}
\inf_{\psi\in\Psi}\inf_{\gamma\geq 0} \big[&\gamma\epsilon^p + P_0(\mathbf{x}, \widetilde{\mathbf{y}}) \max\{\mathcal{T}(1-\psi(\mathbf{x})), \mathcal{T}(\psi(\mathbf{x})) - \gamma\kappa^p\} \\
&+ P_1(\mathbf{x}, \widetilde{\mathbf{y}}) \max\{\mathcal{T}(1-\psi(\mathbf{x})) - \gamma\kappa^p, \mathcal{T}(\psi(\mathbf{x}))\}\big],
\end{aligned} \tag{8}$$

https://doi.org/10.52202/079017-2627

and let $\psi^\star$ denote the solution of (8). For ϵ and κ described in Theorem 2.2, let $\varrho(\epsilon) \triangleq \epsilon^p/\kappa^p$. The following theorem shows that (8) has a closed-form solution, with its form varying based on whether $\mathcal{T}$ is concave or convex.

Theorem 3.1 (Optimal Action for Single Data Point: Binary Case)**.** *Let $\mathbf{x} \in \mathcal{X}$ and $\widetilde{\mathbf{y}} \in \mathcal{Y}^R$ be given. Then, for a concave function $\mathcal{T}$, the optimal solution for (8) is given by:*

$$\psi^\star(\mathbf{x}) = \begin{cases} j, \ \textit{if}\ P_j(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \varrho(\epsilon) + \varpi_1 \ \textit{for}\ j = 0, 1; \\ 1/2, \ \textit{otherwise}, \end{cases}$$

with $\varpi_1 = \{\mathcal{T}(0) - \mathcal{T}(1/2)\}/\{\mathcal{T}(0) - \mathcal{T}(1)\} \in (0, 1/2]$; and for a convex function $\mathcal{T}$, the optimal solution of (8) is given by:

$$\psi^\star(\mathbf{x}) = \begin{cases} j, \ \textit{if}\ P_j(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \varrho(\epsilon) + \varpi_2 \,\textit{for}\ j = 0, 1, \\ t_j^*, \ \textit{if}\ \varrho(\epsilon) + 1/2 < P_j(\mathbf{x}, \widetilde{\mathbf{y}}) < \varrho(\epsilon) + \varpi_2 \,\textit{for}\ j = 0, 1, \\ 1/2, \ \textit{otherwise}, \end{cases}$$

where $\varpi_2 = \{\mathcal{T}'(0)\}/\{\mathcal{T}'(0) + \mathcal{T}'(1)\} \in [1/2, 1)$, t_0^* is the unique solution of $\{P_0(\mathbf{x}, \widetilde{\mathbf{y}}) - \varrho(\epsilon)\}\mathcal{T}'(1-t) = \{P_1(\mathbf{x}, \widetilde{\mathbf{y}}) + \varrho(\epsilon)\}\mathcal{T}'(t)$ for $t \in (0, \frac{1}{2})$, and t_1^* is the unique solution of $\{P_0(\mathbf{x}, \widetilde{\mathbf{y}}) + \varrho(\epsilon)\}\mathcal{T}'(1-t) = \{P_1(\mathbf{x}, \widetilde{\mathbf{y}}) - \varrho(\epsilon)\}\mathcal{T}'(t)$ for $t \in (\frac{1}{2}, 1)$.

Remark 3.1. The optimal solution $\psi^\star(\mathbf{x})$ in Theorem 3.1 can also be expressed in a likelihood ratio format, which naturally leads to a novel algorithm for assigning *robust pseudo-labels*. Specifically, when $\mathcal{T}$ is concave, the optimal solution can be expressed as: $\psi^\star(\mathbf{x}) = 0$ if $P_0(\mathbf{x}, \widetilde{\mathbf{y}})/P_1(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \mathcal{C}_1$; $\psi^\star(\mathbf{x}) = 1$ if $P_1(\mathbf{x}, \widetilde{\mathbf{y}})/P_0(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \mathcal{C}_1$; and $\psi^\star(\mathbf{x}) = 1/2$ otherwise, where $\mathcal{C}_1 \triangleq (\varrho(\epsilon) + \varpi_1)/\{1 - (\varrho(\epsilon) + \varpi_1)\} > 1$ serves as a threshold for the likelihood ratio test. Consequently, for a data point $(\mathbf{x}, \widetilde{\mathbf{y}})$, if $P_1(\mathbf{x}, \widetilde{\mathbf{y}})/P_0(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \mathcal{C}_1$, we assign a robust pseudo-label $\mathrm{y}^\star = 1$; if $P_0(\mathbf{x}, \widetilde{\mathbf{y}})/P_1(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \mathcal{C}_1$, we assign $\mathrm{y}^\star = 0$. Leveraging the likelihood ratio format also facilitates extending the robust pseudo-label selection method to the multi-class case by considering pairwise comparisons. Specifically, if $P_{k^\star}(\mathbf{x}, \widetilde{\mathbf{y}})/\max_{j \neq k^\star} P_j(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \mathcal{C}_1$, we assign the pseudo-label $\mathrm{y}^\star = k^\star$ to the instance.

Remark 3.2. Existing pseudo-labeling methods [9, 19] typically identify the underlying true label as the one with the highest probability in the approximated true label posterior. In contrast, the proposed approach in Remark 3.1 considers both the highest and second-highest predicted probabilities. A pseudo-label is assigned only if the ratio of these probabilities exceeds a specified threshold. This strategy ensures that pseudo-labels are assigned to instances with high confidence, effectively filtering out uncertain data.

Remark 3.3. In the special case where $P_j(\mathbf{x}, \widetilde{\mathbf{y}}) \propto \tau_j(\widetilde{\mathbf{y}}; \mathbf{x}) P_j(\mathbf{x})$, with $\tau_j(\widetilde{\mathbf{y}}; \mathbf{x}) = P^*(\widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}} | \mathrm{Y} = j, \mathbf{X} = \mathbf{x})$ denoting the noisy label transition probability and $P_j(\mathbf{x})$ representing a proper prior for $\mathrm{Y} = j$ conditional on $\mathbf{x}$ for $j = 0, 1$, previous studies have indicated the existence of a Chernoff information-type bound on the probability of error for robust pseudo-label selection, as described in Remark 3.1 [10, 20]. Specifically, for a fixed instance $\mathbf{x}$, let a pseudo-label $\mathrm{Y}^\star$ be generated as described in Remark 3.1, which depends on $\mathbf{x}$ and the corresponding noisy label vector $\widetilde{\mathbf{Y}}$. Consider the Bayes error, defined as $\Re_{\text{Bayes}} \triangleq \sum_{j=0,1} P_j(\mathbf{x}) P^*(\mathrm{Y}^\star \neq j | \mathrm{Y} = j, \mathbf{x})$. According to Section 11.9 of [20], $\Re_{\text{Bayes}} \leq \exp\{-C(\tau_0(\cdot; \mathbf{x}), \tau_1(\cdot; \mathbf{x}))\}$, where $C(\cdot, \cdot)$ represents the Chernoff information between two distributions.

Remark 3.4. In practice, one can use either uninformative priors, such as a uniform prior for each class, or informative priors derived from pre-trained or concurrently trained models for $P_j(\mathbf{x})$ as discussed in Remark 3.3 [10, 13]. Moreover, the estimation of $P_j(\mathbf{x}, \widetilde{\mathbf{y}})$ is not limited to Bayes's rule. For example, [11] proposed aggregating data and noisy label information by maximizing the f-mutual information gain.

Remark 3.5. Theorem 3.1 is developed based on the assumption that the function $\mathcal{T}$ is convex or concave. In our experiments, we use the cross-entropy loss for ℓ, meaning $\mathcal{T}(t) = -\log t$ for $t > 0$. To meet the required conditions, we clip its input to $[0.01, 1 - 0.01]$ to ensure $\mathcal{T}(\cdot)$ remains bounded.

Next, we extend the preceding development for binary classification to multi-class scenarios with $K > 2$. Letting $\mathcal{T}(\cdot)$ in (7) be specified as $\mathcal{T}(t) = 1 - t$, we extend loss function form (7) to facilitate the worst-case misclassification probability in multi-class scenarios: $\ell(\psi(\mathbf{x}), \mathrm{y}) = \sum_{j=1}^{K} \mathbf{1}(\mathrm{y} = j)\{1 - \psi(\mathbf{x})_j\}$. For ease of presentation, we sometimes omit the dependence on $\mathbf{x}$ and $\widetilde{\mathbf{y}}$ in the notation. Specifically, for $j \in [K]$, we let $P_j \triangleq P_j(\mathbf{x}, \widetilde{\mathbf{y}}) \triangleq P(\mathrm{Y} = j|\mathbf{x}, \widetilde{\mathbf{y}})$ and $\psi_j \triangleq \psi(\mathbf{x})_j$. In a manner similar to deriving (8), given $\mathbf{x}$, minimizing (5) with respect to $\psi(\mathbf{x})$ can be expressed as:

$$\inf_{\psi\in\Psi}\inf_{\gamma\geq 0}\Big[\gamma\epsilon^p + \sum_{j=1}^{K} P_j \max\{1-\psi_1-\gamma\kappa^p,\ldots,1-\psi_{j-1}-\gamma\kappa^p, \\ 1-\psi_j, 1-\psi_{j+1}-\gamma\kappa^p,\ldots,1-\psi_K-\gamma\kappa^p\}\Big]. \tag{9}$$

Theorem 3.2 (Optimal Action for Single Data Point: Multi-class Case)**.** *Let $\{P_1,\ldots,P_K\}$ be arranged in decreasing order, denoted $P^{(1)} \geq \ldots \geq P^{(K)}$, with the associated indexes denoted $\chi(1),\ldots,\chi(K)$. Let $\psi^\star$ denote the solution of the outer optimization problem in (9). For $j \in [K]$, let $\psi^{\star(j)}$ denote the $\chi(j)$-th component of $\psi(\mathbf{x})$ corresponding to $P^{(j)}$. Then, the elements of $\psi^\star$ are given as follows:*

(a). If $\frac{1}{K} \geq \frac{1}{k}\sum_{j=1}^{k} P^{(j)} - \frac{1}{k}\varrho(\epsilon)$ for all $k \in [K-1]$, then $\psi^{\star(j)} = \frac{1}{K}$ for all $j \in [K]$.

(b). If there exists some $k_0 \in [K-1]$ such that $\frac{1}{k_0}\sum_{j=1}^{k_0} P^{(j)} - \frac{1}{k_0}\varrho(\epsilon) > \frac{1}{K}$, and $\frac{1}{k_0}\sum_{j=1}^{k_0} P^{(j)} - \frac{1}{k_0}\varrho(\epsilon) \geq \frac{1}{k}\sum_{j=1}^{k} P^{(j)} - \frac{1}{k}\varrho(\epsilon)$ for all $k \in [K-1]$, then $\psi^{\star(j)} = \frac{1}{k_0}$ for $j \in [k_0]$ and $\psi^{\star(j)} = 0$ for $j = k_0+1,\ldots,K$.

Remark 3.6. The robust pseudo-labeling method described in Remark 3.1 can also be extended from Theorem 3.2. Specifically, by Theorem 3.2, if $P^{(1)} \geq \max\{\frac{1}{K} + \varrho(\epsilon), P^{(2)} + \varrho(\epsilon)\}$, then the optimal solution is: $\psi^{\star(1)} = 1$ and $\psi^{\star(j)} = 0$ for $j = 2,\ldots,K$, which can also be expressed in a likelihood ratio format and applied to assign robust pseudo-labels.

3.2 Closed-Form Robust Risk

We investigate the empirical robust risk (5) by examining its closed form expression. For $i \in [K]$ and $j \in [K]$, we let $P_{i,j} \triangleq P_j(\mathbf{x}_i, \widetilde{\mathbf{y}}_i) \triangleq P(\mathrm{Y} = j|\mathbf{X} = \mathbf{x}_i, \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}_i)$ and $\psi_{i,j} \triangleq \psi(\mathbf{x}_i)_j$. For simplicity, we denote the Wasserstein robust loss in (5) and the nominal loss respectively as:

$$\widehat{\mathfrak{R}}_\epsilon = \inf_{\gamma\geq 0}\Big[\gamma\epsilon^p + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} P_{i,j}\max\big\{\mathcal{T}(\psi_{i,1}) - \gamma\kappa^p,\ldots,\mathcal{T}(\psi_{i,j-1}) - \gamma\kappa^p \\ \mathcal{T}(\psi_{i,j}), \mathcal{T}(\psi_{i,j+1}) - \gamma\kappa^p,\ldots,\mathcal{T}(\psi_{i,K}) - \gamma\kappa^p\big\}\Big] \text{ for } \epsilon > 0; \tag{10}$$

$$\widehat{\mathfrak{R}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} P_{i,j}\mathcal{T}(\psi_{i,j}). \tag{11}$$

For given $\mathbf{x}_i$, we sort $\{\psi_{i,1},\ldots,\psi_{i,K}\}$ in decreasing order, denoted as $\psi_i^{(1)} \geq \ldots \geq \psi_i^{(K)}$. Let $\alpha_{i,j} \triangleq \mathcal{T}(\psi_i^{(K)}) - \mathcal{T}(\psi_{i,j})$ for $i \in [n]$ and $j \in [K]$, and sort $\{\alpha_{i,j} : i \in [n], j \in [K]\}$ in decreasing order, denoted as $\alpha^{(1)} \geq \ldots \geq \alpha^{(nK)}$. Correspondingly, the $P_{i,j}$ values with the associated indexes are denoted as $P^{(1)},\ldots,P^{(nK)}$. For any $\varrho(\epsilon)$, define an associated positive integer $s^* \in [nK+1]$ as follows: if $\frac{1}{n}P^{(1)} < \varrho(\epsilon) < \frac{1}{n}\sum_{t=1}^{nK} P^{(t)}$, then there exists $s^* \in \{2,\ldots,nK\}$ such that $\frac{1}{n}\sum_{t=1}^{s} P^{(t)} < \varrho(\epsilon)$ for $s < s^*$, and $\frac{1}{n}\sum_{t=1}^{s} P^{(t)} \geq \varrho(\epsilon)$ for $s \geq s^*$; if $\varrho(\epsilon) \leq \frac{1}{n}P^{(1)}$, then s^* is set as 1; if $\varrho(\epsilon) \geq \frac{1}{n}\sum_{t=1}^{nK} P^{(t)}$, then s^* is set as $nK+1$.

Let $\gamma_\psi^\star$ denote the optimal value of γ in $\widehat{\mathfrak{R}}_\epsilon$ in (10), where its dependence on ϵ is implicit, but its dependence on ψ is explicit. The following theorem presents this value, based on which we demonstrate that the Wasserstein robust loss $\widehat{\mathfrak{R}}_\epsilon$ can be expressed as the nominal loss plus an additional term $\frac{1}{n}\sum_{t=1}^{s^*-1} P^{(t)}\alpha^{(t)}\mathbf{1}(s^* > 1)$ that prevents the classifier from becoming overly certain on the data.

Theorem 3.3 (Closed-Form Robust Risk)**.** *The optimal value of γ in (10) is given by $\gamma_\psi^\star \triangleq \alpha^{(s^*)}/\kappa^p$, and the resulting robust risk is expressed as*

$$\widehat{\mathfrak{R}}_\epsilon = \widehat{\mathfrak{R}} + \frac{1}{n}\sum_{t=1}^{s^*-1} P^{(t)}\alpha^{(t)}\mathbf{1}(s^* > 1) + O\left(\frac{1}{n}\right)\alpha^{(s^*)}.$$

Remark 3.7. Theorem 3.3 shows that minimizing the Wasserstein robust loss $\widehat{\mathfrak{R}}_\epsilon$ in (10) effectively minimizes the nominal loss $\widehat{\mathfrak{R}}$ in (11) while simultaneously penalizing terms associated with $|\alpha_i|$ values exceeding a certain threshold $|\alpha^{s^*}|$, weighted by the corresponding reference probability values. This minimization prevents the classifier from becoming overly confident in certain data points, particularly when there are potential misspecifications in the approximated true label posterior.

Remark 3.8. As suggested by Remark 2.2, Theorem 3.3 provides a guideline for balancing robustness and model fitting by deriving the optimal value for γ in (10). In Section 3.3, we develop a one-step update method for determining $\gamma^\star_\psi$.

3.3 Training using Conditional Distributionally Robust True Label Posterior

In this subsection, we outline the steps for approximating the true label posterior, constructing the pseudo-empirical distribution as the reference distribution for solving the robust risk minimization problem (5), and subsequently training classifiers robustly. The pseudo code for the training process is provided in Algorithm 1 in Appendix B.1. Here we elaborate on the details.

Approximating noise transitions probabilities. We begin by warming up the classifiers on the noisy training data, denoted $\check{\mathcal{D}} = \{\mathbf{x}_i, \check{\mathrm{y}}_i\}_{i=1}^n$, where $\check{\mathrm{y}}_i$ represents the majority vote label for instance $\mathbf{x}_i$, determined by the label that receives the highest number of votes from the annotators. After warming up the classifiers for 20-30 epochs, we sort the dataset by the cross-entropy loss values and collect a subset of size m with the smallest m losses, denoted as $\mathcal{D}^\star_0 = \{\mathbf{x}_i, \check{\mathrm{y}}_i\}_{i=1}^m$, where $m \ll n$, and the ratio of m to n is set to 1 minus the estimated noise rate. Next, we estimate the noise transition probabilities by $\widehat{\tau}_j(\widetilde{\mathbf{y}}) = \sum_{i=1}^m \mathbf{1}(\widetilde{\mathbf{y}} = \widetilde{\mathbf{y}}_i, \check{\mathrm{y}}_i = j)/\sum_{i=1}^m \mathbf{1}(\check{\mathrm{y}}_i = j)$ for $\widetilde{\mathbf{y}} \in [K]^R$ and $j \in [K]$ (Line 1 of Algorithm 1). With $\widehat{\tau}_j(\widetilde{\mathbf{y}})$, we then iteratively update the approximated true label posterior, construct the pseudo-empirical distribution, and robustly train the classifiers (Lines 2-13 of Algorithm 1). Here, we employ the straightforward frequency-counting method for noise transition estimation for simplicity. However, our approach is versatile and can be integrated with various methods for estimating the noise transition matrices or the true label posterior. Additional experimental results using more advanced transition matrix estimation methods are provided in Appendix B.

Constructing a pseudo-empirical distribution. We train two classifiers, $\psi^{(1)}$ and $\psi^{(2)}$, in parallel each serving as an informative prior for the other. In the tth epoch, the approximated true label posterior with prior $\psi^{(\iota)}$ is updated as $\widehat{P}^{(\iota)}_j(\mathbf{x}, \widetilde{\mathbf{y}}) \triangleq \widehat{P}^{(\iota)}(\mathrm{Y} = j | \widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}, \mathbf{X} = \mathbf{x}) \propto \psi^{(\iota)}_j(\mathbf{x}) \cdot \widehat{\tau}_j(\widetilde{\mathbf{y}})$ (Line 4 of Algorithm 1), where $\psi^{(\iota)}_j(\mathbf{x})$ denotes the jth element of the vector-valued function $\psi^{(\iota)}(\mathbf{x})$ for $j \in [K]$ and $\iota = 1, 2$. As described in Remark 3.1, for $i \in [n]$, if $\widehat{P}^{(\iota)}_{k^\star}(\mathbf{x}_i, \widetilde{\mathbf{y}}_i)/\max_{j \neq k^\star} \widehat{P}^{(\iota)}_j(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \mathcal{C}$ for a pre-specified threshold $\mathcal{C} > 1$, we assign the robust pseudo-label $\mathrm{y}^\star_i = k^\star$ to the instance and collect it into $\mathcal{D}^\star_{t,\iota}$ (Lines 5-7 of Algorithm 1). The pseudo-empirical distribution $P^\star_{t,\iota}$ is updated based on $\mathcal{D}^\star_{t,\iota}$ (Line 8 of Algorithm 1).

Robustly training the classifiers. For $\iota = 1, 2$, let $\backslash\iota \triangleq 1$ if $\iota = 2$; and $\backslash\iota \triangleq 2$ if $\iota = 1$. With the updated pseudo-empirical distribution, the classifier $\psi^{(\iota)}$ is then trained by minimizing the empirical robust risk (5) with the reference distribution $P^\star_{t,\backslash\iota}$ (Line 9 of Algorithm 1). After updating the classifier $\psi^{(\iota)}$ with $\gamma^{(\iota)}_{t-1}$ from the previous iteration, we take one step to update the γ value $\gamma^{(\iota)}_t$. In particular, as suggested by Theorem 3.3, we use $\gamma_{0,t} = |\alpha^{(s^*)}|/\kappa^p$ as a reference value for γ (Lines 11-12 f Algorithm 1). We then update $\gamma^{(\iota)}_t$ by minimizing $\left[\gamma\{\epsilon^p - \mathbb{E}_{P^\star_{t,\backslash\iota}} c^p(y', \mathrm{Y})\} + \frac{\lambda}{2}(\gamma - \gamma_0)^2\right]$ (Line 13 of Algorithm 1) with respect to γ, where y' is determined by (5) after updating $\psi^{(\iota)}$, and $\lambda > 0$ is a positive constant that determines the learning rate of γ.

4 Experimental Results

Datasets and model architectures. We evaluate the performance of the proposed AdaptCDRP on two datasets, CIFAR-10 and CIFAR-100 [21], by generating synthetic noisy labels (details provided below), as well as four datasets, CIFAR-10N [22], CIFAR-100N [22], LabelMe [23, 24], and Animal-10N [25], which contain human annotations. For all datasets except LabelMe, we set aside 10% of the original data, together with the corresponding synthetic or human annotated noisy labels, to validate the model selection procedure. We use the ResNet-18 architecture [26] for CIFAR-10 and

Table 1: Average accuracies (with associated standard errors expressed after the ± signs) for learning the CIFAR-10 and CIFAR-100 datasets ($R = 5$).

Method	CIFAR-10			CIFAR-100		
	IDN-LOW	IDN-MID	IDN-HIGH	IDN-LOW	IDN-MID	IDN-HIGH
CE (Clean)		$88.60_{\pm 0.79}$			$58.75_{\pm 0.55}$	
CE (MV)	$80.90_{\pm 0.88}$	$76.05_{\pm 0.70}$	$69.65_{\pm 1.73}$	$50.96_{\pm 0.49}$	$44.80_{\pm 0.99}$	$38.51_{\pm 0.66}$
CE (EM) [7]	$81.15_{\pm 0.74}$	$75.84_{\pm 0.97}$	$69.85_{\pm 1.43}$	$51.29_{\pm 1.00}$	$45.24_{\pm 0.41}$	$38.01_{\pm 0.90}$
Co-teaching [30]	$83.08_{\pm 0.52}$	$80.58_{\pm 0.36}$	$\mathbf{81.30}_{\pm 0.82}$	$53.10_{\pm 0.98}$	$47.25_{\pm 0.82}$	$44.11_{\pm 0.31}$
Co-teaching+ [31]	$81.17_{\pm 0.55}$	$78.23_{\pm 0.43}$	$71.84_{\pm 1.13}$	$53.10_{\pm 0.64}$	$47.92_{\pm 0.76}$	$41.33_{\pm 0.81}$
CoDis [32]	$85.33_{\pm 0.39}$	$\mathbf{82.02}_{\pm 0.41}$	$78.67_{\pm 0.46}$	$\mathbf{58.66}_{\pm 0.44}$	$52.27_{\pm 0.64}$	$46.12_{\pm 0.66}$
LogitClip [33]	$\mathbf{85.39}_{\pm 0.35}$	$80.87_{\pm 0.42}$	$75.36_{\pm 0.79}$	$57.79_{\pm 0.77}$	$\mathbf{53.14}_{\pm 0.37}$	$\mathbf{49.00}_{\pm 0.35}$
DoctorNet [34]	$81.85_{\pm 0.41}$	$78.69_{\pm 0.75}$	$76.26_{\pm 1.28}$	$52.61_{\pm 0.70}$	$47.80_{\pm 0.86}$	$43.50_{\pm 0.53}$
MBEM [9]	$82.37_{\pm 0.77}$	$78.05_{\pm 0.83}$	$71.43_{\pm 2.43}$	$52.20_{\pm 0.07}$	$45.26_{\pm 0.50}$	$38.92_{\pm 0.69}$
CrowdLayer [27]	$83.98_{\pm 0.35}$	$77.76_{\pm 1.06}$	$67.77_{\pm 1.69}$	$51.28_{\pm 0.64}$	$45.28_{\pm 0.64}$	$38.93_{\pm 0.76}$
TraceReg [12]	$80.72_{\pm 0.79}$	$77.71_{\pm 1.36}$	$67.86_{\pm 1.77}$	$51.43_{\pm 0.61}$	$45.08_{\pm 0.57}$	$38.69_{\pm 1.01}$
Max-MIG [11]	$81.00_{\pm 0.72}$	$75.90_{\pm 0.52}$	$70.96_{\pm 0.96}$	$51.76_{\pm 1.11}$	$44.93_{\pm 0.71}$	$38.70_{\pm 0.49}$
CoNAL [35]	$81.60_{\pm 0.82}$	$76.02_{\pm 0.79}$	$69.50_{\pm 1.89}$	$51.61_{\pm 1.14}$	$44.19_{\pm 0.62}$	$38.24_{\pm 0.29}$
CCC [36]	$84.81_{\pm 0.89}$	$81.29_{\pm 0.66}$	$77.28_{\pm 1.05}$	$56.65_{\pm 0.55}$	$50.68_{\pm 0.40}$	$43.94_{\pm 0.95}$
Ours (AdaptCDRP)	$\mathbf{88.09}_{\pm 0.37}$	$\mathbf{87.37}_{\pm 0.29}$	$\mathbf{86.62}_{\pm 0.45}$	$\mathbf{60.20}_{\pm 0.15}$	$\mathbf{56.65}_{\pm 1.03}$	$\mathbf{54.24}_{\pm 0.99}$

CIFAR-10N, and the ResNet-34 architecture [26] for CIFAR-100 and CIFAR-100N. Following [27], we employ a pretrained VGG-16 model with a 50% dropout rate for the LabelMe dataset. In line with [25], the VGG19-BN architecture [28] is used for the Animal-10N dataset. Further details on the datasets and experimental setup are provided in Appendix B.1.

Noise generation. We generate synthetic annotations on the CIFAR-10 and CIFAR-100 datasets using Algorithm 2 from [29]. Three groups of annotators, labeled as IDN-LOW, IDN-MID, and IDN-HIGH, are considered, with average labeling error rates of approximately 20%, 35%, and 50%, respectively, representing low, intermediate, and high error rates. Each group consists of $R = 5$ annotators. To assess the algorithms in an incomplete labeling setting, we randomly select only one annotation per instance from the R annotators for the training dataset rather than using all available annotations [10]. Further details on noise generation are provided in Appendix B.1.

Comparison with SOTA methods. We compare our method with a comprehensive set of state-of-the-art approaches, including: (1) CE (Clean) with clean labels; (2) CE (MV) with majority vote labels; (3) CE (EM) [7]; (4) Co-teaching [30]; (5) Co-teaching+ [31]; (6) CoDis [32]; (7) LogitClip [33]; (8) DoctorNet [34]; (9) MBEM [9]; (10) CrowdLayer [27]; (11) TraceReg [12]; (12) Max-MIG [11]; (13) CoNAL [35]; and (14) CCC [36]. We report the average test accuracy over five repeated experiments, each with a different random seed, on synthetic datasets, CIFAR-10 and CIFAR-100, with instance-dependent label noise introduced at low, intermediate, and high error rates. Standard errors are shown following the plus/minus sign (±), and the two highest accuraries are highlighted in bold. Table 2 presents evaluation results on four real-world datasets. As shown, our AdaptCDRP consistently outperforms competing methods across all scenarios. To further explore the impact of annotation sparsity, we conduct additional experiments with the number of annotators ranging from 5 to 100, with each instance labeled only once. Figure 1 illustrates the average accuracy across different numbers of annotators on CIFAR-10, highlighting the advantages of the proposed method under diverse settings. The results for the CIFAR-100 dataset are shown in Figure 3 in Appendix B.2.

Hyper-parameter analysis. We investigate the impact of the hyperparameter ϵ in the empirical robust risk (5). Under our experiment setup, ϵ should be chosen within $(0, 1/K)$ for a K-class classification problem with $K \geq 2$ as demonstrated in the proof of Theorem 3.3 in Appendix A.8. Hence, we take $\epsilon \in (0, 0.1)$ for CIFAR-10 and $\epsilon \in (0, 0.01)$ for CIFAR-100, with the results presented in Figure 2. The results suggest that setting ϵ near zero leads to relatively low test accuracies, highlighting the importance of CDRO under model specification when handling noisy labels. Furthermore, continually increasing ϵ eventually results in a drop in accuracy due to excessive noise injection into the data.

Additional experimental results. To further evaluate the performance of the proposed method across various scenarios, we conducted additional experiments, detailed in Appendix B.2. Specifically, we compare different annotation aggregation methods, present average test accuracies and robust pseudo-label accuracies during training, assess sensitivity to the number of warm-up epochs, explore different noise transition estimation methods, and examine the impact of sparse annotation.

Table 2: Average accuracies (with associated standard errors expressed after the ± signs) for learning the CIFAR-10N, CIFAR-100N, LabelMe, and Animal-10N datasets.

Method	CIFAR-10N	CIFAR-100N	LabelMe	Animal-10N
CE (MV)	$82.82_{\pm 0.05}$	$46.26_{\pm 0.81}$	$79.49_{\pm 0.48}$	$79.88_{\pm 0.38}$
CE (EM) [7]	$83.14_{\pm 0.80}$	$46.14_{\pm 0.69}$	$80.64_{\pm 0.55}$	$80.18_{\pm 0.34}$
Co-teaching [30]	$85.66_{\pm 0.54}$	$52.34_{\pm 0.31}$	$79.71_{\pm 0.55}$	$\mathbf{81.96}_{\pm 0.48}$
Co-teaching+ [31]	$82.25_{\pm 0.21}$	$50.52_{\pm 0.40}$	$81.55_{\pm 0.92}$	$81.24_{\pm 0.22}$
CoDis [32]	$\mathbf{87.23}_{\pm 0.45}$	$\mathbf{52.66}_{\pm 0.44}$	$81.85_{\pm 0.49}$	$73.08_{\pm 0.35}$
LogitClip [33]	$86.37_{\pm 0.43}$	$51.50_{\pm 0.58}$	$81.75_{\pm 0.90}$	$70.89_{\pm 0.65}$
DoctorNet [34]	$84.52_{\pm 0.69}$	$46.21_{\pm 0.81}$	$79.09_{\pm 0.40}$	$79.96_{\pm 0.55}$
MBEM [9]	$85.49_{\pm 0.43}$	$46.74_{\pm 0.69}$	$80.10_{\pm 1.09}$	$76.96_{\pm 3.17}$
CrowdLayer [27]	$82.84_{\pm 0.24}$	$47.43_{\pm 0.59}$	$82.95_{\pm 0.21}$	$79.70_{\pm 0.35}$
TraceReg [12]	$82.94_{\pm 0.27}$	$47.71_{\pm 0.70}$	$83.10_{\pm 0.15}$	$80.34_{\pm 0.66}$
Max-MIG [11]	$85.12_{\pm 0.36}$	$46.56_{\pm 0.64}$	$\mathbf{83.25}_{\pm 0.26}$	$79.78_{\pm 0.80}$
CoNAL [35]	$83.01_{\pm 0.21}$	$49.37_{\pm 0.48}$	$82.96_{\pm 0.30}$	$80.45_{\pm 0.49}$
CCC [36]	$86.45_{\pm 0.53}$	$48.57_{\pm 0.58}$	$83.18_{\pm 0.38}$	$78.36_{\pm 0.35}$
Ours (AdaptCDRP)	$\mathbf{88.25}_{\pm 0.34}$	$\mathbf{53.42}_{\pm 0.64}$	$\mathbf{83.36}_{\pm 0.68}$	$\mathbf{83.08}_{\pm 0.39}$

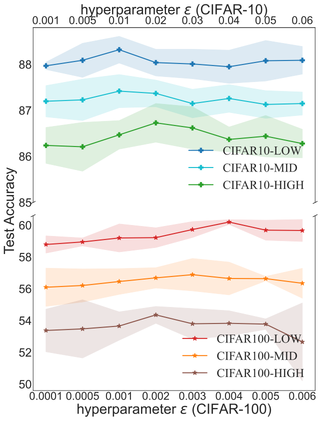
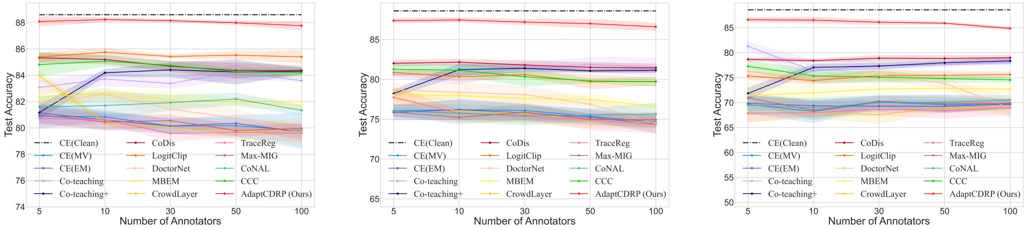


Figure 2: Average accuracy on the CIFAR-10 and CIFAR-100 datasets ($R = 5$) for different ϵ values.



(a) IDN-LOW

(b) IDN-MID

(c) IDN-HIGH

Figure 1: Average test accuracy on the CIFAR-10 dataset with varying numbers of annotators. The shaded areas are constructed using the associated standard deviations.

5 Conclusion

In this paper, we address the challenge of learning from noisy annotations by estimating true label posteriors using the CDRO framework. We formulate the problem as minimizing the worst-case risk within a distance-based ambiguity set, which constrains the conditional distributional uncertainty around a reference distribution. By deriving the dual form of the worst-case risk and finding the analytical solution to the robust risk minimization problem for each data point, we propose a novel approach for determining robust pseudo-labels using the likelihood ratio test. This approach further leads to the construction of a pseudo-empirical distribution that serves as a robust reference probability distribution in CDRO. We also derive a closed-form expression of the empirical robust risk and identify the optimal Lagrange multiplier for the dual problem. This leads to a guideline for balancing robustness and model fitting in a principled way and inspires an efficient one-step update method for the Lagrange multiplier.

Limitations and Extensions. Our development here does not focus on precisely estimating the noise transition matrix or the true label posterior. Further research may be conducted to address the sparse annotation problem and improve estimates of the true label posterior. This can be accomplished through several approaches: (1) employing regularization techniques to mitigate the impact of small sample sizes by smoothing estimates and reducing sensitivity to outliers; (2) leveraging subgroup structures among annotators to capture additional nuances; and (3) directly modeling the true label posterior by integrating both data and noisy label information, moving beyond the limitation of purely applying Bayes's rule.

https://doi.org/10.52202/079017-2627

Acknowledgements

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

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

[2] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[3] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 233–242, 2017.

[4] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2006.

[5] Grace Y Yi. *Statistical Analysis with Measurement Error or Misclassification*. Springer, 2017.

[6] Grace Y Yi, Aurore Delaigle, and Paul Gustafson. *Handbook of Measurement Error Models*. CRC Press, 2021.

[7] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C*, 28(1):20–28, 1979.

[8] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pages 2035–2043, 2009.

[9] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017.

[10] Hui Guo, Boyu Wang, and Grace Y Yi. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. In *Advances in Neural Information Processing Systems*, volume 36, pages 347–386, 2023.

[11] Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Max-mig: an information theoretic approach for joint learning from crowds. *arXiv preprint arXiv:1905.13436*, 2019.

[12] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019.

[13] Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew G Wilson. Pre-train your loss: Easy bayesian transfer learning with informative priors. In *Advances in Neural Information Processing Systems*, volume 35, pages 27706–27715, 2022.

[14] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing Systems*, volume 32, pages 6838–6849, 2019.

[15] Hisham Husain and Jeremias Knoblauch. Adversarial interpretation of bayesian inference. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167, pages 553–572. Proceedings of Machine Learning Research, 2022.

[16] Alexander Shapiro and Alois Pichler. Conditional distributionally robust functionals. *Operations Research*, 2023.

[17] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[18] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

https://doi.org/10.52202/079017-2627

[19] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.

[20] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[21] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[22] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

[23] Filipe Rodrigues, Mariana Lourenco, Bernardete Ribeiro, and Francisco C Pereira. Learning supervised topic models for classification and regression from crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2409–2422, 2017.

[24] Antonio Torralba, Bryan C. Russell, and Jenny Yuen. Labelme: Online image annotation and applications. *Proceedings of the IEEE*, 98(8):1467–1484, 2010.

[25] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5907–5915, 2019.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[27] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 1611–1618, 2018.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[29] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, volume 33, pages 7597–7610, 2020.

[30] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, pages 8527–8537, 2018.

[31] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7164–7173, 2019.

[32] Xiaobo Xia, Bo Han, Yibing Zhan, Jun Yu, Mingming Gong, Chen Gong, and Tongliang Liu. Combating noisy labels with sample selection by mining high-discrepancy examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1843, 2023.

[33] Hongxin Wei, Huiping Zhuang, Renchunzi Xie, Lei Feng, Gang Niu, Bo An, and Yixuan Li. Mitigating memorization of noisy labels by clipping the model prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 36868–36886, 2023.

[34] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, pages 3109–3118, 2018.

[35] Zhendong Chu, Jing Ma, and Hongning Wang. Learning from crowds by modeling common confusions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5832–5840, 2021.

[36] Hansong Zhang, Shikun Li, Dan Zeng, Chenggang Yan, and Shiming Ge. Coupled confusion correction: Learning from crowds with sparse annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16732–16740, 2024.

[37] David G Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*, volume 2. Springer, 1984.

[38] Stephen P Bradley, Arnoldo C Hax, and Thomas L Magnanti. *Applied Mathematical Programming*. Addison-Wesley Publishing Company, 1977.

https://doi.org/10.52202/079017-2627

[39] Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[40] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 31, pages 2687–2696, 2018.

[41] John Hunter. The supremum and infimum. `https://www.math.ucdavis.edu/~hunter/m125b/ch2.pdf`.

[42] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

[43] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[44] Hongwei Li and Bin Yu. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086*, 2014.

[45] Maria Sofia Bucarelli, Lucas Cassano, Federico Siciliano, Amin Mantrach, and Fabrizio Silvestri. Leveraging inter-rater agreement for classification in the presence of noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3439–3448, 2023.

[46] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11447–11457, 2020.

[47] Shahana Ibrahim, Tri Nguyen, and Xiao Fu. Deep learning from crowdsourced labels: Coupled cross-entropy minimization, identifiability, and regularization. In *The Eleventh International Conference on Learning Representations*, 2023.

https://doi.org/10.52202/079017-2627

SUPPLEMENTARY MATERIAL

A Technical Details

A.1 Preliminaries about Linear Programming and Concentration Bounds

A *linear program* (LP) is an optimization problem of the form

$$\begin{aligned} \max_{\mathbf{x}\in\mathbb{R}^n} \ & \mathbf{c}^\top \mathbf{x} \\ s.t. \ & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \mathbf{x} \ge 0, \end{aligned} \tag{A1}$$

where $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$ are given, and $\mathbf{A}$ is a specified $m \times n$ matrix. Here, "$\le$" represents elementwise inequality for vectors. The expression $\mathbf{c}^\top \mathbf{x}$ is called the *objective function*, and the set $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}, \mathbf{x} \ge 0\}$ defines the *feasible region* of the linear program. By introducing slack variables, any linear program can be converted to the following *standard form*:

$$\begin{aligned} \max_{\mathbf{x}\in\mathbb{R}^n} \ & \mathbf{c}^\top \mathbf{x} \\ s.t. \ & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \ge 0. \end{aligned} \tag{A2}$$

We begin by introducing the concept of extreme point of related properties.

Definition A.1 ([37, Chapter 2]). A point $\mathbf{z}$ in a convex set Θ is called an *extreme point* of Θ if there do not exist two distinct points $\mathbf{z}', \mathbf{z}'' \in \Theta$ and a scalar ν with $0 < \nu < 1$ such that $\mathbf{z} = \nu\mathbf{z}' + (1-\nu)\mathbf{z}''$.

The following lemma shows that optimal solutions of a linear program are located among the extreme points.

Lemma 1 ([37, Chapter 2]). *If a linear programming problem has a finite optimal solution (i.e., a feasible solution that optimizes the objective function), then there is a finite optimal solution that is an extreme point of the constraint set.*

The following lemma on strong duality in linear programming will be used in the proof of Proposition 2.1 and Remark 2.3.

Lemma 2 (Duality Theorem of Linear Programming; [38, Chapter 4]). *Let* $\mathbf{c} = (c_1, \dots, c_n)^\top$ *and* $\mathbf{b} = (b_1, \dots, b_m)^\top$ *be given vectors, and let* $\mathbf{A} = [a_{ij}]$ *be a given* $m \times n$ *matrix with* a_{ij} *being its* (i, j) *element. Define the* ***primal*** *problem as:*

$$\begin{cases} \displaystyle\max_{x_1,\dots,x_n\in\mathbb{R}} \ \mathsf{P}, \ \textit{with } \mathsf{P} \triangleq \sum_{j=1}^{n} c_j x_j, \\ \displaystyle s.t. \ \sum_{j=1}^{n} a_{ij} x_j \le b_i \ \textit{for} \ i \in [m], \\ \quad x_j \ge 0 \ \textit{for} \ j \in [n], \end{cases}$$

or equivalently written in compact form:

$$\begin{cases} \displaystyle\max_{\mathbf{x}\in\mathbb{R}^n} \ \mathsf{P}, \ \textit{with } \mathsf{P} \triangleq \mathbf{c}^\top \mathbf{x} \ \textit{and} \ \mathbf{x} = (x_1, \dots, x_n)^\top, \\ s.t. \ \mathbf{A}\mathbf{x} \ge \mathbf{b}, \\ \quad \mathbf{x} \ge 0. \end{cases}$$

The corresponding ***dual*** *linear problem is:*

$$\begin{cases} \displaystyle\min_{y_1,\dots,y_m\in\mathbb{R}} \ \mathsf{D}, \ \textit{with } \mathsf{D} \triangleq \sum_{i=1}^{m} b_i y_i, \\ \displaystyle s.t. \ \sum_{i=1}^{m} a_{ij} y_i \ge c_j \ \textit{for} \ j \in [n], \\ \quad y_i \ge 0 \ \textit{for} \ i \in [m], \end{cases}$$

or equivalently,

$$\begin{cases} \min\limits_{\mathbf{y}\in\mathbb{R}^m} \ \mathsf{D}, \textit{ with } \mathsf{D} \triangleq \mathbf{y}^\top \mathbf{b} \textit{ and } \mathbf{y} = (y_1, \ldots, y_m)^\top, \\ s.t.\ \mathbf{y}^\top \mathbf{A} \le \mathbf{c}^\top, \\ \quad \mathbf{y} \ge 0. \end{cases}$$

If the primal (or dual) problem has a finite optimal solution, then the dual (or primal) problem also has a finite solution, and the optimal values of the primal and dual problems are equal.

Next we introduce a concentration bound along with associated concepts, which will be used in the proof of Theorem 2.2.

Let Ω denote a subset of $\mathbb{R}$ and $f : \Omega^n \to \mathbb{R}$. We say that a function f satisfies the *bounded difference inequality* [39] with parameters $\{L_1, \ldots, L_n\}$ if for any $k \in [n]$,

$$\sup_{s_1,\ldots,s_n,s'_k\in\Omega} |f(s_1, \ldots, s_k, \ldots, s_n) - f(s_1, \ldots, s'_k, \ldots, s_n)| \le L_k. \tag{A3}$$

That is, for any $k \in [n]$, if we substitute s_k with s'_k while keeping other s_j fixed for all $j \ne k$, the function f changes by at most L_k.

Lemma 3 (Bounded Differences Inequality; [39, Corollary 2.21])**.** *Let $\boldsymbol{S} = (S_1, \ldots, S_n)^\top$ represent a random vector with independent components defined on a sample space Ω^n, and $f(\boldsymbol{S}) \triangleq f(S_1, \ldots, S_n)$ for a function $f : \Omega^n \to \mathbb{R}$. Suppose that f satisfies the bounded difference property with parameters $\{L_1, \ldots, L_n\}$. Then, for any $t \ge 0$,*

$$\mathbb{P}\left[f(\boldsymbol{S}) - \mathbb{E}\{f(\boldsymbol{S})\} \ge t\right] \le \exp\left\{-\frac{2t^2}{\sum_{i=1}^n L_i^2}\right\};$$
$$\mathbb{P}\left[f(\boldsymbol{S}) - \mathbb{E}\{f(\boldsymbol{S})\} \le -t\right] \le \exp\left\{-\frac{2t^2}{\sum_{i=1}^n L_i^2}\right\}.$$

We also introduce the following definitions and lemmas, which will be used to characterize the complexity of the function class.

Definition A.2 (Covering number; [39, Definition 5.1])**.** Let Θ denote a set and ρ a metric on Θ. For $t > 0$, a *t-cover* of Θ with respect to ρ is a set $\{\theta_i \in \Theta : i = 1, \ldots, N\}$ such that for each $\theta \in \Theta$, there exists some $i \in [N]$ such that $\rho(\theta, \theta_i) \le t$. The *$t$-covering number $N(t; \Theta, \rho)$* is the cardinality of the smallest t-cover.

Lemma 4. *Let Θ_j be a set equipped with a metric ρ_j for $j = 1, 2$, and define $\Theta = \Theta_1 \times \Theta_2$. Given $\alpha_1, \alpha_2 > 0$, define the metric ρ on Θ as $\rho(\theta, \theta') \triangleq \alpha_1\rho_1(\theta^1, \theta^{1'}) + \alpha_2\rho_2(\theta^2, \theta^{2'})$ for any $\theta \triangleq (\theta^1, \theta^2) \in \Theta$ and $\theta' \triangleq (\theta^{1'}, \theta^{2'}) \in \Theta$. Then for $t > 0$,*

$$N(t; \Theta, \rho) \le N(t/(2\alpha_1); \Theta_1, \rho_1) \times N(t/(2\alpha_2); \Theta_2, \rho_2).$$

Proof. For $j = 1, 2$, let $\bar{\Theta}_j \triangleq \{\theta^j_1, \ldots, \theta^j_{N_j}\}$ denote the smallest $t/(2\alpha_j)$-cover of Θ_j with respect to ρ_j. Then, by Definition A.2, $N(t/(2\alpha_j); \Theta_j, \rho_j) = N_j$ for $j = 1, 2$. For any $\theta = (\theta^1, \theta^2) \in \Theta$, by definition A.2, there exists $i_1 \in [N_1]$ and $i_2 \in [N_2]$ such that $\rho_1(\theta^1, \theta^1_{i_1}) \le t/(2\alpha_1)$ and $\rho_2(\theta^2, \theta^2_{i_2}) \le t/(2\alpha_2)$. Then, $\theta_i \triangleq (\theta^1_{i_1}, \theta^2_{i_2}) \in \bar{\Theta}_1 \times \bar{\Theta}_2 \subset \Theta$, and

$$\rho(\theta, \theta_i) = \alpha_1\rho_1(\theta^1, \theta^1_{i_1}) + \alpha_2\rho_2(\theta^2, \theta^2_{i_2}) \le t.$$

Hence, by Definition A.2, $\bar{\Theta}_1 \times \bar{\Theta}_2$ is a t-cover of Θ with respect to ρ and $N(t; \Theta, \rho) \le |\bar{\Theta}_1 \times \bar{\Theta}_2| = N_1 \times N_2 = N(t/(2\alpha_1); \Theta_1, \rho_1) \times N(t/(2\alpha_2); \Theta_2, \rho_2)$, where $|\cdot|$ represents the cardinality of a set. □

Lemma 5. *Let $\mathcal{I} \triangleq [a, b] \subset \mathbb{R}$ denote a closed interval on $\mathbb{R}$ with $a < b$. Define the metric ρ on $\mathcal{I}$ as $\rho(x, x') = |x - x'|$ for any $x, x' \in \mathcal{I}$. Then, for any $t > 0$, $N(t; \mathcal{I}, \rho) \le \frac{b-a}{2t} + 1$ if $t < \frac{b-a}{2}$, and $N(t; \mathcal{I}, \rho) = 1$ if $t \ge \frac{b-a}{2}$.*

https://doi.org/10.52202/079017-2627

Proof. Let $n_t = \lfloor \frac{b-a}{2} \rfloor$, where $\lfloor \cdot \rfloor$ represents the floor function. To prove the first result for $b-a \geq 2t$, we construct the following subset $\overline{\mathcal{I}}$ of $\mathcal{I}$:

$$\overline{\mathcal{I}} \triangleq \Big\{ a+t, a+t+2t, a+t+4t, \ldots, a+t+(n_t-1)\cdot 2t, \min(b, a+t+n_t \cdot 2t) \Big\}.$$

Clearly, $\mathcal{I} \subset \cup_{k \in [n_t]} [a+2t(k-1), a+2tk] \cup [a+2tn_t, b]$. Next, we verify that $\overline{\mathcal{I}}$ is a t-cover of $\mathcal{I}$ with respect to metric ρ. Specifically, for any $x \in \mathcal{I}$, there exists $k \in [n_t]$ such that $x \in [a+2t\cdot(k-1), a+2t\cdot k]$, or $x \in [a+2tn_t, b]$. For the former case,

$$\rho(x, a+t+2t\cdot(K-1)) \leq t;$$

and for the latter case,

$$\rho(x, \min(b, a+t+n_t \cdot 2t)) \leq t.$$

Hence, by Definition A.2, $\overline{\mathcal{I}}$ is a t-cover of $\mathcal{I}$, therefore $N(t; \mathcal{I}, \rho) \leq |\overline{\mathcal{I}}| = n_t + 1 \leq \frac{b-a}{2t} + 1$. The first result is then established.

The second result for $b-a \leq 2t$ follows from the fact that $\{a + \frac{b-a}{2}\}$ is a t-cover of $\mathcal{I}$ since $\rho(x, a+\frac{b-a}{2}) \leq \frac{b-a}{2} \leq t$ for any $x \in \mathcal{I}$. □

Definition A.3 ([39, Definition 5.16])**.** A collection of zero-mean random variables $\{S_\theta : \theta \in \Theta\}$ is a *sub-Gaussian* process with respect to a metric ρ on Θ if, for all $\theta_1, \theta_2 \in \Theta$ and $t \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left\{t(S_{\theta_1} - S_{\theta_2})\right\}\right] \leq \exp\left\{\frac{t^2\rho^2(\theta_1, \theta_2)}{2}\right\}.$$

Lemma 6 (Dudley's Entropy Integral Bound; modified from Theorem 5.22 of [39])**.** *Let $\{S_\theta : \theta \in \Theta\}$ be a zero-mean sub-Gaussian process with respect to a metric ρ on Θ. Then,*

$$\mathbb{E}\left(\sup_{\theta \in \Theta} S_\theta\right) \leq 8\sqrt{2}\int_0^\infty \sqrt{\log N(t; \Theta, \rho)}dt,$$

where $N(t; \Theta, \rho)$ represents the t-covering number of Θ with respect to ρ.

A.2 Proof of Proposition 2.1

Strong duality can be established using Theorem 1 in [18], which applies to general cases. However, by capitalizing the discrete nature of the sample space $\mathcal{Y}$, we can present a more concise result. To see this, here we provide an alternative proof of strong duality using the duality principle in finite-dimensional linear programming, as detailed below.

For every fixed $\mathbf{x} \in \mathcal{X}$, $\widetilde{\mathbf{y}} \in \mathcal{Y}^R$, $\psi \in \Psi$, by re-writing the the constraint $Q_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}} \in \Gamma_\epsilon(P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$ in (1) using Definition 2.1, we re-express the primal problem (1) as:

$$\overline{\mathsf{P}}_\epsilon \triangleq \sup_{\Pi \in \mathcal{P}(\mathcal{Y}^2)} \Big\{ \int \ell(\psi(\mathbf{x}), \mathrm{y}) d\Pi(\mathrm{y}, \mathrm{y}') : \int c^p(\mathrm{y}, \mathrm{y}') d\Pi(\mathrm{y}, \mathrm{y}') \leq \epsilon^p,\ \Pi(\mathcal{Y}, \cdot) = P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}(\cdot) \Big\}.$$

Given the discrete nature of the sample space $\mathcal{Y}$, $\overline{\mathsf{P}}_\epsilon$ can be reformulated as the following finite-dimensional linear program:

$$\overline{\mathsf{P}}_\epsilon = \begin{cases} \max\limits_{\pi_{jk} \in \mathbb{R} \text{ with } j,k \in [K]} \Big\{ \sum\limits_{j,k \in [K]} \ell(\psi(\mathbf{x}), j)\pi_{jk} \Big\}, \\ s.t. \ \sum\limits_{j,k \in [K]} c^p(j,k)\pi_{jk} \leq \epsilon^p, \\ \qquad \sum\limits_{j=1}^{K} \pi_{jk} = P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}(k) \text{ for } k \in [K], \\ \qquad \pi_{jk} \geq 0 \text{ for } j,k \in [K], \end{cases}$$

where $\pi_{jk} = \Pi(\mathrm{Y} = j, \mathrm{Y}' = k)$ for $j, k \in [K]$. By Lemma 2, each primal constraint corresponds to a dual variable. Introducing the dual variables γ and τ_k for $k \in [K]$, the dual linear program for $\overline{\mathsf{P}}_\epsilon$ is then expressed as

$$\overline{\mathsf{D}}_\epsilon = \begin{cases} \min\limits_{\gamma, \tau_k \in \mathbb{R} \text{ with } k \in [K]} \left\{ \gamma \epsilon^p + \sum\limits_{k=1}^{K} P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}(k) \tau_k \right\}, \\ s.t.\ \gamma c^p(j,k) + \tau_k \geq \ell(\psi(\mathbf{x}), j) \text{ for } j, k \in [K], \\ \qquad \gamma \geq 0, \end{cases} \tag{A4}$$

where first constraint can be written as $\tau_k \geq \max_{j \in [K]} \left\{ \ell(\psi(\mathbf{x}), j) - \gamma c^p(j,k) \right\}$. Therefore, to minimize the objective function in (A4), the value of the dual variable τ_k should be taken as $\max_{j \in [K]} \left\{ \ell(\psi(\mathbf{x}), j) - \gamma c^p(j,k) \right\}$ for each $\gamma \geq 0$ and $k \in [K]$. Hence, the proof is established.

A.3 Proof of Remark 2.3

As in the proof of Proposition 2.1, the optimization problem in (6) can be written as

$$\begin{aligned} \overline{\mathsf{P}}_\epsilon &\triangleq \sup \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}_{Q_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}} \left\{ \ell(\psi(\mathbf{x}_i), \mathrm{Y}) \right\} \right] : \frac{1}{n} \sum_{i=1}^{n} \mathscr{d}(Q_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}, P_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}) \leq \epsilon \right\} \\ &= \sup_{\Pi^{(i)} \in \mathcal{P}(\mathcal{Y}^2), i \in [n]} \left\{ \frac{1}{n} \sum_{i=1}^{n} \int \ell(\psi(\mathbf{x}_i), \mathrm{y}) d\Pi^{(i)}(\mathrm{y}, \mathrm{y}') : \frac{1}{n} \sum_{i=1}^{n} \int c(\mathrm{y}, \mathrm{y}') d\Pi^{(i)}(\mathrm{y}, \mathrm{y}') \leq \epsilon, \right. \\ &\qquad\qquad \left. \Pi^{(i)}(\mathcal{Y}, \cdot) = P_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}(\cdot) \text{ for } i \in [n] \right\}, \end{aligned}$$

where the first step is due to the definition of the empirical distribution $P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}$, and the second step holds by re-writing the constraint $\frac{1}{n} \sum_{i=1}^{n} \mathscr{d}(Q_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}, P_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}) \leq \epsilon$ using the definition of the the Wasserstein distance $\mathscr{d}(\cdot, \cdot)$ given in Definition 2.1.

$\overline{\mathsf{P}}_\epsilon$ can be further expressed as the finite- dimensional linear program:

$$\overline{\mathsf{P}}_\epsilon = \begin{cases} \max\limits_{\pi^{(i)}_{jk} \in \mathbb{R} \text{ with } i \in [n], j,k \in [K]} \left\{ \frac{1}{n} \sum\limits_{i=1}^{n} \sum\limits_{j,k \in [K]} \ell(\psi(\mathbf{x}_i), j) \pi^{(i)}_{jk} \right\}, \\ s.t.\ \frac{1}{n} \sum\limits_{i=1}^{n} \sum\limits_{j,k \in [K]} c(j,k) \pi^{(i)}_{jk} \leq \epsilon, \\ \sum\limits_{j=1}^{K} \pi^{(i)}_{jk} = P_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}(k) \text{ for } i \in [n] \text{ and } k \in [K], \\ \pi^{(i)}_{jk} \geq 0 \text{ for } i \in [n] \text{ and } j, k \in [K], \end{cases}$$

where $\pi^{(i)}_{jk} = \Pi^{(i)}(\mathrm{Y} = j, \mathrm{Y}' = k)$ for $i \in [n]$ and $j, k \in [K]$.

By Lemma 2 and introducing dual variables γ and $\tau^{(i)}_k$ with $i \in [n]$ and $k \in [K]$, the dual linear program for $\overline{\mathsf{P}}_\epsilon$ is expressed as

$$\begin{aligned} \overline{\mathsf{D}}_\epsilon &= \begin{cases} \min\limits_{\gamma, \tau^{(i)}_k \in \mathbb{R} \text{ for } i \in [n] \text{ and } k \in [K]} \left\{ \gamma \epsilon + \sum\limits_{i=1}^{n} \sum\limits_{k=1}^{K} P_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}(k) \tau^{(i)}_k \right\}, \\ s.t.\ \frac{1}{n} \gamma c(j,k) + \tau^{(i)}_k \geq \frac{1}{n} \ell(\psi(\mathbf{x}_i), j) \text{ for } i \in [n] \text{ and } j, k \in [K], \\ \gamma \geq 0, \end{cases} \\ &= \begin{cases} \min\limits_{\gamma, \tau^{(i)}_k \in \mathbb{R} \text{ for } i \in [n] \text{ and } k \in [K]} \left\{ \gamma \epsilon + \frac{1}{n} \sum\limits_{i=1}^{n} \sum\limits_{k=1}^{K} P_{\mathrm{y}|\mathbf{x}_i,\widetilde{\mathbf{y}}_i}(k) \widetilde{\tau}^{(i)}_k \right\}, \\ s.t.\ \gamma c(j,k) + \widetilde{\tau}^{(i)}_k \geq \ell(\psi(\mathbf{x}_i), j) \text{ for } i \in [n] \text{ and } j, k \in [K], \\ \gamma \geq 0, \end{cases} \end{aligned}$$

where we let $\widetilde{\tau}_k^{(i)} = n\tau_k^{(i)}$ for $i \in [n]$ and $k \in [K]$ in the second step. Thus, the proof is completed.

A.4 Proof of Theorem 2.2

We begin by demonstrating that, for various choices of the reference distribution, the optimal value for γ in the relaxed dual problem, as stated in (4), is constrained to a compact set. The proof techniques in [40] are used.

Specifically, for a given $\epsilon > 0$ and $\psi \in \Psi$, let

$$\gamma^* \in \arg\inf_{\gamma \geq 0} \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\bigg(\gamma\epsilon^p + \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}\Big[\sup_{y' \in \mathcal{Y}}\big\{\ell(\psi(\mathbf{X}), y') - \gamma c^p(y', \mathrm{Y})\Big]\big\}\bigg).$$

Noting that for any $\mathbf{x} \in \mathcal{X}$ and $\mathrm{y} \in \mathcal{Y}$, $\sup_{y' \in \mathcal{Y}}\big\{\ell(\psi(\mathbf{x}), y') - \ell(\psi(\mathbf{x}), \mathrm{y}) - \gamma^* c^p(y', \mathrm{y})\big\} \geq \big\{\ell(\psi(\mathbf{x}), y') - \ell(\psi(\mathbf{x}), \mathrm{y}) - \gamma^* c^p(y', \mathrm{y})\big\}\big|_{y'=\mathrm{y}} = 0$, we obtain that for any $\gamma \geq 0$,

$$\begin{aligned}
\gamma^*\epsilon^p \leq& \gamma^*\epsilon^p + \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\bigg(\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}\Big[\sup_{y' \in \mathcal{Y}}\big\{\ell(\psi(\mathbf{X}), y') - \ell(\psi(\mathbf{X}), \mathrm{Y}) - \gamma^* c^p(y', \mathrm{Y})\big\}\Big]\bigg) \\
\leq& \gamma\epsilon^p + \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\bigg(\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}\Big[\sup_{y' \in \mathcal{Y}}\big\{\ell(\psi(\mathbf{X}), y') - \ell(\psi(\mathbf{X}), \mathrm{Y}) - \gamma c^p(y', \mathrm{Y})\big\}\Big]\bigg) \\
\leq& \gamma\epsilon^p + \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\bigg(\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}\Big[\sup_{y' \in \mathcal{Y}}\big\{L \cdot c(y', \mathrm{Y}) - \gamma c^p(y', \mathrm{Y})\big\}\Big]\bigg) \\
\leq& \gamma\epsilon^p + \sup_{t \geq 0}\big\{Lt - \gamma t^p\big\},
\end{aligned} \tag{A5}$$

where the second inequality is due to the definition of γ^*; the third inequality comes from the Lipschitz property in the assumption; and the last inequality holds because $c(y', \mathrm{Y}) = \kappa\mathbf{1}(y' \neq \mathrm{Y})$ takes values in $\{0, \kappa\}$, leading to $\sup_{y' \in \mathcal{Y}}\big\{L \cdot c(y', \mathrm{Y}) - \gamma \cdot c^p(y', \mathrm{Y})\big\} = \sup_{t \in \{0,\kappa\}}\big\{Lt - \gamma t^p\big\} \leq \sup_{t \geq 0}\big\{Lt - \gamma t^p\big\}$, which is a constant.

Now, we show that

$$\gamma^* \leq L\epsilon^{-(p-1)} \triangleq M^*. \tag{A6}$$

Indeed, when $p = 1$, then taking $\gamma = L$ in (A5) shows (A6). When $p > 1$, then $Lt - \gamma t^p$ in (A5) takes its maximum value at $t^* = \{L/(p\gamma)\}^{1/(p-1)}$, and hence, (A5) yields that for any $\gamma \geq 0$,

$$\begin{aligned}
\gamma^*\epsilon^p \leq& \gamma\epsilon^p + L \cdot \{L/(p\gamma)\}^{1/(p-1)} - \gamma\{L/(p\gamma)\}^{p/(p-1)} \\
=& \gamma\epsilon^p + L^{p/(p-1)}\gamma^{-1/(p-1)}p^{-p/(p-1)}(p-1).
\end{aligned}$$

Therefore, taking $\gamma = L/(p\epsilon^{p-1})$ leads to $\gamma^*\epsilon^p \leq L\epsilon$, i.e., (A6) holds.

Next, let $\ell_{\gamma,\psi}(\mathbf{x}, \widetilde{\mathbf{y}}) \triangleq \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}\Big[\sup_{y' \in \mathcal{Y}}\{\ell(\psi(\mathbf{x}), y') - \gamma c^p(y', \mathrm{Y})\}\Big]$ for any $(\mathbf{x}, \widetilde{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}^R$. For every $\psi \in \Psi$, we have that

$$\begin{aligned}
&\Big|\mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \widehat{\mathfrak{R}}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})\Big| \\
=&\Big|\inf_{\gamma \geq 0} \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\bigg(\gamma\epsilon^p + \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}\Big[\sup_{y' \in \mathcal{Y}}\{\ell(\psi(\mathbf{X}), y') - \gamma c^p(y', \mathrm{Y})\}\Big]\bigg) \\
&\quad - \inf_{\gamma \geq 0} \mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}}\bigg(\gamma\epsilon^p + \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}}\Big[\sup_{y' \in \mathcal{Y}}\{\ell(\psi(\mathbf{X}), y') - \gamma c^p(y', \mathrm{Y})\}\Big]\bigg)\Big| \\
=&\Big|\inf_{0 \leq \gamma \leq M^*} \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\Big\{\gamma\epsilon^p + \ell_{\gamma,\psi}(\mathbf{X}, \widetilde{\mathbf{Y}})\Big\} - \inf_{0 \leq \gamma \leq M^*} \mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}}\Big\{\gamma\epsilon^p + \ell_{\gamma,\psi}(\mathbf{X}, \widetilde{\mathbf{Y}})\Big\}\Big| \\
\leq& \sup_{0 \leq \gamma \leq M^*}\Big|\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\Big\{\ell_{\gamma,\psi}(\mathbf{X}, \widetilde{\mathbf{Y}})\Big\} - \mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}}\Big\{\ell_{\gamma,\psi}(\mathbf{X}, \widetilde{\mathbf{Y}})\Big\}\Big| \\
\triangleq& \Phi(\mathcal{D})
\end{aligned} \tag{A7}$$

where the first equality comes from (4) and (5), the second equality follows from (A6) and the definition of $\ell_{\gamma,\psi}$, and the third step is due to the fact that $|\inf_{v \in A} f(v) - \inf_{v \in A} g(v)| \leq \sup_{v \in A}|f(v) - g(v)|$ for bounded functions $f, g : A \to \mathbb{R}$ [41, Proposition 2.18].

In (A7), we use $\mathcal{D}$ to stress the dependence of $\mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\}$ on the observed data of size n, as defined in Section 2.1, and let Φ represent the resulting function mapping from $\left(\mathcal{X}\times\mathcal{Y}^R\right)^n$ to $\mathbb{R}$, with $\Phi(\mathcal{D})$ being the value for data $\mathcal{D}$, where $\left(\mathcal{X}\times\mathcal{Y}^R\right)^n$ is the Cartesian product of multiplying $\mathcal{X}\times\mathcal{Y}^R$ n times. The function $\Phi:\left(\mathcal{X}\times\mathcal{Y}^R\right)^n\to\mathbb{R}$ defined in (A7) satisfies the bounded difference property (A3) with parameters $\left\{\frac{M}{n},\ldots,\frac{M}{n}\right\}$, where M represents the upper bound of the loss function ℓ in the assumption of Theorem 2.2. Indeed, for any $k\in[n]$,

$$\begin{aligned}
&\sup_{z_1,\ldots,z_n,z_k'\in\mathcal{X}\times\mathcal{Y}^R}\left|\Phi(z_1,\ldots,z_k,\ldots,z_n)-\Phi(z_1,\ldots,z_k',\ldots,z_n)\right|\\
=&\sup_{z_1,\ldots,z_n,z_k'\in\mathcal{X}\times\mathcal{Y}^R}\left|\sup_{0\le\gamma\le M^*}\left|\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\}-\frac{1}{n}\sum_{i=1}^{n}\ell_{\gamma,\psi}(z_i)\right|\right.\\
&\qquad\left.-\sup_{0\le\gamma\le M^*}\left|\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\}-\frac{1}{n}\sum_{i=1}^{n}\ell_{\gamma,\psi}(z_i)-\frac{1}{n}\ell_{\gamma,\psi}(z_k')+\frac{1}{n}\ell_{\gamma,\psi}(z_k)\right|\right|\\
\le&\sup_{z_1,\ldots,z_n,z_k'\in\mathcal{X}\times\mathcal{Y}^R,0\le\gamma\le M^*}\left|\left|\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\}-\frac{1}{n}\sum_{i=1}^{n}\ell_{\gamma,\psi}(z_i)\right|\right.\\
&\qquad\left.-\left|\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\}-\frac{1}{n}\sum_{i=1}^{n}\ell_{\gamma,\psi}(z_i)-\frac{1}{n}\ell_{\gamma,\psi}(z_k')+\frac{1}{n}\ell_{\gamma,\psi}(z_k)\right|\right|\\
\le&\sup_{z_1,\ldots,z_n,z_k'\in\mathcal{X}\times\mathcal{Y}^R,0\le\gamma\le M^*}\left|\frac{1}{n}\ell_{\gamma,\psi}(z_k')-\frac{1}{n}\ell_{\gamma,\psi}(z_k)\right|\\
\le&\frac{M}{n},
\end{aligned}$$

where the second step holds since $|\sup_{v\in A}f(v)-\sup_{v\in A}g(v)|\le\sup_{v\in A}|f(v)-g(v)|$ for bounded functions $f,g:A\to\mathbb{R}$ [41, Proposition 2.18], the third step is due to the triangle inequality for absolute values, and the last step holds since $\ell_{\gamma,\psi}\in[0,M]$ by definition.

Thus, by letting $t=M\sqrt{\frac{\log(1/\eta)}{2n}}$ in lemma 3, we have that, with probability at least $1-\eta$,

$$\Phi(\mathcal{D})\le\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\left\{\Phi(\mathcal{D})\right\}+M\sqrt{\frac{\log(1/\eta)}{2n}}.\tag{A8}$$

Similar to the derivations for (3.8)-(3.13) in the proof of Theorem 3.3 in [42], we obtain that

$$\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\left\{\Phi(\mathcal{D})\right\}\le2\mathbb{E}\left[\sup_{\gamma\in[0,M^*]}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell_{\gamma,\psi}(\mathbf{X}_i,\widetilde{\mathbf{Y}}_i)\right],\tag{A9}$$

where $\{\sigma_i\}_{i=1}^n$ are independent random variables chosen from $\{-1,+1\}$ with equal probability, and the expectation is taken with respect to all involved random variables.

Applying (A8) and (A9) to (A7) gives that with probability at least $1-\eta$,

$$\left|\mathfrak{R}_\epsilon(\psi;P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})-\widehat{\mathfrak{R}}_\epsilon(\psi;P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})\right|\le2\mathbb{E}\left[\sup_{\gamma\in[0,M^*]}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell_{\gamma,\psi}(\mathbf{X}_i,\widetilde{\mathbf{Y}}_i)\right]+M\sqrt{\frac{\log(1/\eta)}{2n}}.\tag{A10}$$

Now we identify an entropy based upper bound for the right-hand side of (A9) using the proof techniques in [40] and Example 5.24 of [39]. Specifically, for a given ψ, we define the random process $\left\{S_\gamma\triangleq\frac{1}{\sqrt{n}}\sum_{i=1}^n\sigma_i\ell_{\gamma,\psi}(\mathbf{X}_i,\widetilde{\mathbf{Y}}_i):\gamma\in[0,M^*]\right\}$. By using $\mathbb{E}(\sigma_i)=0$ and the independence between σ_i and $(\mathbf{X}_i,\widetilde{\mathbf{Y}}_i)$, we obtain that $\mathbb{E}(S_\gamma)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbb{E}(\sigma_i)\mathbb{E}\left\{\ell_{\gamma,\psi}(\mathbf{X}_i,\widetilde{\mathbf{Y}}_i)\right\}=0$. For any

$\gamma_1, \gamma_2 \in [0, M^*]$,

$$
\begin{aligned}
&\left|\ell_{\gamma_1,\psi}(\mathbf{x}, \widetilde{\mathbf{y}}) - \ell_{\gamma_2,\psi}(\mathbf{x}, \widetilde{\mathbf{y}})\right| \\
=&\left|\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}} \sup_{y' \in \mathcal{Y}} \{\ell(\psi(\mathbf{x}), y') - \gamma_1 c^p(y', \mathrm{Y})\} - \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}} \sup_{y' \in \mathcal{Y}} \{\ell(\psi(\mathbf{x}), y') - \gamma_2 c^p(y', \mathrm{Y})\}\right| \\
\leq&\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}} \left|\sup_{y' \in \mathcal{Y}} \{\ell(\psi(\mathbf{x}), y') - \gamma_1 c^p(y', \mathrm{Y})\} - \sup_{y' \in \mathcal{Y}} \{\ell(\psi(\mathbf{x}), y') - \gamma_2 c^p(y', \mathrm{Y})\}\right| \\
\leq&\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}} \sup_{y' \in \mathcal{Y}} \left|\gamma_1 c^p(y', \mathrm{Y}) - \gamma_2 c^p(y', \mathrm{Y})\right| \\
=&\kappa^p |\gamma_1 - \gamma_2|,
\end{aligned} \tag{A11}
$$

where the second inequality is due to the fact that $|\sup_A f - \sup_A g| \leq \sup_A |f - g|$ for bounded functions $f, g : A \to \mathbb{R}$ [41, Proposition 2.18], and the last step is due to the fact that $c(y', \mathrm{Y})$ can only take values in $\{0, \kappa\}$.

Hence, for $t \in \mathbb{R}$, we have that

$$
\begin{aligned}
\mathbb{E}\left\{e^{t(S_{\gamma_1} - S_{\gamma_2})}\right\} =&\mathbb{E}\left\{\exp\left[\frac{t}{\sqrt{n}} \sum_{i=1}^{n} \sigma_i \left\{\ell_{\gamma_1,\psi}(\mathbf{X}_i, \widetilde{\mathbf{Y}}_i) - \ell_{\gamma_2,\psi}(\mathbf{X}_i, \widetilde{\mathbf{Y}}_i)\right\}\right]\right\} \\
=&\left\{\mathbb{E}\left(\exp\left[\frac{t}{\sqrt{n}} \sigma_1 \left\{\ell_{\gamma_1,\psi}(\mathbf{X}_1, \widetilde{\mathbf{Y}}_1) - \ell_{\gamma_2,\psi}(\mathbf{X}_1, \widetilde{\mathbf{Y}}_1)\right\}\right]\right)\right\}^n \\
\leq&\exp\left\{\frac{t^2(\kappa^p |\gamma_1 - \gamma_2|)^2}{2}\right\},
\end{aligned} \tag{A12}
$$

where the inequality is due to Hoeffding's lemma and (A11). Thus, $\{S_\gamma : \gamma \in [0, M^*]\}$ is a zero-mean sub-Gaussian process with respect to metric ρ_γ, defined as $\rho_\gamma(\gamma_1, \gamma_2) = \kappa^p |\gamma_1 - \gamma_2|$ for any $\gamma_1, \gamma_2 \in [0, M^*]$.

Therefore, by Lemma 6, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\gamma \in [0, M^*]} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell_{\gamma,\psi}(\mathbf{X}_i, \widetilde{\mathbf{Y}}_i)\right] \\
=&\frac{1}{\sqrt{n}} \mathbb{E}\left(\sup_{\gamma \in [0, M^*]} S_\gamma\right) \\
\leq&\frac{8\sqrt{2}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(t; [0, M^*], \rho_\gamma)} dt \\
=&\frac{8\sqrt{2}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(t/\kappa^p; [0, M^*], |\cdot|)} dt \\
=&\frac{8\sqrt{2}\kappa^p}{\sqrt{n}} \int_0^\infty \sqrt{\log N(s; [0, M^*], |\cdot|)} ds \\
\leq&\frac{8\sqrt{2}\kappa^p}{\sqrt{n}} \int_0^{M^*/2} \sqrt{\log\left(\frac{M^*}{2s} + 1\right)} ds \\
\leq&\frac{8\sqrt{2}\kappa^p}{\sqrt{n}} \int_0^{M^*/2} \sqrt{\log\left(\frac{M^*}{s}\right)} ds \\
=&\frac{8\sqrt{2}}{\sqrt{n}} M^* \kappa^p \int_0^{1/2} \sqrt{\log(1/u)} du \\
=&\frac{4\sqrt{2}\{\sqrt{\log 2} + \sqrt{\pi}\mathrm{erfc}(\sqrt{\log 2})\}}{\sqrt{n}} M^* \kappa^p,
\end{aligned} \tag{A13}
$$

where the third step comes from the fact that, by the definition of ρ_γ, $\rho_\gamma(\gamma_1, \gamma_2) \leq t$ if and only if $|\gamma_1 - \gamma_2| \leq t/\kappa^p$; the fourth step holds by letting $s = t/\kappa^p$; the fifth step follows from Lemma 5; the sixth step comes from the fact that $1 < \frac{M^*}{2s} + 1 = \frac{M^*+2s}{2s} \leq \frac{M^*}{s}$ for $s \in [0, M^*/2]$; the penultimate step holds by letting $u = s/M^*$; and the last step arises from the fact that

$$\begin{aligned}&\int_0^{1/2}\sqrt{\log(1/u)}du = \int_0^{1/2}\sqrt{-\log u}du = \int_0^{1/2}\int_0^{-\log(u)}\frac{1}{2\sqrt{s}}dsdu\\ =&\int_0^{\log 2}\int_0^{1/2}\frac{1}{2\sqrt{s}}duds + \int_{\log 2}^{\infty}\int_0^{e^{-s}}\frac{1}{2\sqrt{s}}duds = \frac{\sqrt{\log 2}}{2} + \int_{\log 2}^{\infty}\frac{e^{-s}}{2\sqrt{s}}ds\\ =&\frac{\sqrt{\log 2}}{2} + \int_{\sqrt{\log 2}}^{\infty}\frac{e^{-w^2}}{2w}2wdw = \frac{\sqrt{\log 2}}{2} + \frac{\sqrt{\pi}}{2}\mathrm{erfc}(\sqrt{\log 2}),\end{aligned}$$

where $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty} e^{-w^2}dw$.

Applying (A13) and (A6) to (A10) gives that with probability at least $1-\eta$,

$$\begin{aligned}\left|\mathfrak{R}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \widehat{\mathfrak{R}}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})\right| \leq&\frac{8\sqrt{2}(\sqrt{\log 2} + \sqrt{\pi}\mathrm{erfc}(\sqrt{\log 2}))}{\sqrt{n}} \cdot \frac{L\kappa^p}{\epsilon^{p-1}} + M\sqrt{\frac{\log(1/\eta)}{2n}}\\ <&\frac{15L\kappa^p}{\epsilon^{p-1}\sqrt{n}} + M\sqrt{\frac{\log(1/\eta)}{2n}},\end{aligned}$$

where the last step is due to the fact that $8\sqrt{2}(\sqrt{\log 2} + \sqrt{\pi}\mathrm{erfc}(\sqrt{\log 2})) \approx 14.2 < 15$. Hence, the proof is completed.

A.5 Proof of Corollary 2.3

By the definition of $\inf_{\psi\in\Psi}\mathfrak{R}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})$, for any $\zeta > 0$, there exists $\psi_{\zeta} \in \Psi$ such that $\mathfrak{R}_{\epsilon}(\psi_{\zeta}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \leq \inf_{\psi\in\Psi}\mathfrak{R}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) + \zeta$. Therefore,

$$\begin{aligned}&\mathfrak{R}_{\epsilon}(\widehat{\psi}_{\epsilon,n}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \inf_{\psi\in\Psi}\mathfrak{R}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})\\ \leq&\mathfrak{R}_{\epsilon}(\widehat{\psi}_{\epsilon,n}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \mathfrak{R}_{\epsilon}(\psi_{\zeta}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) + \zeta\\ \leq&\mathfrak{R}_{\epsilon}(\widehat{\psi}_{\epsilon,n}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \widehat{\mathfrak{R}}_{\epsilon}(\widehat{\psi}_{\epsilon,n}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) + \widehat{\mathfrak{R}}_{\epsilon}(\psi_{\zeta}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \mathfrak{R}_{\epsilon}(\psi_{\zeta}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) + \zeta\\ \leq&2\sup_{\psi\in\Psi}\left|\mathfrak{R}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \widehat{\mathfrak{R}}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})\right| + \zeta.\end{aligned}$$

Since the inequality above is true for all $\zeta > 0$, we have that

$$\begin{aligned}&\mathfrak{R}_{\epsilon}(\widehat{\psi}_{\epsilon,n}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \inf_{\psi\in\Psi}\mathfrak{R}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})\\ \leq&2\sup_{\psi\in\Psi}\left|\mathfrak{R}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \widehat{\mathfrak{R}}_{\epsilon}(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}})\right|\\ \leq&2\sup_{\psi\in\Psi,0\leq\gamma\leq M^*}\left|\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\} - \mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\}\right|,\end{aligned}\tag{A14}$$

where the second inequality arises from (A7).

Similar to the proof of Theorem 2.2 in Appendix A.4, we can derive that

$$\begin{aligned}&\sup_{\psi\in\Psi,0\leq\gamma\leq M^*}\left|\mathbb{E}_{\mathbf{x},\widetilde{\mathbf{y}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\} - \mathbb{E}_{P^{(n)}_{\mathbf{x},\widetilde{\mathbf{y}}}}\left\{\ell_{\gamma,\psi}(\mathbf{X},\widetilde{\mathbf{Y}})\right\}\right|\\ \leq&2\mathbb{E}\left[\sup_{\gamma\in[0,M^*],\psi\in\Psi}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell_{\gamma,\psi}(\mathbf{X}_i,\widetilde{\mathbf{Y}}_i)\right] + M\sqrt{\frac{\log(1/\eta)}{2n}}.\end{aligned}\tag{A15}$$

For any $\gamma_1, \gamma_2 \in [0, M^*]$ and $\psi_1, \psi_2 \in \Psi$,

$$
\begin{aligned}
&\left|\ell_{\gamma_1,\psi_1}(\mathbf{x}, \widetilde{\mathbf{y}}) - \ell_{\gamma_2,\psi_2}(\mathbf{x}, \widetilde{\mathbf{y}})\right| \\
=&\left|\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}} \sup_{y' \in \mathcal{Y}} \left\{\ell(\psi_1(\mathbf{x}), y') - \gamma_1 c^p(y', \mathrm{Y})\right\} - \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}} \sup_{y' \in \mathcal{Y}} \left\{\ell(\psi_2(\mathbf{x}), y') - \gamma_2 c^p(y', \mathrm{Y})\right\}\right| \\
\leq&\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}} \left|\sup_{y' \in \mathcal{Y}} \left\{\ell(\psi_1(\mathbf{x}), y') - \gamma_1 c^p(y', \mathrm{Y})\right\} - \sup_{y' \in \mathcal{Y}} \left\{\ell(\psi_2(\mathbf{x}), y') - \gamma_2 c^p(y', \mathrm{Y})\right\}\right| \\
\leq&\mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}} \sup_{y' \in \mathcal{Y}} \left\{\left|\gamma_1 c^p(y', \mathrm{Y}) - \gamma_2 c^p(y', \mathrm{Y})\right| + \left|\ell(\psi_1(\mathbf{x}), y') - \ell(\psi_2(\mathbf{x}), y')\right|\right\} \\
\leq&\kappa^p |\gamma_1 - \gamma_2| + L' \|\psi_1 - \psi_2\|_\infty,
\end{aligned}
$$

where the first step is due to the definition of $\ell_{\gamma,\psi}$ defined after (A6), the second step is due to Jensen's inequality, and the last step is due to the Lipschitz property with respect to the cost function $c(\cdot, \cdot)$ defined in Theorem 2.2 in the assumption, and $\|\psi_1 - \psi_2\|_\infty \triangleq \sup_{\mathbf{x} \in \mathcal{X}} \|\psi_1(\mathbf{x}) - \psi_2(\mathbf{x})\|$ for some norm $\|\cdot\|$.

Allowing ψ to vary, we modify the discussion for the random process $\{S_\gamma : \gamma \in [0, M^*]\}$ in Appendix A.4, and consider the collection of random variables $\Big\{S_{\gamma,\psi} \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n \sigma_i \ell_{\gamma,\psi}(\mathbf{X}_i, \widetilde{\mathbf{Y}}_i) :$ $\gamma \in [0, M^*], \psi \in \Psi\Big\}$. Clearly, $\mathbb{E}(S_{\gamma,\psi}) = 0$. Modifying the metric $\rho_\gamma(\gamma_1, \gamma_2)$ in Appendix A.4, we define the metric $\rho_{\gamma,\psi}((\gamma_1, \psi_1), (\gamma_2, \psi_2)) \triangleq \kappa^p |\gamma_1 - \gamma_2| + L' \|\psi_1 - \psi_2\|_\infty$ for any $\gamma_1, \gamma_2 \in [0, M^*]$ and $\psi_1, \psi_2 \in \Psi$. Similar to deriving (A12), we obtain that for $t \in \mathbb{R}$,

$$
\mathbb{E}\left\{e^{t\left(S_{\gamma_1,\psi_1} - S_{\gamma_2,\psi_2}\right)}\right\} \leq \exp\left[\frac{t^2 \left\{\rho_{\gamma,\psi}((\gamma_1, \psi_1), (\gamma_2, \psi_2))\right\}^2}{2}\right].
$$

Thus, $\Big\{S_{\gamma,\psi} : \gamma \in [0, M^*], \psi \in \Psi\Big\}$ is a zero-mean sub-Gaussian process with respect to metric $\rho_{\gamma,\psi}$.

Let the Cartesian product $[0, M^*] \times \Psi$ denote the "parameter" space of (γ, ψ). Then, by Lemma 6, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\gamma \in [0, M^*], \psi \in \Psi} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell_{\gamma,\psi}(\mathbf{X}_i, \widetilde{\mathbf{Y}}_i)\right] \\
=&\frac{1}{\sqrt{n}} \mathbb{E}\left(\sup_{\gamma \in [0, M^*], \psi \in \Psi} S_{\gamma,\psi}\right) \\
\leq&\frac{8\sqrt{2}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(t; [0, M^*] \times \Psi, \rho_{\gamma,\psi})} dt \\
\leq&\frac{8\sqrt{2}}{\sqrt{n}} \int_0^\infty \sqrt{\log \left\{N(t/(2\kappa^p); [0, M^*], |\cdot|) \times N(t/(2L'); \Psi, \|\cdot\|_\infty)\right\}} dt \\
\leq&\frac{8\sqrt{2}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(t/(2\kappa^p); [0, M^*], |\cdot|)} dt + \frac{8\sqrt{2}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(t/(2L'); \Psi, \|\cdot\|_\infty)} dt \\
\leq&\frac{16\sqrt{2}\kappa^p}{\sqrt{n}} \int_0^\infty \sqrt{\log N(s; [0, M^*], |\cdot|)} ds + \frac{16\sqrt{2}L'}{\sqrt{n}} \int_0^\infty \sqrt{\log N(s; \Psi, \|\cdot\|_\infty)} ds \\
\leq&\frac{8\sqrt{2}(\sqrt{\log 2} + \sqrt{\pi}\mathrm{erfc}(\sqrt{\log 2}))}{\sqrt{n}} \cdot M^* \kappa^p + \frac{16\sqrt{2}L'}{\sqrt{n}} \int_0^\infty \sqrt{\log N(s; \Psi, \|\cdot\|_\infty)} ds, \qquad \text{(A16)}
\end{aligned}
$$

Here, for any set Ω and metric ρ on Ω, $N(t; \Omega, \rho)$ denotes the t-covering number for $t > 0$ as defined in Definition A.2. In the derivation of (A16), the third step is due to Lemma 4; the fourth step is due to $\sqrt{\log(ab)} \leq \sqrt{\log a} + \sqrt{\log b}$ for $a \geq 1$ and $b \geq 1$; the fifth step results from a change of variable; and the last line can be similarly proved as (A13).

By (A14)-(A16), we have that, with probability at least $1-\eta$,

$$
\begin{aligned}
&\mathfrak{R}_\epsilon(\widehat{\psi}_{\epsilon,n}; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) - \inf_{\psi\in\Psi} \mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\widetilde{\mathbf{y}}}) \\
\leq &\left\{ 57\frac{L\kappa^p}{\epsilon^{p-1}} + 91L' \int_0^\infty \sqrt{\log N(s;\Psi,\|\cdot\|_\infty)}ds \right\} \cdot \frac{1}{\sqrt{n}} + 2M\sqrt{\frac{\log(1/\eta)}{2n}}.
\end{aligned}
$$

Therefore, the proof is completed.

A.6 Proof of Theorem 3.1

For ease of presentation, in this proof we omit the dependence on $\mathbf{x}$ and $\widetilde{\mathbf{y}}$ in the notation. In particular, we let $P_0 \triangleq P_0(\mathbf{x},\widetilde{\mathbf{y}})$, $P_1 \triangleq P_1(\mathbf{x},\widetilde{\mathbf{y}})$, and $\psi \triangleq \psi(\mathbf{x})$. Let the objective function in (8) be denoted as

$$
g(\gamma;\psi) \triangleq \gamma\epsilon^p + P_0 \max\{\mathcal{T}(1-\psi), \mathcal{T}(\psi) - \gamma\kappa^p\} + P_1 \max\{\mathcal{T}(1-\psi) - \gamma\kappa^p, \mathcal{T}(\psi)\}. \quad \text{(A17)}
$$

To complete the proof, we begin by investigating the *inner optimization problem* in (8) by finding the optimal value of γ, γ_ψ^*, as defined in Section 3.2, that minimizes $g(\gamma;\psi)$ for each given ψ, and then address the *outer optimization problem* in (8) by finding the optimal value $\psi^\star$ that minimizes $g(\gamma_\psi^*;\psi)$. To this end, we eliminate the max operators in $g(\gamma;\psi)$ based on the values of ψ, and use the assumption that $\mathcal{T}$ is a decreasing function in (7). As ψ takes its value in $[0,1]$, we re-write the optimization problem (8) as

$$
\begin{aligned}
&\inf_{\psi\in[0,\frac{1}{2}]\cup[\frac{1}{2},1]} \inf_{\gamma\geq 0} g(\gamma;\psi) \\
=&\min\left\{ \min_{\psi\in[0,\frac{1}{2}]} \inf_{\gamma\geq 0} g(\gamma;\psi), \min_{\psi\in[\frac{1}{2},1]} \inf_{\gamma\geq 0} g(\gamma;\psi) \right\}, \\
\triangleq&\min\left\{ g(\gamma^*_{\psi_1^*};\psi_1^*), g(\gamma^*_{\psi_2^*};\psi_2^*) \right\}, \qquad \text{(A18)}
\end{aligned}
$$

where $(\gamma^*_{\psi_1^*},\psi_1^*)$ and $(\gamma^*_{\psi_2^*},\psi_2^*)$ are the arguments of $\min_{\psi_1\in[0,\frac{1}{2}]}\inf_{\gamma\geq 0} g(\gamma;\psi_1)$ and $\min_{\psi_2\in[\frac{1}{2},1]}\inf_{\gamma\geq 0} g(\gamma;\psi_2)$, respectively. We complete the proof by considering the following two cases.

Case 1: $\psi_1 \in [0,\frac{1}{2}]$.

In this case, $\mathcal{T}(\psi_1) \geq \mathcal{T}(\frac{1}{2}) \geq \mathcal{T}(1-\psi_1) \geq \mathcal{T}(1-\psi_1) - \gamma\kappa^p$. Let

$$
\gamma_0 = \frac{\mathcal{T}(\psi_1) - \mathcal{T}(1-\psi_1)}{\kappa^p}.
$$

Consequently, we have that

$$
g(\gamma;\psi_1) = \begin{cases} \gamma\epsilon^p + P_0\{\mathcal{T}(\psi_1) - \gamma\kappa^p\} + P_1\mathcal{T}(\psi_1) = \mathcal{T}(\psi_1) + \gamma\kappa^p(\varrho(\epsilon) - P_0), \text{ if } \gamma \leq \gamma_0; \\ \gamma\epsilon^p + P_0\mathcal{T}(1-\psi_1) + P_1\mathcal{T}(\psi_1), \text{ if } \gamma > \gamma_0. \end{cases} \quad \text{(A19)}
$$

For given ψ_1,

$$
\lim_{\gamma\to\gamma_0^+} g(\gamma;\psi_1) = \lim_{\gamma\to\gamma_0^-} g(\gamma;\psi_1) = \varrho(\epsilon)\left\{\mathcal{T}(\psi_1) - \mathcal{T}(1-\psi_1)\right\} + P_0\mathcal{T}(1-\psi_1) + P_1\mathcal{T}(\psi_1),
$$

showing that $g(\gamma;\psi_1)$ is continuous at γ_0. Therefore, for any given ψ_1, $g(\gamma;\psi_1)$ is continuous in γ over $\mathbb{R}^+$.

Then, for any given $\psi_1 \in [0,\frac{1}{2}]$, corresponding to the first term in (A18), we obtain that

$$
\begin{aligned}
\inf_{\gamma\geq 0} g(\gamma;\psi_1) &= \min\left\{ \inf_{\gamma>\gamma_0} g(\gamma;\psi_1), \inf_{\gamma\in[0,\gamma_0]} g(\gamma;\psi_1) \right\} \\
&= \min\left\{ g(\gamma_0;\psi_1), \min_{\gamma\in[0,\gamma_0]} g(\gamma;\psi_1) \right\} \\
&= \min_{\gamma\in[0,\gamma_0]} g(\gamma;\psi_1) \\
&\triangleq g(\gamma^*_{\psi_1};\psi_1), \qquad \text{(A20)}
\end{aligned}
$$

where we use the continuity of $g(\gamma;\psi_1)$ in γ, the fact that $g(\gamma;\psi_1)$ is increasing in γ when $\gamma > \gamma_0$, and the fact that a continuous function attains its infimum within any closed and bounded set in $\mathbb{R}$. Here,

$$\gamma^*_{\psi_1} \triangleq \arg\min_{\gamma\in[0,\gamma_0]} g(\gamma;\psi_1) \tag{A21}$$

for any $\psi_1 \in [0, \frac{1}{2}]$.

We complete the proof by the following two steps to examine the range of P_0.

Step 1: If $P_0 < \varrho(\epsilon)$, then, by (A19), for any given $\psi_1 \in [0, \frac{1}{2}]$, $g(\gamma;\psi_1)$ is increasing in γ over $[0,\gamma_0]$, showing that the optimal value in (A21) is $\gamma^*_{\psi_1} = 0$. Furthermore, because $g(\gamma^*_{\psi_1};\psi_1) = \mathcal{T}(\psi_1)$ for any $\psi_1 \in [0, \frac{1}{2}]$, we obtain that

$$\begin{aligned}
&\min_{\psi_1\in[0,\frac{1}{2}]} \inf_{\gamma\geq 0} g(\gamma;\psi)\\
=&\min_{\psi_1\in[0,\frac{1}{2}]} g(\gamma^*_{\psi_1};\psi_1)\\
=&\min_{\psi_1\in[0,\frac{1}{2}]} \mathcal{T}(\psi_1)\\
=&\mathcal{T}(1/2).
\end{aligned}$$

Consequently, $(0, \frac{1}{2})$ minimizes $g(\gamma;\psi)$ over $[0,\gamma_0]\times[0,\frac{1}{2}]$, i.e., $\psi_1^* = \frac{1}{2}$ and $\gamma^*_{\psi_1^*} = 0$.

Step 2: If $P_0 \geq \varrho(\epsilon)$, then, by (A19), $g(\gamma;\psi_1)$ is decreasing in γ when $\gamma \leq \gamma_0$, showing that the optimal value in (A21) is

$$\gamma^*_{\psi_1} = \gamma_0, \text{ with } \gamma_0 = \frac{\mathcal{T}(\psi_1) - \mathcal{T}(1-\psi_1)}{\kappa^p}, \tag{A22}$$

and thus,

$$g(\gamma^*_{\psi_1};\psi_1) = (P_1 + \varrho(\epsilon))\mathcal{T}(\psi_1) + (P_0 - \varrho(\epsilon))\mathcal{T}(1-\psi_1) \text{ for any } \psi_1 \in [0, 1/2].$$

Consequently, the derivative of $g(\gamma^*_{\psi_1};\psi_1)$ with respect to ψ_1 is

$$g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1) = (P_1 + \varrho(\epsilon))\mathcal{T}'(\psi_1) - (P_0 - \varrho(\epsilon))\mathcal{T}'(1-\psi_1),$$

leading to $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)\big|_{\psi_1=0} = (1-P_0+\varrho(\epsilon))\mathcal{T}'(0)-(P_0-\varrho(\epsilon))\mathcal{T}'(1)$ and $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)\big|_{\psi_1=1/2} = (1+2\varrho(\epsilon)-2P_0)\mathcal{T}'(\frac{1}{2})$. Solving

$$g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)\big|_{\psi_1=0} = 0 \text{ and } g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)\big|_{\psi_1=1/2} = 0$$

for P_0 leads to solutions

$$P_0^{(1)} \triangleq \varrho(\epsilon) + \frac{\mathcal{T}'(0)}{\mathcal{T}'(1)+\mathcal{T}'(0)} \text{ and } P_0^{(2)} \triangleq \varrho(\epsilon) + \frac{1}{2},$$

respectively.

Next, we identify ψ_1^* and $\gamma^*_{\psi_1^*}$ by examining P_0 relative to $P_0^{(1)}$ and $P_0^{(2)}$, in combination with the convexity or concavity of function $\mathcal{T}$ by the following two steps.

Step 2.1: Assume $\mathcal{T}$ is concave. Then by twice differentiability of $\mathcal{T}$, $\mathcal{T}''(\psi_1) \leq 0$ for $\psi_1 \in [0,1]$, leading to $\mathcal{T}'(1) \leq \mathcal{T}'(0) < 0$, and hence $P_0^{(1)} \leq P_0^{(2)}$. Additionally, $g''_{\psi_1}(\gamma^*_{\psi_1};\psi_1) = (P_1 + \varrho(\epsilon))\mathcal{T}''(\psi_1) + (P_0 - \varrho(\epsilon))\mathcal{T}''(1-\psi_1) \leq 0$, and thus, $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)$ is non-increasing in ψ_1 for $\psi_1 \in [0, \frac{1}{2}]$.

- If $\varrho(\epsilon) \leq P_0 \leq P_0^{(1)}$, then $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1) \leq 0$ for $\psi_1 \in [0, \frac{1}{2}]$, and thus, $g(\gamma^*_{\psi_1};\psi_1)$ is non-increasing in ψ_1 for $\psi_1 \in [0, \frac{1}{2}]$. Therefore, $\inf_{0\leq\psi_1\leq 1/2} g(\gamma^*_{\psi_1};\psi_1) = g(\gamma^*_{1/2};\frac{1}{2}) = \mathcal{T}(\frac{1}{2})$. Thus, $\psi_1^* = \frac{1}{2}$ and $\gamma^*_{\psi_1^*} = 0$ by (A22).

- If $P_0 \geq P_0^{(2)}$, then $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1) \geq 0$, showing that $g(\gamma^*_{\psi_1};\psi_1)$ is non-decreasing in ψ_1 for $\psi_1 \in [0,\frac{1}{2}]$. Therefore, $\inf_{0\leq\psi_1\leq 1/2} g(\gamma^*_{\psi_1};\psi_1) = g(\gamma^*_0;0) = (P_1+\varrho(\epsilon))\mathcal{T}(0) + (P_0-\varrho(\epsilon))\mathcal{T}(1)$. Thus, $\psi_1^* = 0$ and $\gamma^*_{\psi_1^*} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$ by (A22).
- If $P_0^{(1)} < P_0 < P_0^{(2)}$, then $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)$ is non-increasing in ψ_1 for $\psi_1 \in [0,\frac{1}{2}]$ with $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)\big|_{\psi_1=0} > 0$ and $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)\big|_{\psi_1=1/2} < 0$. Therefore, $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1) = 0$ has a unique solution on $[0,\frac{1}{2}]$, denoted $\psi_1^\diamond$, and furthermore, $g(\gamma^*_{\psi_1};\psi_1)$ is increasing in ψ_1 for $\psi_1 \in [0,\psi_1^\diamond]$ and decreasing on $[\psi_1^\diamond,\frac{1}{2}]$. Therefore, the infimum of $g_{\psi_1}(\gamma^*_{\psi_1};\psi_1)$ on $\psi_1 \in [0,\frac{1}{2}]$ is taken at $\psi_1 = 0$ or $\psi_1 = \frac{1}{2}$. If $g(\gamma^*_{1/2};\frac{1}{2}) \leq g(\gamma^*_0;0)$, i.e., $P_0 \leq \varrho(\epsilon) + \frac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}$, the optimal value for ψ_1 in Case 1 is $\psi_1^* = \frac{1}{2}$ with $\gamma^*_{\psi_1^*} = 0$; otherwise, the optimal value is $\psi_1^* = 0$ with $\gamma^*_{\psi_1^*} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$ by (A22).

Summarizing the discussion in Step 2.1, we obtain that when $\psi_1 \in [0,\frac{1}{2}]$ and $\mathcal{T}$ is concave,

(i) if $P_0 > \varrho(\epsilon) + \frac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}$, $(\psi_1^*,\gamma^*_{\psi_1^*})$ in (A18) is given by $\psi_1^* = 0$ and $\gamma^*_{\psi_1^*} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$, yielding $g(\gamma^*_{\psi_1^*};\psi_1^*) = (P_1+\varrho(\epsilon))\mathcal{T}(0) + (P_0-\varrho(\epsilon))\mathcal{T}(1)$;

(ii) otherwise, $\psi_1^* = \frac{1}{2}$ and $\gamma^*_{\psi_1^*} = 0$, yielding $g(\gamma^*_{\psi_1^*};\psi_1^*) = \mathcal{T}(\frac{1}{2})$.

Step 2.2: Assume $\mathcal{T}$ is convex. Then by twice differentiability of $\mathcal{T}$, $\mathcal{T}''(\psi_1) \geq 0$ for $\psi_1 \in [0,1]$, leading to $\mathcal{T}'(0) \leq \mathcal{T}'(1) < 0$, and hence $P_0^{(2)} \leq P_0^{(1)}$. Additionally, $g''_{\psi_1}(\gamma^*_{\psi_1};\psi_1) = (P_1+\varrho(\epsilon))\mathcal{T}''(\psi_1) + (P_0-\varrho(\epsilon))\mathcal{T}''(1-\psi_1) \geq 0$, and thus, $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)$ is non-decreasing in ψ_1 for $\psi_1 \in [0,\frac{1}{2}]$.

- If $P_0 \geq P_0^{(1)}$, then $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1) \geq 0$ for $\psi_1 \in [0,\frac{1}{2}]$, and thus, $g(\gamma^*_{\psi_1};\psi_1)$ is non-decreasing in ψ_1 for $\psi_1 \in [0,\frac{1}{2}]$. Therefore, $\inf_{0\leq\psi_1\leq 1/2} g(\gamma^*_{\psi_1};\psi_1) = g(\gamma^*_0;0) = (P_1+\varrho(\epsilon))\mathcal{T}(0) + (P_0-\varrho(\epsilon))\mathcal{T}(1)$. Thus, $\psi_1^* = 0$ and $\gamma^*_{\psi_1^*} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$ by (A22).
- If $\varrho(\epsilon) \leq P_0 \leq P_0^{(2)}$, then $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1) \leq 0$ for $\psi_1 \in [0,\frac{1}{2}]$, and thus, $g(\gamma^*_{\psi_1};\psi_1)$ is non-increasing in ψ_1 for $\psi_1 \in [0,\frac{1}{2}]$. Therefore, $\inf_{0\leq\psi_1\leq\frac{1}{2}} g(\gamma^*_{\psi_1};\psi_1) = g(\gamma^*_{1/2};\frac{1}{2}) = \mathcal{T}(\frac{1}{2})$. Thus, $\psi_1^* = \frac{1}{2}$ and $\gamma^*_{\psi_1^*} = 0$ by (A22).
- If $P_0^{(2)} < P_0 < P_0^{(1)}$, then $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)$ is non-decreasing in ψ_1 for $\psi_1 \in [0,\frac{1}{2}]$ with $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)\big|_{\psi_1=0} < 0$ and $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1)\big|_{\psi_1=1/2} > 0$. Therefore, $g'_{\psi_1}(\gamma^*_{\psi_1};\psi_1) = 0$ has a unique solution on $[0,\frac{1}{2}]$, denoted $\psi_1^\diamond$, and furthermore, $g(\gamma^*_{\psi_1};\psi_1)$ is decreasing in ψ_1 for $\psi_1 \in [0,\psi_1^\diamond]$ and increasing on $[\psi_1^\diamond, 1/2]$. Then, the infimum of $g_{\psi_1}(\gamma^*_{\psi_1};\psi_1)$ on $\psi_1 \in [0,\frac{1}{2}]$ is taken at $\psi_1 = \psi_1^\diamond$, that is, $\inf_{0\leq\psi_1\leq 1/2} g(\gamma^*_{\psi_1};\psi_1) = g(\gamma^*_{\psi_1^\diamond};\psi_1^\diamond) = (P_1+\varrho(\epsilon))\mathcal{T}(\psi_1^\diamond) + (P_0-\varrho(\epsilon))\mathcal{T}(1-\psi_1^\diamond)$. Thus, $\psi_1^* = \psi_1^\diamond$ with $\gamma^*_{\psi_1^*} = \frac{\mathcal{T}(\psi_1^\diamond)-\mathcal{T}(1-\psi_1^\diamond)}{\kappa^p}$ by (A22).

Case 2: $\psi_2 \in [\frac{1}{2},1]$**.**

In this case, we set $\overline{\psi}_2 \triangleq 1-\psi_2$, yielding $\overline{\psi}_2 \in [0,\frac{1}{2}]$, and the objective function $g(\gamma;\psi_2)$ defined in (A17) can be written as

$$\begin{aligned} g(\gamma;\psi_2) =& \gamma\epsilon^p + P_0\max\{\mathcal{T}(1-\psi_2),\mathcal{T}(\psi_2)-\gamma\kappa^p\} + P_1\max\{\mathcal{T}(1-\psi_2)-\gamma\kappa^p,\mathcal{T}(\psi_2)\} \\ =& \gamma\epsilon^p + P_1\max\{\mathcal{T}(1-\overline{\psi}_2),\mathcal{T}(\overline{\psi}_2)-\gamma\kappa^p\} + P_0\max\{\mathcal{T}(1-\overline{\psi}_2)-\gamma\kappa^p,\mathcal{T}(\overline{\psi}_2)\}. \end{aligned}$$

Hence, the derivation in Case 1 for any ψ_1 in $[0,\frac{1}{2}]$ can be applied to $\overline{\psi}_2$ by modifying the derivations based on the range of P_0 to be that for P_1, as outlined below.

- **Step 1:** If $P_1 < \varrho(\epsilon)$, then following the results for ψ_1^* and $\gamma^*_{\psi_1^*}$ in Case 1, with only ψ_1^*, P_1, and P_0 there replaced by $\overline{\psi}_2^*$, P_0, and P_1, respectively, we obtain that $\overline{\psi}_2^* = \frac{1}{2}$ and $\gamma^*_{\overline{\psi}_2^*} = 0$. Hence, ψ_2^* is taken as $1-\overline{\psi}_2^* = \frac{1}{2}$ and $\gamma^*_{\psi_2^*} = 0$, yielding $g(\gamma^*_{\psi_2^*};\psi_2^*) = g(0;\frac{1}{2}) = \mathcal{T}(\frac{1}{2})$.

- **Step 2:** If $P_1 \geq \varrho(\epsilon)$, then, by (A22), $\gamma^*_{\psi^*_2}$ is set as $\gamma^*_{\psi^*_2} = \frac{\mathcal{T}(\overline{\psi}^*_2) - \mathcal{T}(1-\overline{\psi}^*_2)}{\kappa^p} = \frac{\mathcal{T}(1-\psi^*_2) - \mathcal{T}(\psi^*_2)}{\kappa^p}$.
- ***Step 2.1:*** Assume $\mathcal{T}$ is concave. We can directly derive the following result from the summary in Step 2.1 of Case 1.
  - If $P_1 \leq \varrho(\epsilon) + \frac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}$, then $\overline{\psi}^*_2 = \frac{1}{2}$ and $\gamma^*_{\psi^*_2} = 0$. Hence, $\psi^*_2 = 1 - \overline{\psi}^*_2 = \frac{1}{2}$, and $g(\gamma^*_{\psi^*_2}; \psi^*_2) = g(0; \frac{1}{2}) = \mathcal{T}(\frac{1}{2})$.
  - If $P_1 > \varrho(\epsilon) + \frac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}$, then $\overline{\psi}^*_2 = 0$ and $\gamma^*_{\psi^*_2} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$. Hence, $\psi^*_2 = 1 - \overline{\psi}^*_2 = 1$, and $g(\gamma^*_{\psi^*_2}; \psi^*_2) = g(\frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}; 1) = (P_0 + \varrho(\epsilon))\mathcal{T}(0) + (P_1 - \varrho(\epsilon))\mathcal{T}(1)$.
- ***Step 2.2:*** Assume $\mathcal{T}$ is convex. From the results on ψ^*_1 and $\gamma^*_{\psi^*_1}$ in Step 2.2 of Case 1, we obtain the following conclusion.
  - If $P_1 \geq \varrho(\epsilon) + \frac{\mathcal{T}'(0)}{\mathcal{T}'(1)+\mathcal{T}'(0)}$, then $\overline{\psi}^*_2 = 0$ and $\gamma^*_{\psi^*_2} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$. Hence, $\psi^*_2 = 1 - \overline{\psi}^*_2 = 1$, and $g(\gamma^*_{\psi^*_2}; \psi^*_2) = g(\frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}; 1) = (P_0 + \varrho(\epsilon))\mathcal{T}(0) + (P_1 - \varrho(\epsilon))\mathcal{T}(1)$.
  - If $\varrho(\epsilon) \leq P_1 \leq \varrho(\epsilon) + \frac{1}{2}$, then $\overline{\psi}^*_2 = \frac{1}{2}$ and $\gamma^*_{\psi^*_2} = 0$. Hence, $\psi^*_2 = 1 - \overline{\psi}^*_2 = \frac{1}{2}$, and $g(\gamma^*_{\psi^*_2}; \psi^*_2) = g(0; \frac{1}{2}) = \mathcal{T}(\frac{1}{2})$.
  - If $\varrho(\epsilon) + \frac{1}{2} < P_1 < \varrho(\epsilon) + \frac{\mathcal{T}'(0)}{\mathcal{T}'(1)+\mathcal{T}'(0)}$, then $\overline{\psi}^*_2 = \overline{\psi}^\diamond_2$ and $\gamma^*_{\psi^*_2} = \frac{\mathcal{T}(\overline{\psi}^\diamond_2) - \mathcal{T}(1-\overline{\psi}^\diamond_2)}{\kappa^p}$, where $\overline{\psi}^\diamond_2$ is the unique solution to $(P_0 + \varrho(\epsilon))\mathcal{T}'(\overline{\psi}_2) - (P_1 - \varrho(\epsilon))\mathcal{T}'(1 - \overline{\psi}_2) = 0$ on $[0, \frac{1}{2}]$. Hence, $\psi^*_2 = \psi^\diamond_2$ and $\gamma^*_{\psi^*_2} = \frac{\mathcal{T}(1-\psi^\diamond_2) - \mathcal{T}(\psi^\diamond_2)}{\kappa^p}$, where $\psi^\diamond_2 = 1 - \overline{\psi}^\diamond_2$ is the unique solution to $-(P_0 + \varrho(\epsilon))\mathcal{T}'(1 - \psi_2) + (P_1 - \varrho(\epsilon))\mathcal{T}'(\psi_2) = 0$ on $[\frac{1}{2}, 1]$. Then $g(\gamma^*_{\psi^*_2}; \psi^*_2) = (P_0 + \varrho(\epsilon))\mathcal{T}(1 - \psi^\diamond_2) + (P_1 - \varrho(\epsilon))\mathcal{T}(\psi^\diamond_2)$.

In summary, we present the derived results in Tables 3 and 4 for the scenarios where $\mathcal{T}$ is concave and convex, respectively.

		ψ^*_j	$\gamma^*_{\psi^*_j}$	robust risk $g(\gamma^*_{\psi^*_j}; \psi^*_j)$
Case 1	$P_0 \geq \varrho(\epsilon) + \frac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}$	$\psi^*_1 = 0$	$\gamma^*_{\psi^*_1} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$	$(P_1 + \varrho(\epsilon))\mathcal{T}(0) + (P_0 - \varrho(\epsilon))\mathcal{T}(1)$
	$P_0 < \varrho(\epsilon) + \frac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}$	$\psi^*_1 = \frac{1}{2}$	$\gamma^*_{\psi^*_1} = 0$	$\mathcal{T}(\frac{1}{2})$
Case 2	$P_1 \geq \varrho(\epsilon) + \frac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}$	$\psi^*_2 = 1$	$\gamma^*_{\psi^*_2} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$	$(P_0 + \varrho(\epsilon))\mathcal{T}(0) + (P_1 - \varrho(\epsilon))\mathcal{T}(1)$
	$P_1 < \varrho(\epsilon) + \frac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}$	$\psi^*_2 = \frac{1}{2}$	$\gamma^*_{\psi^*_2} = 0$	$\mathcal{T}(\frac{1}{2})$

Table 3: Summarized results in two cases when $\mathcal{T}$ is concave.

		ψ^*_j	$\gamma^*_{\psi^*_j}$	robust risk $g(\gamma^*_{\psi^*_j}; \psi^*_j)$
Case 1	$P_0 \geq \varrho(\epsilon) + \frac{\mathcal{T}'(0)}{\mathcal{T}'(0)+\mathcal{T}'(1)}$	$\psi^*_1 = 0$	$\gamma^*_{\psi^*_1} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$	$(P_1 + \varrho(\epsilon))\mathcal{T}(0) + (P_0 - \varrho(\epsilon))\mathcal{T}(1)$
	$P_0 \leq \varrho(\epsilon) + \frac{1}{2}$	$\psi^*_1 = \frac{1}{2}$	$\gamma^*_{\psi^*_1} = 0$	$\mathcal{T}(\frac{1}{2})$
	$\varrho(\epsilon) + \frac{1}{2} < P_0 < \varrho(\epsilon) + \frac{\mathcal{T}'(0)}{\mathcal{T}'(0)+\mathcal{T}'(1)}$	$\psi^*_1 = \psi^\diamond_1$	$\gamma^*_{\psi^*_1} = \frac{\mathcal{T}(\psi^\diamond_1)-\mathcal{T}(1-\psi^\diamond_1)}{\kappa^p}$	$(P_1 + \varrho(\epsilon))\mathcal{T}(\psi^\diamond_1) + (P_0 - \varrho(\epsilon))\mathcal{T}(1 - \psi^\diamond_1)$
Case 2	$P_1 \geq \varrho(\epsilon) + \frac{\mathcal{T}'(0)}{\mathcal{T}'(0)+\mathcal{T}'(1)}$	$\psi^*_2 = 1$	$\gamma^*_{\psi^*_2} = \frac{\mathcal{T}(0)-\mathcal{T}(1)}{\kappa^p}$	$(P_0 + \varrho(\epsilon))\mathcal{T}(0) + (P_1 - \varrho(\epsilon))\mathcal{T}(1)$
	$P_1 \leq \varrho(\epsilon) + \frac{1}{2}$	$\psi^*_2 = \frac{1}{2}$	$\gamma^*_{\psi^*_2} = 0$	$\mathcal{T}(\frac{1}{2})$
	$\varrho(\epsilon) + \frac{1}{2} < P_1 < \varrho(\epsilon) + \frac{\mathcal{T}'(0)}{\mathcal{T}'(0)+\mathcal{T}'(1)}$	$\psi^*_2 = \psi^\diamond_2$	$\gamma^*_{\psi^*_2} = \frac{\mathcal{T}(1-\psi^\diamond_2)-\mathcal{T}(\psi^\diamond_2)}{\kappa^p}$	$(P_0 + \varrho(\epsilon))\mathcal{T}(\psi^\diamond_2) + (P_1 - \varrho(\epsilon))\mathcal{T}(1 - \psi^\diamond_2)$

Table 4: Summarized results in two cases when $\mathcal{T}$ is convex.

Finally, for any given input $\mathbf{x}$, applying the preceding results to the optimal solution $(\gamma^*_{\psi^*}, \psi^*) = \arg\inf_{\psi \in [0,1]} \inf_{\gamma \geq 0} g(\gamma, \psi)$ in (A18), we obtain that the optimal action on the given instance $\mathbf{x}$ is given as below: for concave $\mathcal{T}$,

$$\psi^\star(\mathbf{x}) = \begin{cases} 0, \text{ if } P_0 \geq \varrho(\epsilon) + \dfrac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}; \\ 1, \text{ if } P_1 \geq \varrho(\epsilon) + \dfrac{\mathcal{T}(0)-\mathcal{T}(1/2)}{\mathcal{T}(0)-\mathcal{T}(1)}; \\ 1/2, \text{ otherwise}; \end{cases}$$

and for convex $\mathcal{T}$,

$$\psi^\star(\mathbf{x}) = \begin{cases} 0, \text{ if } P_0 \geq \varrho(\epsilon) + \dfrac{\mathcal{T}'(0)}{\mathcal{T}'(0)+\mathcal{T}'(1)}; \\ t_0^*, \text{ if } \varrho(\epsilon) + 1/2 < P_0 < \varrho(\epsilon) + \dfrac{\mathcal{T}'(0)}{\mathcal{T}'(0)+\mathcal{T}'(1)}; \\ 1, \text{ if } P_1 \geq \ \varrho(\epsilon) + \dfrac{\mathcal{T}'(0)}{\mathcal{T}'(0)+\mathcal{T}'(1)}; \\ t_1^*, \text{ if } \varrho(\epsilon) + 1/2 < P_1 < \varrho(\epsilon) + \dfrac{\mathcal{T}'(0)}{\mathcal{T}'(0)+\mathcal{T}'(1)}; \\ 1/2, \text{ otherwise,} \end{cases}$$

where t_0^* is the unique solution of $(P_0 - \varrho(\epsilon))\mathcal{T}'(1-t) = (P_1 + \varrho(\epsilon))\mathcal{T}'(t)$ on $t \in (0, \frac{1}{2})$, and t_1^* is the unique solution of $(P_0 + \varrho(\epsilon))\mathcal{T}'(1-t) = (P_1 - \varrho(\epsilon))\mathcal{T}'(t)$ on $t \in (\frac{1}{2}, 1)$. Hence, the proof is established.

A.7 Proof of Theorem 3.2

For ease of presentation, we omit the dependence on $\mathbf{x}$ and $\widetilde{\mathbf{y}}$ in the notation for now. Specifically, for $j \in [K]$, we let $P_j \triangleq P_j(\mathbf{x}, \widetilde{\mathbf{y}}) \triangleq P(\mathrm{Y} = j|\mathbf{x}, \widetilde{\mathbf{y}})$ and $\psi_j \triangleq \psi(\mathbf{x})_j$. Let the objective function in (9) be denoted as

$$g(\gamma;\psi) \triangleq \gamma\epsilon^p + \sum_{j=1}^K P_j \max\{1-\psi_1-\gamma\kappa^p, \ldots, 1-\psi_{j-1}-\gamma\kappa^p, \\ 1-\psi_j, 1-\psi_{j+1}-\gamma\kappa^p, \ldots, 1-\psi_K-\gamma\kappa^p\}. \qquad \text{(A23)}$$

We complete the proof in four steps. In Step 1, for each given ψ, we investigate the *inner optimization problem* in (9) by finding the optimal value of γ, defined as $\gamma_\psi^\star \triangleq \arg\min_{\gamma\geq 0} g(\gamma;\psi)$. Then, in Step 2, by substituting $\gamma_\psi^\star$ into $g(\gamma;\psi)$, we find that the *outer optimization problem* in (9) can be written in a linear programming format under certain transformations. Next, in Step 3, we find the extreme points of the associated linear programming, and finally, in Step 4, we obtain the solution format of the optimal action $\psi^\star$.

Step 1: For any $\psi \in \Psi$, finding the optimal value of γ, defined as $\gamma_\psi^\star \triangleq \arg\min_{\gamma\geq 0} g(\gamma;\psi)$. Given ψ and $\mathbf{x}$, we sort $\{\psi_1, \ldots, \psi_K\}$ in an decreasing order, denoted $\psi^{(1)} \geq \ldots \geq \psi^{(K)}$, and hence, $1-\psi^{(1)} \leq \ldots \leq 1-\psi^{(K)}$. Assume that $\{\psi^{(1)}, \ldots, \psi^{(K)}\}$ corresponds to $\{\psi_1, \ldots, \psi_K\}$ via a permutation χ, that is, $\psi^{(j)} = \psi_{\chi(j)}$ for $j \in [K]$. Correspondingly, the P_j's with the associated indexes are denoted $P^{(j)} \triangleq P_{\chi(j)}$ for $j \in [K]$. Then, for the $\chi(j)$-th element in the summation of (A23), the maximum is taken between $1-\psi^{(K)}-\gamma\kappa^p$ and $1-\psi^{(j)}$.

First, for given ψ, we examine the continuity of $g(\gamma;\psi)$ in γ by eliminating the max operators in (A23), which is conducted by comparing $1-\psi^{(K)}-\gamma\kappa^p$ and $1-\psi^{(j)}$ for $j \in [K]$ as follows.

If $1-\psi^{(1)} \geq 1-\psi^{(K)}-\gamma\kappa^p$, i.e., $\gamma \geq \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}$, then $1-\psi_j \geq 1-\psi^{(1)} \geq 1-\psi^{(K)}-\gamma\kappa^p \geq 1-\psi^{(j')}-\gamma\kappa^p$ for $j, j' \in [K]$, and hence, (A23) becomes

$$g(\gamma;\psi) = \gamma\epsilon^p + \sum_{j=1}^K P_j(1-\psi_j), \qquad \text{(A24)}$$

which is continuous in γ for $\gamma \geq \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}$.

On the other hand, if $1-\psi^{(1)} < 1-\psi^{(K)}-\gamma\kappa^p$, i.e., $0 \leq \gamma < \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}$, then we express the range of γ as:

$$\left[0, \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}\right) = \cup_{s\in[K-1]}\left[\frac{\psi^{(s+1)}-\psi^{(K)}}{\kappa^p}, \frac{\psi^{(s)}-\psi^{(K)}}{\kappa^p}\right).$$

Then we consider γ in each interval $\Big[\frac{\psi^{(s+1)}-\psi^{(K)}}{\kappa^p}, \frac{\psi^{(s)}-\psi^{(K)}}{\kappa^p}\Big)$ for $s \in [K-1]$. In this case, $1-\psi^{(s)} < 1-\psi^{(K)}-\gamma\kappa^p \leq 1-\psi^{(s+1)}$, and (A23) becomes

$$\begin{aligned} g(\gamma;\psi) &= \gamma\epsilon^p + \sum_{j=1}^{s} P^{(j)}(1-\psi^{(K)}-\gamma\kappa^p) + \sum_{j=s+1}^{K} P^{(j)}(1-\psi^{(j)}) \\ &= \sum_{j=1}^{s} P^{(j)}(1-\psi^{(K)}) + \sum_{j=s+1}^{K} P^{(j)}(1-\psi^{(j)}) + \gamma\kappa^p\Big\{\varrho(\epsilon) - \sum_{j=1}^{s} P^{(j)}\Big\}, \end{aligned} \tag{A25}$$

where the last step holds by re-arranging the arguments and using the definition of $\varrho(\epsilon)$ given after (8). Consequently,

$$\begin{aligned} &\lim_{\gamma\to((\psi^{(s)}-\psi^{(K)})/\kappa^p)-} g(\gamma;\psi) \\ =&\sum_{j=1}^{s} P^{(j)}(1-\psi^{(K)}) + \sum_{j=s+1}^{K} P^{(j)}(1-\psi^{(j)}) + \big(\psi^{(s)}-\psi^{(K)}\big)\Big\{\varrho(\epsilon) - \sum_{j=1}^{s} P^{(j)}\Big\} \\ =&\sum_{j=1}^{s} P^{(j)}(1-\psi^{(K)}) + \sum_{j=s+1}^{K} P^{(j)}(1-\psi^{(j)}) + \big(\psi^{(s)}-\psi^{(K)}\big)\Big\{\varrho(\epsilon) - \sum_{j=1}^{s-1} P^{(j)}\Big\} \\ &- P^{(s)}\big\{(1-\psi^{(K)}) - (1-\psi^{(s)})\big\} \\ =&\sum_{j=1}^{s-1} P^{(j)}(1-\psi^{(K)}) + \sum_{j=s}^{K} P^{(j)}(1-\psi^{(j)}) + \big(\psi^{(s)}-\psi^{(K)}\big)\Big\{\varrho(\epsilon) - \sum_{j=1}^{s-1} P^{(j)}\Big\} \\ =&g((\psi^{(s)}-\psi^{(K)})/\kappa^p;\psi), \end{aligned}$$

where the last step comes from the expression (A25) for $g(\gamma;\psi)$ when $\frac{\psi^{(s)}-\psi^{(K)}}{\kappa^p} \leq \gamma < \frac{\psi^{(s-1)}-\psi^{(K)}}{\kappa^p}$. Thus, $g(\gamma;\psi)$ is continuous in γ for $\gamma \in \Big[\frac{\psi^{(s+1)}-\psi^{(K)}}{\kappa^p}, \frac{\psi^{(s)}-\psi^{(K)}}{\kappa^p}\Big]$ with $s \in [K-1]$. Consequently, $g(\gamma;\psi)$ is continuous in γ for $0 \leq \gamma \leq \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}$.

Therefore, combining the discussion regrading (A24) and(A25), we obtain that given ψ, $g(\gamma;\psi)$ is continuous in γ for $\gamma \geq 0$.

Next, for each given ψ, we examine the monotonicity of $g(\gamma;\psi)$ in γ to find the γ that minimizes $g(\gamma;\psi)$. To this end, we consider the following three cases by the values of $\varrho(\epsilon)$.

Case 1: If $P^{(1)} < \varrho(\epsilon) < \sum_{j=1}^{K} P^{(j)}$,

then there exists an $s^* \in \{2,\ldots,K\}$ such that $\sum_{j=1}^{s^*-1} P^{(j)} \leq \varrho(\epsilon) \leq \sum_{j=1}^{s^*} P^{(j)}$. Then, by (A25), $g(\gamma;\psi)$ is decreasing in γ for $\gamma \in [0, \frac{\psi^{(s^*)}-\psi^{(K)}}{\kappa^p}]$ and increasing for $\gamma \in [\frac{\psi^{(s^*)}-\psi^{(K)}}{\kappa^p}, \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}]$; and by (A24), $g(\gamma;\psi)$ is increasing in γ for $\gamma \geq \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}$. Therefore, $\gamma_\psi^\star = \frac{\psi^{(s^*)}-\psi^{(K)}}{\kappa^p}$.

Case 2: If $\varrho(\epsilon) \leq P^{(1)}$,

then by (A25), $g(\gamma;\psi)$ is decreasing in γ for $\gamma \in [0, \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}]$; and by (A24), increasing in γ for $\gamma \geq \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}$. Therefore, $\gamma_\psi^\star = \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}$.

Case 3: If $\varrho(\epsilon) \geq \sum_{j=1}^{K} P^{(j)}$,

then by (A25), $g(\gamma;\psi)$ is increasing in γ for $\gamma \in [0, \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}]$; and by (A24), increasing in γ for $\gamma \geq \frac{\psi^{(1)}-\psi^{(K)}}{\kappa^p}$. Therefore, $\gamma_\psi^\star = 0$.

Therefore, we conclude that

$$\gamma_{\psi}^{\star} = \begin{cases} \dfrac{\psi^{(1)} - \psi^{(K)}}{\kappa^{p}} \text{ if } \varrho(\epsilon) \leq P^{(1)} \\ \dfrac{\psi^{(s^*)} - \psi^{(K)}}{\kappa^{p}} \text{ if } \displaystyle\sum_{j=1}^{s^*-1} P^{(j)} \leq \varrho(\epsilon) \leq \sum_{j=1}^{s^*} P^{(j)} \text{ with } s^* \in \{2, \ldots, K\} \\ 0 \text{ if } \varrho(\epsilon) \geq \displaystyle\sum_{j=1}^{K} P^{(j)}. \end{cases}$$

Step 2: Linear programming format.
For *each fixed permutation* χ, we now find the optimal ψ that minimizes $g(\gamma_{\psi}^{\star}; \psi)$ by examining the three cases in Step 1.

In Case 3 of Step 1, $g(\gamma_{\psi}^{\star}; \psi) = g(0; \psi) = 1 - \psi^{(K)} \geq 1 - 1/K$. Then the corresponding optimal action is $\psi^{(1)} = \ldots = \psi^{(K)} = 1/K$.

In Case 1 of Step 1, for a single data point $(\mathbf{x}, \widetilde{\mathbf{y}})$, by substituting $\gamma_{\psi}^{\star} = \frac{\psi^{(s^*)} - \psi^{(K)}}{\kappa^{p}}$ with $s^* \in \{2, \ldots, K\}$ into (A25), we obtain that

$$\begin{aligned} g(\gamma_{\psi}^{\star}; \psi) =& \sum_{j=1}^{s^*-1} P^{(j)}(1 - \psi^{(K)}) + \sum_{j=s^*}^{K} P^{(j)}(1 - \psi^{(j)}) + (\psi^{(s^*)} - \psi^{(K)})\Big\{\varrho(\epsilon) - \sum_{j=1}^{s^*-1} P^{(j)}\Big\} \\ =&(1 - \psi^{(K)}) \sum_{j=1}^{s^*-1} P^{(j)} + P^{(s^*)}(1 - \psi^{(s^*)}) + P^{(K)}(1 - \psi^{(K)}) + \sum_{j=s^*+1}^{K-1} P^{(j)}(1 - \psi^{(j)})\mathbf{1}(s^* < K - 1) \\ &+ (1 - \psi^{(K)})\Big\{\varrho(\epsilon) - \sum_{j=1}^{s^*-1} P^{(j)}\Big\} - (1 - \psi^{(s^*)})\Big\{\varrho(\epsilon) - \sum_{j=1}^{s^*-1} P^{(j)}\Big\} \\ =&\Big\{\sum_{j=1}^{s^*} P^{(j)} - \varrho(\epsilon)\Big\}(1 - \psi^{(s^*)}) + \sum_{j=s^*+1}^{K-1} P^{(j)}(1 - \psi^{(j)})\mathbf{1}(s^* < K - 1) \\ &+ \Big\{P^{(K)} + \varrho(\epsilon)\Big\}(1 - \psi^{(K)}). \end{aligned} \tag{A26}$$

To find the optimal value that minimizes $g(\gamma_{\psi}^{\star}; \psi)$ in (A26), we link it with a linear programming problem. Specifically, for $j \in [K]$, let $z_j \triangleq 1 - \psi^{(j)}$ and $\mathbf{z} \triangleq (z_1, \ldots, z_K)^{\top}$. Define

$$a_j = \begin{cases} \displaystyle\sum_{j=1}^{s^*} P^{(j)} - \varrho(\epsilon) & \text{if } j = s^* \\ P^{(K)} + \varrho(\epsilon) & \text{if } j = K \\ P^{(j)} & \text{if } s^* < j < K \text{ when } s^* \neq K - 1. \end{cases}$$

When $s^* = K - 1$, only the entries for $j = s^* = K - 1$ and $j = K$ need to be considered. Let $\mathsf{V}(\mathbf{z}) = \sum_{j=s^*}^{K} a_j z_j$. Then, the optimal ψ that minimizes $g(\gamma_{\psi}^{*}; \psi)$ in (A26) can be derived by solving the linear programming problem:

$$\begin{cases} \displaystyle\min_{z_1, \ldots, z_K} \mathsf{V}(\mathbf{z}), \\ \quad s.t. \displaystyle\sum_{j=1}^{K}(1 - z_j) = 1, \\ \qquad 0 \leq z_1 \leq \ldots \leq z_K \leq 1, \end{cases} \tag{A27}$$

where the constraint $\sum_{j=1}^{K}(1 - z_j) = 1$ is due to $\sum_{j=1}^{K}(1 - z_j) = \sum_{j=1}^{K} \psi^{(j)} = 1$ by the definitions of $\mathbf{z}$ and ψ, and the constraint $0 \leq z_1 \leq \ldots \leq z_K \leq 1$ reflects the definition of $\psi^{(j)}$ for $j \in [K]$.

Similarly, in Case 2 of Step 1, by substituting $\gamma_{\psi}^{\star}=\frac{\psi^{(1)}-\psi^{(K)}}{\kappa^{p}}$ into (A24), we obtain that

$$
\begin{aligned}
g(\gamma_{\psi}^{\star};\psi) =&\varrho(\epsilon)\{\psi^{(1)}-\psi^{(K)}\}+\sum_{j=1}^{K}P_j(1-\psi_j)\\
=&\varrho(\epsilon)\Big[\{1-\psi^{(K)}\}-\{1-\psi^{(1)}\}\Big]+\sum_{j=1}^{K}P_j(1-\psi_j)\\
=&\{P^{(1)}-\varrho(\epsilon)\}(1-\psi^{(1)})+\sum_{j=2}^{K-1}P^{(j)}(1-\psi^{(j)})+\{P^{(K)}+\varrho(\epsilon)\}(1-\psi^{(K)}),
\end{aligned}
$$

which is a form similar to (A26) if letting s^* in (A26) equal 1. Hence, its optimal minimizer can be found through a linear programming problem similar to (A27). Consequently, in the next step, our discussion focuses on (A26) only.

Step 3: Extreme points.
The feasible region of (A27), denoted Ξ, can be expressed as follows:

$$
\begin{aligned}
\Xi \triangleq&\Big\{\mathbf{z}:\sum_{j=1}^{K}(1-z_j)=1,\ 0\le z_1\le\ldots\le z_K\le 1\Big\}\\
=&\Big\{\mathbf{z}:\sum_{j=1}^{K}(1-z_j)=1,\ 0\le z_1\le\ldots\le z_K\le 1, 1-z_j\le\frac{1}{j}\text{ for } j\in[K]\Big\}\\
=&\Big\{\mathbf{z}:\sum_{j=1}^{K}z_j=K-1,\ 0\le z_1\le\ldots\le z_K\le 1, z_j\ge 1-\frac{1}{j}\text{ for } j\in[K]\Big\}, \qquad \text{(A28)}
\end{aligned}
$$

where the second step holds since, for $j\in[K]$, $j(1-z_j)=\sum_{t=1}^{j}(1-z_j)\le\sum_{t=1}^{j}(1-z_t)\le 1$ as $z_t\le z_j$ for $t\in[j]$, and the last step holds by rearranging the equality $\sum_{j=1}^{K}(1-z_j)=1$.

We next prove that the following K feasible solutions are the only extreme points of (A27):

$$
\begin{aligned}
&\mathbf{z}_1\triangleq(0,1,1,\ldots,1,1)^{\top},\\
&\mathbf{z}_2\triangleq\left(1-\frac{1}{2},1-\frac{1}{2},1,\ldots,1,1\right)^{\top},\\
&\ldots,\\
&\mathbf{z}_j\triangleq\Bigg(\underbrace{1-\frac{1}{j},\ldots,1-\frac{1}{j}}_{j\text{ elements}},\ \underbrace{1,\ldots,1}_{K-j\text{ elements}}\Bigg)^{\top}\\
&\ldots,\\
&\mathbf{z}_{K-1}\triangleq\left(1-\frac{1}{K-1},1-\frac{1}{K-1},1-\frac{1}{K-1},\ldots,1-\frac{1}{K-1},1\right)^{\top},\\
&\mathbf{z}_K\triangleq\left(1-\frac{1}{K},1-\frac{1}{K},1-\frac{1}{K},\ldots,1-\frac{1}{K},1-\frac{1}{K}\right)^{\top}.
\end{aligned}
$$

We denote $\Xi_0\triangleq\{\mathbf{z}_1,\ldots,\mathbf{z}_K\}$.

Firstly, we prove that each data point in Ξ_0 is an extreme point of (A27). To this end, consider any $\mathbf{z}_j\in\Xi_0$. If there exist $\nu\in(0,1)$, $\mathbf{z}'=(z_1',\ldots,z_K')^{\top}\in\Xi$, and $\mathbf{z}''=(z_1'',\ldots,z_K'')^{\top}\in\Xi$, such that $\mathbf{z}_j=\nu\mathbf{z}'+(1-\nu)\mathbf{z}''$, then $\mathbf{z}'=\mathbf{z}''=\mathbf{z}_j$, as shown below. Let $z_{j,t}$, z_t', and z_t'' represent the tth element of $\mathbf{z}_j$, $\mathbf{z}'$, and $\mathbf{z}''$, respectively.

- If $t=j+1,\ldots,K$: then by $\nu z_t'+(1-\nu)z_t''=z_{j,t}$, $z_{j,t}=1$, and $z_t',z_t''\le 1$, we have that $z_t'=z_t''=z_{j,t}=1$;

- If $t = j$: then $\nu z'_j + (1-\nu)z''_j = z_{j,j} = 1 - \frac{1}{j}$, and $z'_j, z''_j \geq 1 - \frac{1}{j}$ by (A28). Thus, we obtain that $z'_j = z''_j = z_{j,j}$.
- If $t = 1, \ldots, j-1$: then $z'_t \leq z'_j = 1 - \frac{1}{j}$, $z''_t \leq z''_j = 1 - \frac{1}{j}$, and $\nu z'_t + (1-\nu)z''_t = z_{j,t} = 1 - \frac{1}{j}$. Thus, we can also obtain that $z'_t = z''_t = z_{j,t}$.

Therefore, $\mathbf{z}' = \mathbf{z}'' = \mathbf{z}_j$, and hence, $\mathbf{z}_j$ is an extreme point of (A27) by Definition A.1.

Next, for any point $\widetilde{\mathbf{z}} \triangleq (\widetilde{z}_1, \ldots, \widetilde{z}_K)^\top \in \Xi \backslash \Xi_0$, we prove that $\widetilde{\mathbf{z}}$ is not an extreme point of (A27) by construction. Specifically, we have the following claims for $\widetilde{\mathbf{z}}$.

- *Claim 1: $\widetilde{z}_t > 1 - \frac{1}{t}$ for $t \in [K]$:*
  This claim can be proved by contradiction. Assume there exists $t_0 \in [K]$ such that $\widetilde{z}_{t_0} \leq 1 - \frac{1}{t_0}$. As $\widetilde{\mathbf{z}} \in \Xi$, by (A28) and the assumption, we have $\widetilde{z}_{t_0} = 1 - \frac{1}{t_0}$. Since $\widetilde{\mathbf{z}} \notin \Xi_0$, one of the following statements must hold: (1) there exists $j < t_0$ such that $\widetilde{z}_j < 1 - \frac{1}{t_0}$; or (2) there exists $j > t_0$ such that $\widetilde{z}_j < 1$ for some $j > t_0$. Therefore, $\sum_{j=1}^{K} \widetilde{z}_j = \sum_{j=1}^{t_0} \widetilde{z}_j + \sum_{j=t_0+1}^{K} \widetilde{z}_j < \sum_{j=1}^{t_0} \widetilde{z}_{t_0} + \sum_{j=t_0+1}^{K} 1 = t_0 \cdot \widetilde{z}_{t_0} + (K - t_0) \cdot 1 = K - 1$ since $\widetilde{z}_1 \leq \ldots \leq \widetilde{z}_K$ and $\widetilde{z}_t \leq 1$ for $t \in [K]$ by (A28), where the strict inequality arises from the fact that either statement (1) or (2) holds. This conclusion contradicts the condition that $\widetilde{\mathbf{z}} \in \Xi$ by (A28).
- *Claim 2: There exists $t_1 \in [K]$ such that $\widetilde{z}_{t_1-1} < \widetilde{z}_{t_1} < 1$:*
  This claim can be proved by contradiction:
  - On one hand, if there exists $t' \in [K]$ such that $\widetilde{z}_{t'-1} < \widetilde{z}_{t'}$, then we must have $\widetilde{z}_t = \widetilde{z}_{t'-1}$ for $t \leq t'-1$; otherwise, by letting $t'' = \arg\max\{t : \widetilde{z}_t < \widetilde{z}_{t'-1}, t < t'-1\}$, we obtain that $\widetilde{z}_{t''} < \widetilde{z}_{t''+1} = \widetilde{z}_{t'} < 1$ and hence, t_1 can be set as $t''+1$, which contradicts the assumption. Additionally, we have $\widetilde{z}_t = 1$ for $t \geq t'$; otherwise, $\widetilde{z}_{t'-1} < \widetilde{z}_{t'} < 1$ and t_1 can be set as t', which contradicts the assumption. Summarizing the discussion for $t \leq t'-1$ and $t \geq t'$, we have $\widetilde{\mathbf{z}} \in \Xi_0$.
  - On the other hand, if $\widetilde{z}_{t-1} = \widetilde{z}_t$ for all $t \in [K]$, then $\widetilde{\mathbf{z}} = \mathbf{z}_K \in \Xi_0$.

  In both cases, $\widetilde{\mathbf{z}} \in \Xi_0$, contradicting the fact that $\widetilde{\mathbf{z}} \notin \Xi_0$. Hence, Claim 2 holds.

Let $t_2 \triangleq \max\{t \in [K] : \widetilde{z}_t < 1\}$. Then, $t_2 \geq t_1$. Let

$$\not{c}_1 \triangleq \min\{\frac{\widetilde{z}_{t_1} - \widetilde{z}_{t_1-1}}{2}, \widetilde{z}_t - (1 - \frac{1}{t}) \text{ for } t \leq t_1 - 1\} \text{ and}$$
$$\not{c}_2 \triangleq \min\{\frac{\widetilde{z}_{t_1} - \widetilde{z}_{t_1-1}}{2}, \widetilde{z}_{t_1} - (1 - \frac{1}{t_1}), 1 - \widetilde{z}_t \text{ for } t_1 \leq t \leq t_2\}.$$

By Claims 1 and 2, we have that $\not{c}_1 > 0$ and $\not{c}_2 > 0$. Let $\overline{\not{c}} \triangleq \min\{(t_1 - 1)\not{c}_1, (t_2 - t_1 + 1)\not{c}_2\}$, $\overline{\not{c}}_1 \triangleq \overline{\not{c}}/(t_1 - 1)$, and $\overline{\not{c}}_2 \triangleq \overline{\not{c}}/(t_2 - t_1 + 1)$. Then we construct two points in Ξ:

$$\mathbf{z}' \triangleq (\widetilde{z}_1 + \overline{\not{c}}_1, \ldots, \widetilde{z}_{t_1-1} + \overline{\not{c}}_1, \widetilde{z}_{t_1} - \overline{\not{c}}_2, \ldots, \widetilde{z}_{t_2} - \overline{\not{c}}_2, \ldots, \widetilde{z}_K)^\top \text{ and}$$
$$\mathbf{z}'' \triangleq (\widetilde{z}_1 - \overline{\not{c}}_1, \ldots, \widetilde{z}_{t_1-1} - \overline{\not{c}}_1, \widetilde{z}_{t_1} + \overline{\not{c}}_2, \ldots, \widetilde{z}_{t_2} + \overline{\not{c}}_2, \ldots, \widetilde{z}_K)^\top.$$

Therefore, $\widetilde{\mathbf{z}} = \frac{1}{2}\mathbf{z}' + \frac{1}{2}\mathbf{z}''$, and hence, $\widetilde{\mathbf{z}}$ is not an extreme point of (A27).

Step 4: Solution format and optimal action.
By Steps 2 and 3, we obtain that*for each fixed χ and s^**, the extreme points of the linear programming problem are given in Ξ_0. By Lemma 1, every linear program has an extreme point that is an optimal solution. Hence, by the format of the K extreme points in Ξ_0, we obtain that at least one optimal action of ψ can be found in the format:

$$\psi^{(j)} = \frac{1}{k^*} \text{ for } j \leq k^* \text{ and } \psi^{(j)} = 0 \text{ for } j \geq k^* \tag{A29}$$

for some $k^* \in [K]$.

If $k^* = K$, by (A23), we have that $g(\gamma; \psi) = \gamma \epsilon^p + \sum_{j=1}^{K} P_j \cdot (1 - \frac{1}{K})$, and hence, the robust risk is $g(\gamma^\star_\psi; \psi) = 1 - \frac{1}{K}$ by taking $\gamma^\star_\psi = 0$.

If $k^* < K$, we obtain that

$$\begin{aligned} g(\gamma;\psi) =& \gamma\epsilon^p + \sum_{j=1}^{k^*} P^{(j)} \max\Big(1-\gamma\kappa^p, 1-\frac{1}{k^*}\Big) + \sum_{j=k^*+1}^{K} P^{(j)}\cdot 1 \\ =& \begin{cases} 1+\gamma\kappa^p\Big\{\varrho(\epsilon)-\sum_{j=1}^{k^*}P^{(j)}\Big\}, \text{ if } 0\le\gamma\le\dfrac{1}{k^*\kappa^p}; \\ \gamma\epsilon^p+1-\dfrac{1}{k^*}\sum_{j=1}^{k^*}P^{(j)}, \text{ if } \gamma\ge\dfrac{1}{k^*\kappa^p}. \end{cases} \end{aligned}$$

Hence, for $k^* < K$, the robust risk is the minimum of $g(\gamma_\psi^\star;\psi) = 1$ by taking $\gamma_\psi^\star = 0$ and $g(\gamma_\psi^\star;\psi) = 1 + \frac{1}{k^*}\Big\{\varrho(\epsilon) - \sum_{j=1}^{k^*} P^{(j)}\Big\}$ by taking $\gamma_\psi^\star = \frac{1}{k^*\kappa^p}$. Additionally, we observe that we should take the highest k^* values of $\{P_1,\dots,P_K\}$ as $P^{(1)},\dots,P^{(k^*)}$ to minimize $g(\gamma_\psi^\star;\psi)$. Hence, we take the permutation χ such that $P^{(1)}\ge\dots\ge P^{(K)}$.

In summary, the optimal action $\psi^\star$ that minimizes $g(\gamma_\psi^\star;\psi)$ is given as below.

- If $\frac{1}{K}\ge\frac{1}{k^*}\sum_{j=1}^{k^*}P^{(j)}-\frac{1}{k^*}\varrho(\epsilon)$ for all $k^*\in[K-1]$, then $\psi_j^\star=\frac{1}{K}$ for $j\in[K]$.
- If there exists some $k_0\in[K-1]$, $\frac{1}{k_0}\sum_{j=1}^{k_0}P^{(j)}-\frac{1}{k_0}\varrho(\epsilon)>\frac{1}{K}$, and $\frac{1}{k_0}\sum_{j=1}^{k_0}P^{(j)}-\frac{1}{k_0}\varrho(\epsilon)\ge\frac{1}{k^*}\sum_{j=1}^{k^*}P^{(j)}-\frac{1}{k^*}\varrho(\epsilon)$ for all $k^*\in[K-1]$, then $\psi^{\star(j)}=\frac{1}{k_0}$ for $j\in[k_0]$ and $\psi^{\star(j)}=0$ for $j=k_0+1,\dots,K$.

In particular, if $P^{(1)}\ge\max\{\frac{1}{K}+\varrho(\epsilon),P^{(2)}+\varrho(\epsilon)\}$, then the optimal action is given as: $\psi^{\star(1)}=1$ and $\psi^{\star(j)}=0$ for $j=2,\dots,K$. Thus, the proof is complete.

A.8 Proof of Theorem 3.3

For ease of presentation, we omit the dependence on $\mathbf{x}_i$ and $\widetilde{\mathbf{y}}_i$ in the notation for now. Specifically, for $i\in[K]$ and $j\in[K]$, let $P_{i,j}\triangleq P_j(\mathbf{x}_i,\widetilde{\mathbf{y}}_i)\triangleq P(\mathrm{Y}=j|\mathbf{x}_i,\widetilde{\mathbf{y}}_i)$ and $\psi_{i,j}\triangleq\psi(\mathbf{x}_i)_j$. For given $\mathbf{x}_i$, we sort the K elements of $\psi(\mathbf{x}_i)$, $\{\psi_{i,1},\dots,\psi_{i,K}\}$, in a decreasing order, denoted $\psi_i^{(1)}\ge\dots\ge\psi_i^{(K)}$.

We first consider the difference between the Wasserstein robust loss (10) and the nominal loss (11):

$$\begin{aligned} &\widehat{\mathfrak{R}}_\epsilon-\widehat{\mathfrak{R}} \\ &=\inf_{\gamma\ge0}\Big[\gamma\epsilon^p+\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^K P_{i,j}\max\big\{\mathcal{T}(\psi_{i,1})-\mathcal{T}(\psi_{i,j})-\gamma\kappa^p,\dots,\mathcal{T}(\psi_{i,j-1})-\mathcal{T}(\psi_{i,j})-\gamma\kappa^p,0, \\ &\qquad\qquad \mathcal{T}(\psi_{i,j+1})-\mathcal{T}(\psi_{i,j})-\gamma\kappa^p,\dots,\mathcal{T}(\psi_{i,K})-\mathcal{T}(\psi_{i,j})-\gamma\kappa^p\big\}\Big] \\ &=\inf_{\gamma\ge0}\Big[\gamma\epsilon^p+\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^K P_{i,j}\max\big\{\mathcal{T}(\psi_i^{(K)})-\mathcal{T}(\psi_{i,j})-\gamma\kappa^p,0\big\}\Big] \\ &\triangleq\inf_{\gamma\ge0}\hbar(\gamma), \end{aligned}$$

where the second equality is due to the fact that $\psi_i^{(1)}\ge\dots\ge\psi_i^{(K)}$ and that $\mathcal{T}$ is decreasing.

By definition, $\Big\{\alpha_{i,j}\triangleq\mathcal{T}(\psi_i^{(K)})-\mathcal{T}(\psi_{i,j}):i\in[n],j\in[K]\Big\}$ are ordered as $\alpha^{(1)}\ge\dots\ge\alpha^{(nK)}$, and correspondingly, the $P_{i,j}$'s with the associated indexes are denoted $P^{(1)},\dots,P^{(nK)}$. Consequently, we obtain that

$$\hbar(\gamma)=\gamma\epsilon^p+\frac{1}{n}\sum_{t=1}^{nK}P^{(t)}(\alpha^{(t)}-\gamma\kappa^p)\mathbf{1}(\alpha^{(t)}>\gamma\kappa^p). \tag{A30}$$

https://doi.org/10.52202/079017-2627

Define $\alpha^{(nK+1)} = 0$. Then, $\hbar(\gamma)$ in (A30) can be expressed as

$$\hbar(\gamma) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{t=1}^{s} P^{(t)}\alpha^{(t)} + \gamma\kappa^p \Big\{\varrho(\epsilon) - \dfrac{1}{n}\displaystyle\sum_{t=1}^{s} P^{(t)}\Big\}, \text{ if } \alpha^{(s+1)}/\kappa^p \le \gamma < \alpha^{(s)}/\kappa^p \text{ for } s \in [nK]; \\ \gamma\epsilon^p \text{ if } \gamma \ge \alpha^{(1)}/\kappa^p. \end{cases} \tag{A31}$$

Since

$$\begin{aligned} \lim_{\gamma \to (|\alpha^{(s)}|/\kappa^p)-} \hbar(\gamma) =& \frac{1}{n}\sum_{t=1}^{s} P^{(t)}\alpha^{(t)} + |\alpha^{(s)}|\{\varrho(\epsilon) - \frac{1}{n}\sum_{t=1}^{s} P^{(t)}\} \\ =& |\alpha^{(s)}|\varrho(\epsilon) + \frac{1}{n}\sum_{t=1}^{s} P^{(t)}(|\alpha^{(t)}| - |\alpha^{(s)}|) \\ =& |\alpha^{(s)}|\varrho(\epsilon) + \frac{1}{n}\sum_{t=1}^{s-1} P^{(t)}(|\alpha^{(t)}| - |\alpha^{(s)}|) \\ =& \frac{1}{n}\sum_{t=1}^{s-1} P^{(t)}|\alpha^{(t)}| + |\alpha^{(s)}|\Big\{\varrho(\epsilon) - \frac{1}{n}\sum_{t=1}^{s-1} P^{(t)}\Big\} \\ =& \hbar(|\alpha^{(s)}|/\kappa^p), \end{aligned}$$

and $\hbar(\gamma)$ is right-continuous at $\gamma = |\alpha^{(s)}|/\kappa^p$ by definition (A31), so we conclude that $\hbar(\gamma)$ is continuous at $\gamma = |\alpha^{(s)}|/\kappa^p$ for $s \in [nK]$. Hence, by (A31), $\hbar(\gamma)$ is continuous for $\gamma \ge 0$.

By (A31), $\hbar(\gamma)$ is increasing in γ on $\gamma \in [\alpha^{(s+1)}/\kappa^p, \alpha^{(s)}/\kappa^p)$ if $\varrho(\epsilon) - \frac{1}{n}\sum_{t=1}^{s} P^{(t)} > 0$, and decreasing if $\varrho(\epsilon) - \frac{1}{n}\sum_{t=1}^{s} P^{(t)} < 0$. Hence, to examine the monotonicity of $\hbar(\gamma)$ and find the γ that minimizes $\hbar(\gamma)$, we consider the following three cases by the values of $\varrho(\epsilon)$.

Case 1. If $\frac{1}{n}P^{(1)} < \varrho(\epsilon) < \frac{1}{n}\sum_{t=1}^{nK} P^{(t)}$: then there exists $s^* \in \{2, \ldots, nK\}$ such that $\varrho(\epsilon) > \frac{1}{n}\sum_{t=1}^{s} P^{(t)}$ for $s < s^*$, and $\varrho(\epsilon) \le \frac{1}{n}\sum_{t=1}^{s} P^{(t)}$ for $s \ge s^*$. Hence, by (A31), $\hbar(\gamma)$ is decreasing in γ for $\gamma \in [0, \alpha^{(s^*)}/\kappa^p]$ and increasing for $\gamma \ge \alpha^{(s^*)}/\kappa^p$. Consequently, $\gamma_\psi^\star = \alpha^{(s^*)}/\kappa^p$, and

$$\begin{aligned} \inf_{\gamma \ge 0} \hbar(\gamma) = \hbar(\alpha^{(s^*)}/\kappa^p) =& \frac{1}{n}\sum_{t=1}^{s^*-1} P^{(t)}|\alpha^{(t)}| + |\alpha^{(s^*)}|\Big\{\varrho(\epsilon) - \frac{1}{n}\sum_{t=1}^{s^*-1} P^{(t)}\Big\} \\ =& \frac{1}{n}\sum_{t=1}^{s^*-1} P^{(t)}|\alpha^{(t)}| + O\left(\frac{1}{n}\right)|\alpha^{(s^*)}|. \end{aligned}$$

Here the last step holds because by the definition of s^*, $0 < \varrho(\epsilon) - \frac{1}{n}\sum_{t=1}^{s^*-1} P^{(t)} \le \frac{1}{n}\sum_{t=1}^{s^*} P^{(t)} - \frac{1}{n}\sum_{t=1}^{s^*-1} P^{(t)} = \frac{1}{n}P^{(s^*)} \le \frac{1}{n}$, where the last inequality holds since $P^{(s^*)} \in [0, 1]$.

Case 2. If $\varrho(\epsilon) \le \frac{1}{n}P^{(1)}$: then, by (A31), $\hbar(\gamma)$ is decreasing in γ for $\gamma \in [0, \alpha^{(1)}/\kappa^p]$ and increasing for $\gamma \ge \alpha^{(1)}/\kappa^p$. Therefore, $\gamma_\psi^\star = \alpha^{(1)}/\kappa^p$, and

$$\inf_{\gamma \ge 0} \hbar(\gamma) = \hbar(\alpha^{(1)}/\kappa^p) = \varrho(\epsilon)|\alpha^{(1)}|.$$

Case 3. If $\varrho(\epsilon) \ge \frac{1}{n}\sum_{t=1}^{nK} P^{(t)}$: then, by (A31), $\hbar(\gamma)$ is increasing in γ for $\gamma \ge 0$. Therefore, $\gamma_\psi^\star = 0 = \alpha^{(nK+1)}/\kappa^p$, and

$$\inf_{\gamma \ge 0} \hbar(\gamma) = \hbar(0) = \frac{1}{n}\sum_{t=1}^{nK} P^{(t)}|\alpha^{(t)}|.$$

Hence, summarizing the discussion in the three cases above, we have that $\gamma_\psi^\star = \alpha^{(s^*)}/\kappa^p$, and the robust risk (10) is expressed as

$$\widehat{\mathfrak{R}}_\epsilon = \widehat{\mathfrak{R}} + \frac{1}{n}\sum_{t=1}^{s^*-1} P^{(t)}\alpha^{(t)}\mathbf{1}(s^* > 1) + O\left(\frac{1}{n}\right)\alpha^{(s^*)},$$

where, to provide a unified expression, we define $s^* \triangleq 1$ and $s^* \triangleq nK + 1$ in Case 2 and Case 3, respectively. Thus, the proof is completed.

B Experimental Details

B.1 Implementation Details

Datasets. We evaluate the effectiveness of the proposed AdaptCDRP on CIFAR-10 and CIFAR-100 [21] with synthetic annotations, and on four real-world datasets with human annotations: CIFAR-10N, CIFAR-100N [22], LabelMe [23, 24], and Animal-10N [25]. CIFAR-10 has 10 classes of $32 \times 32 \times 3$ color images, with 50,000 training images and 10,000 test images; CIFAR-10N provides three independent human annotated noisy labels per instance, with a majority vote yielding a 9.03% noise rate. CIFAR-100, with the same number and size of training and test images as CIFAR-10, features 100 fine-grained classes; for each instance in CIFAR-100, CIFAR-100N provides one human annotated noisy label, with a noise rate of 40.20%. LabelMe is an image classification dataset comprising 10,000 training images, 500 validation images, and 1,188 test images. The training set includes noisy and incomplete labels provided by 59 annotators, with each image being labeled an average of 2.547 times. The Animal-10N dataset contains 10 classes of $64 \times 64 \times 3$ color animal images; it includes 5 pairs of similar-looking animals, where each pair consists of two animals that are visually alike. The training dataset contains 50,000 images and the test dataset contains 5,000 images. The noise rate (mislabeling ratio) of the dataset is about 8%. For all the datasets except LabelMe, we allocate 10% of the training data as validation data used for model selection, where we choose the model with the lowest validation accuracy during training. The test data is reserved for final evaluation of the model's performance on unseen data.

Noise generation. We generate synthetic instance-dependent label noise on the CIFAR-10 and CIFAR-100 datasets using Algorihtm 2 in [29]. Each annotator is classified as an IDN-τ annotator if their mislabeling ratio is upper bounded by τ. We simulate R annotators independently, with R taking values from the set $\{5, 10, 30, 50, 100\}$. For each instance, one annotation is randomly selected from those provided by the R annotators' contributions, which evaluates methods under incomplete annotator labeling conditions. Additionally, for each R, we consider three groups of annotators with varying expertise levels, characterized by average mislabeling ratios of approximately 20%, 35%, and 50%. These groups are referred to as IDN-LOW, IDN-MID, and IDN-HIGH, indicating low, medium, and high error rates, respectively. We manually corrupt the datasets according to the following annotator groups:

R=5:

IDN-LOW. *2 IDN-10% annotators, 2 IDN-20% annotators, 1 IDN-30% annotator;*

IDN-MID. *2 IDN-30% annotators, 2 IDN-40% annotators, 1 IDN-50% annotator;*

IDN-HIGH. *2 IDN-50% annotators, 2 IDN-60% annotators, 1 IDN-70% annotator;*

R=10:

IDN-LOW. *4 IDN-10% annotators, 4 IDN-20% annotators, 2 IDN-30% annotators;*

IDN-MID. *4 IDN-30% annotators, 4 IDN-40% annotators, 2 IDN-50% annotators;*

IDN-HIGH. *4 IDN-50% annotators, 4 IDN-60% annotators, 2 IDN-70% annotators;*

R=30:

IDN-LOW. *11 IDN-10% annotators, 11 IDN-20% annotators, 8 IDN-30% annotators;*

IDN-MID. *11 IDN-30% annotators, 11 IDN-40% annotators, 8 IDN-50% annotators;*

IDN-HIGH. *11 IDN-50% annotators, 11 IDN-60% annotators, 8 IDN-70% annotators;*

R=50:

IDN-LOW. *18 IDN-10% annotators, 18 IDN-20% annotators, 14 IDN-30% annotators;*

IDN-MID. *18 IDN-30% annotators, 18 IDN-40% annotators, 14 IDN-50% annotators;*

IDN-HIGH. *18 IDN-50% annotators, 18 IDN-60% annotators, 14 IDN-70% annotators;*

R=100:

IDN-LOW. *35 IDN-10% annotators, 35 IDN-20% annotators, 30 IDN-30% annotators;*

IDN-MID. *35 IDN-30% annotators, 35 IDN-40% annotators, 30 IDN-50% annotators;*

IDN-HIGH. *35 IDN-50% annotators, 35 IDN-60% annotators, 30 IDN-70% annotators.*

Experiment setup. We employ the ResNet-18 architecture for CIFAR-10 and CIFAR-10N, and the ResNet-34 architecture for CIFAR-100 and CIFAR-100N datasets. Following [27], we use a pretrained VGG-16 model with a 50% dropout rate as the backbone for the LabelMe dataset. For the Animal-10N dataset, in line with [25], we use the VGG19-BN architecture [28] as the backbone. A batch size of 128 is maintained across all datasets. We use the Adam optimizer [43] with a weight decay of 5×10^{-4} for CIFAR-10, CIFAR-100, CIFAR-10N, CIFAR-100N, and LabelMe datasets. The initial learning rate for CIFAR-10, CIFAR-100, CIFAR-10N, and CIFAR-100N is set to 10^{-3}, with the networks trained for 120, 150, 120, and 150 epochs respectively. The first 30 epochs serve as a warm-up. For the LabelMe dataset, the model is trained for 100 epochs with an initial learning rate of 10^{-2} and a 20-epoch warm-up. For the Animal-10N dataset, the network is trained for 100 epochs with an initial learning rate of 10^{-1} and a weight decay of 10^{-3}. The learning rate is reduced by a factor of 0.1 at the 50th and 75th epochs, with the first 40 epochs designated as the warm-up stage. Training times are approximately 3 hours on CIFAR-10 and 5.5 hours on CIFAR-100 using an NVIDIA V100 GPU.

Baselines. Our method addresses learning from noisy annotations, particularly when estimated true label posteriors may be misspecified. Thus, we select baselines that either use estimated transition matrices or true label posteriors (MBEM [9], CrowdLayer [27], TraceReg [12], Max-MIG [11], CoNAL [35]). We also include baselines that aggregate labels differently (CE (MV), CE (EM) [7], DoctorNet [34], CCC [36]). Since our theoretical framework applies to both single-annotator and multiple-annotator scenarios, we also include baselines designed for single noisy labels (LogitClip [33]), particularly those employing two networks (Co-teaching [30], Co-teaching+ [31], CoDis [32]), as our method similarly uses two networks that act as priors for each other. Details of the baselines are given as follows.

(1) CE (Clean): Trains the network using the standard cross-entropy loss on clean datasets;

(2) CE (MV): Trains the network using majority voting labels;

(3) CE (EM) [7]: Aggregate labels using the EM algorithm;

(4) Co-teaching [30]: Trains two networks and cross-trains on instances with small loss values;

(5) Co-teaching+ [31]: Combines the "Update by Disagreement" with the Co-teaching method;

(6) CoDis [32]: Selects possibly clean data that have high-discrepancy prediction probabilities between two networks;

(7) LogitClip [33]: Clamps the norm of the logit vector to ensure it is upper bounded by a constant;

(8) DoctorNet [34]: Models individual annotators and learns averaging weights by combining them;

(9) MBEM [9]: Alternates between estimating annotator quality from disagreements with the current model and updating the model by optimizing a loss function that accounts for the current estimate of worker quality;

(10) CrowdLayer [27]: Concatenates the classifier with multiple annotator-specific layers and learns the parameters simultaneously;

(11) TraceReg [12]: Uses a loss function similar to CrowdLayer but adds regularization to establish identifiability of the confusion matrices and the classifier;

(12) Max-MIG [11]: Jointly aggregates noisy crowdsourced labels and trains the classifier;

(13) CoNAL [35]: Decomposes the annotation noise into common and individual confusions;

(14) CCC [36]: Simultaneously trains two models to correct the confusion matrices learned by each other via bi-level optimization.

Among these methods, Co-teaching, Co-teaching+, CoDis, and LogitClip are strong baselines for handling single noisy labels. We adapt them to the multiple annotations setting by using majority vote labels for loss computation. The results demonstrate the effectiveness of the proposed pseudo-label generation method across various scenarios. Results for CE (Clean), CE (MV), CE (EM), DoctorNet, MBEM, CrowdLayer, Max-MIG, and CoNAL in Table 1 are sourced from [10]. Baselines (1)-(3)

Algorithm 1: Learning from Noisy Labels via **C**onditional **D**istributionally **R**obust True Label **P**osterior with an **Adapt**ive Lagrange multiplier (AdaptCDRP)

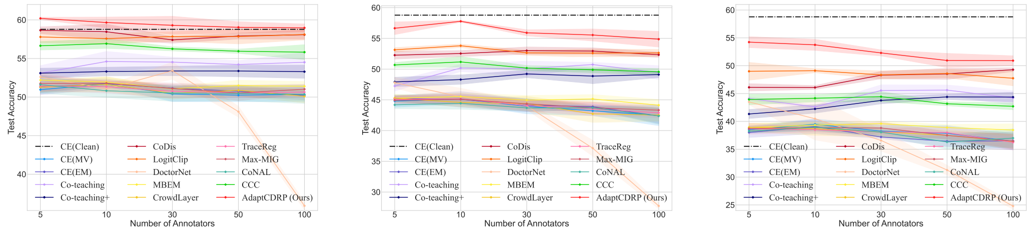
```
Input: $\mathcal{D} = \{\mathbf{x}_i, \tilde{\mathbf{y}}_i\}_{i=1}^n$, $\epsilon \in (0, \frac{1}{K})$, $\kappa > 0$, $\mathcal{C} > 1$, $\lambda > 0$
Warm up classifiers $\psi^{(1)}$ and $\psi^{(2)}$; Approximate noise transition probabilities $\widehat{\tau}_j(\widetilde{\mathbf{y}})$ for $j \in [K]$ using small-loss data;
for epoch $t = 1, ..., T$ do
    // Update the classifiers with pseudo-empirical distribution (Theorem 3.1)
    Update approximated true label posteriors: $\widehat{P}_j^{(\iota)}(\mathbf{x}, \widetilde{\mathbf{y}}) \propto \psi_j^{(\iota)}(\mathbf{x}) \cdot \widehat{\tau}_j(\widetilde{\mathbf{y}})$ for $j \in [K]$;
    for each instance $\mathbf{x}_i$ do
        if $\widehat{P}_{k^\star}^{(\iota)}(\mathbf{x}_i, \widetilde{\mathbf{y}}_i) / \max_{j \neq k^\star} \widehat{P}_j^{(\iota)}(\mathbf{x}, \widetilde{\mathbf{y}}) \geq \mathcal{C}$, let $\mathrm{y}_i^\star = k^\star$ and collect $(\mathbf{x}_i, \widetilde{\mathbf{y}}_i, \mathrm{y}_i^\star)$ into $\mathcal{D}_{t,\iota}^\star$;
    end for
    Update the pseudo-empirical distribution $P_{t,\iota}^\star$ based on $\mathcal{D}_{t,\iota}^\star$;
    Update $\psi^{(\iota)}$ by minimizing the empirical robust risk (5) with the reference distribution $P_{t,\setminus\iota}^\star$ and the Lagrange multiplier $\gamma_{t-1}^{(\iota)}$;
    // Update the Lagrange multiplier (Theorem 3.3)
    Compute $\alpha_i' s$ for $i \in [nK]$ and $s^*$ by Theorem 3.3;
    Compute the reference value for the Lagrange multiplier: $\gamma_{0,t} = |\alpha^{(s^*)}| / \kappa^p$;
    Update the Lagrange multiplier: $\gamma_t^{(\iota)} = \gamma_{0,t} - \frac{1}{\lambda}\{\epsilon^p - \mathbb{E}_{P_{t,\setminus\iota}^\star} c^p(y', \mathrm{Y})\}$
end for
Output : $\psi_1$ and $\psi_2$.
```

are implemented according to their respective algorithms, while for the remaining baseline methods, we adapted the code from the GitHub repositories provided in their original papers, with further modifications to fit our setup.

Pseudo code for the algorithm. The training process described in Section 3.3 is presented in Algorithm 1.

B.2 Additional Experimental Results

Performance on the CIFAR-100 dataset with varying numbers of annotators. We conduct additional experiments on the CIFAR-100 dataset, varying the number of annotators from 5 to 100, with each instance labeled only once. Figure 3 presents the average accuracy across different annotator counts, highlighting the advantages of the proposed method across various settings. As the total number of annotators increases, labeling sparsity becomes more pronounced, which may lead to a performance collapse in methods that do not account for this sparsity, especially in datasets with a large number of classes, such as CIFAR-100.



(a) IDN-LOW

(b) IDN-MID

(c) IDN-HIGH

Figure 3: Average test accuracy on the CIFAR-100 dataset with varying numbers of annotators. The error bars representing standard deviations are shaded.

Performance with varying numbers of annotations per instance. To further evaluate model performance with varying numbers of annotations per instance, we use $R = 30$ annotators and

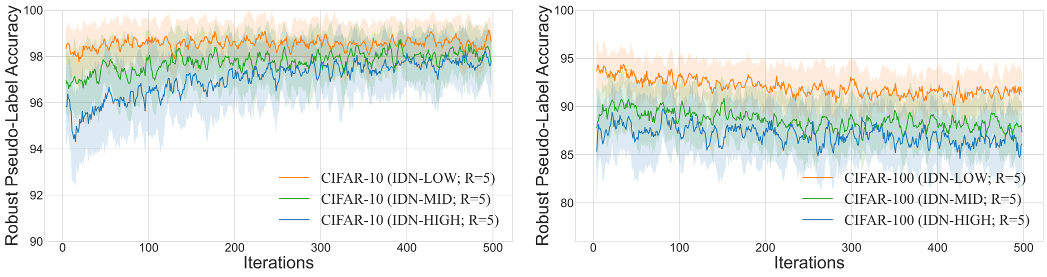
https://doi.org/10.52202/079017-2627

randomly select $l = 1, 3, 5, 7, 9$ labels from these R annotators for each instance. The test accuracies of the proposed method and other annotation aggregation methods are shown in Table 5.

Table 5: Average test accuracies (with associated standard errors expressed after the ± signs) for learning the CIFAR-10 dataset with varying numbers of annotations (denoted l) from $R = 30$ annotators.

		Ours (AdaptCDRP)	CE(MV)	CE(EM) [7]	IWMV [44]	IAA [45]
IDN-LOW	$l = 1$	$88.09_{\pm 0.37}$	$80.90_{\pm 0.88}$	$81.15_{\pm 0.74}$	–	–
	$l = 3$	$88.76_{\pm 0.30}$	$83.00_{\pm 0.40}$	$83.22_{\pm 0.43}$	$83.09_{\pm 0.35}$	$82.99_{\pm 0.40}$
	$l = 5$	$89.05_{\pm 0.36}$	$85.78_{\pm 0.53}$	$85.59_{\pm 0.46}$	$86.63_{\pm 0.29}$	$84.53_{\pm 0.52}$
	$l = 7$	$89.06_{\pm 0.30}$	$87.46_{\pm 0.32}$	$87.93_{\pm 0.53}$	$87.83_{\pm 0.24}$	$85.05_{\pm 0.55}$
	$l = 9$	$89.24_{\pm 0.51}$	$88.30_{\pm 0.25}$	$88.38_{\pm 0.32}$	$88.24_{\pm 0.12}$	$85.55_{\pm 0.46}$
IDN-MID	$l = 1$	$87.37_{\pm 0.29}$	$76.05_{\pm 0.70}$	$75.84_{\pm 0.97}$	–	–
	$l = 3$	$88.47_{\pm 0.19}$	$79.12_{\pm 0.66}$	$79.11_{\pm 0.71}$	$79.72_{\pm 0.44}$	$79.43_{\pm 0.54}$
	$l = 5$	$88.68_{\pm 0.17}$	$81.58_{\pm 0.20}$	$81.90_{\pm 0.67}$	$82.05_{\pm 0.59}$	$81.56_{\pm 0.31}$
	$l = 7$	$88.71_{\pm 0.24}$	$83.24_{\pm 0.34}$	$82.84_{\pm 0.43}$	$83.06_{\pm 0.47}$	$83.28_{\pm 0.22}$
	$l = 9$	$88.89_{\pm 0.29}$	$84.04_{\pm 0.13}$	$83.95_{\pm 0.43}$	$84.28_{\pm 0.22}$	$83.94_{\pm 0.23}$
IDN-HIGH	$l = 1$	$86.62_{\pm 0.45}$	$69.65_{\pm 1.73}$	$69.85_{\pm 1.43}$	–	–
	$l = 3$	$88.37_{\pm 0.19}$	$74.32_{\pm 0.40}$	$74.02_{\pm 1.01}$	$75.00_{\pm 0.68}$	$74.80_{\pm 0.38}$
	$l = 5$	$88.63_{\pm 0.48}$	$77.91_{\pm 1.30}$	$78.42_{\pm 0.42}$	$77.90_{\pm 0.74}$	$78.04_{\pm 0.50}$
	$l = 7$	$88.64_{\pm 0.37}$	$79.75_{\pm 0.81}$	$80.00_{\pm 0.55}$	$79.36_{\pm 0.61}$	$79.72_{\pm 0.69}$
	$l = 9$	$88.78_{\pm 0.26}$	$80.90_{\pm 0.65}$	$81.11_{\pm 0.44}$	$80.53_{\pm 0.88}$	$80.74_{\pm 0.69}$

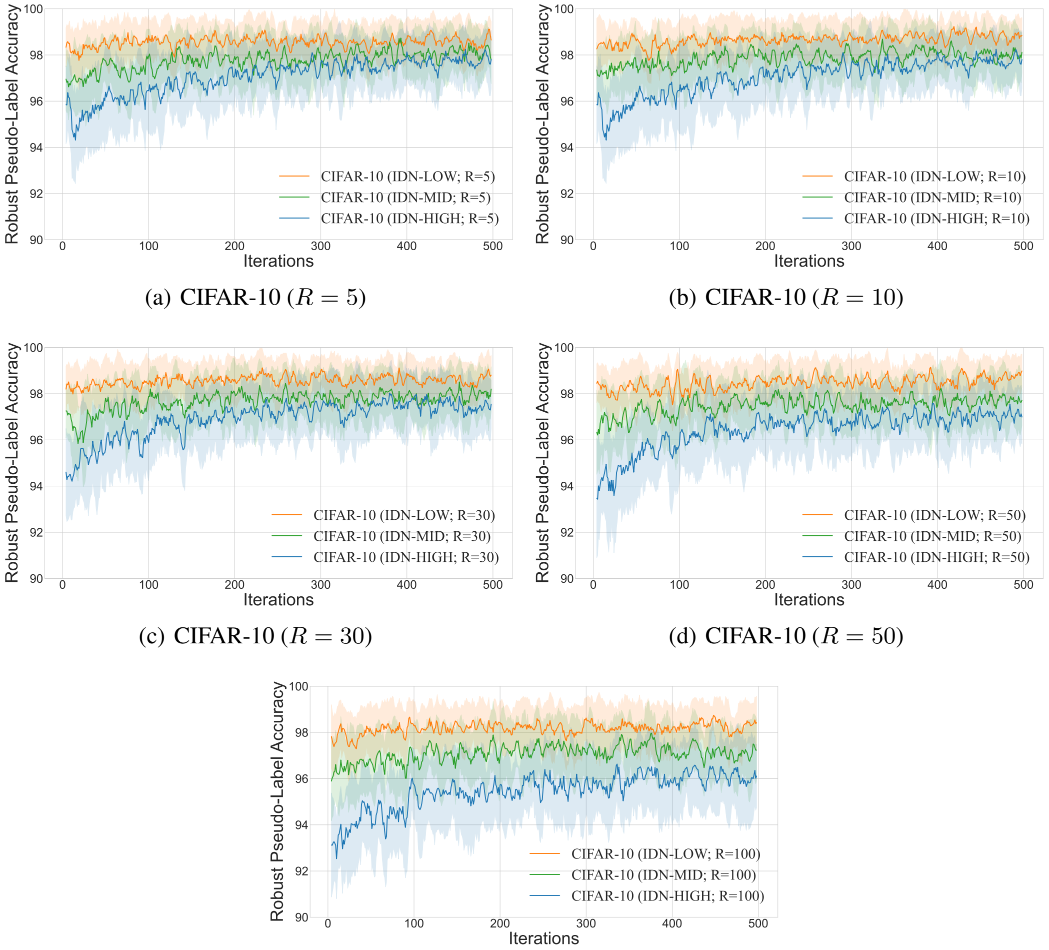
Accuracy of robust pseudo-labels. To enhance the assessment of the effectiveness of the proposed robust pseudo-label generation method, we present the average accuracy of the robust pseudo-labels on the CIFAR-10 and CIFAR-100 datasets during the training process over 5 random trials, as shown in Figure 4. Additionally, the average accuracy of the robust pseudo-labels with varying numbers of annotators on the CIFAR-10 dataset is shown in Figure 5.



(a) CIFAR-10

(b) CIFAR-100

Figure 4: Average accuracy of robust pseudo-labels on the CIFAR-10 and CIFAR-100 datasets ($R = 5$) during the training process.



(a) CIFAR-10 ($R = 5$)

(b) CIFAR-10 ($R = 10$)

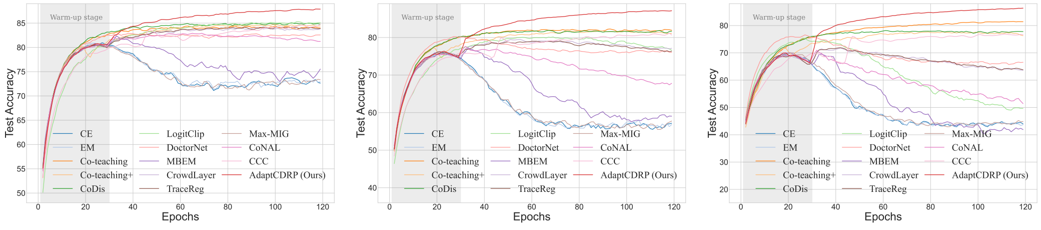
(c) CIFAR-10 ($R = 30$)

(d) CIFAR-10 ($R = 50$)

(e) CIFAR-10 ($R = 100$)

Figure 5: Average accuracy of robust pseudo-labels on the CIFAR-10 dataset with varying number of annotators in the training process.

Test accuracy during the training process. To further assess the effectiveness of the proposed method, we present the average test accuracy for the CIFAR-10 and CIFAR-100 datasets during the training process, as shown in Figures 6 and 7, respectively. The results indicate that the model tends to overfit during the warm-up stage, particularly under higher noise rates. This suggests that the results in Table 1 are not obtained with the optimal number of warm-up epochs. However, following the warm-up phase, the test accuracy of our method steadily improves, outperforming baseline methods across various scenarios.



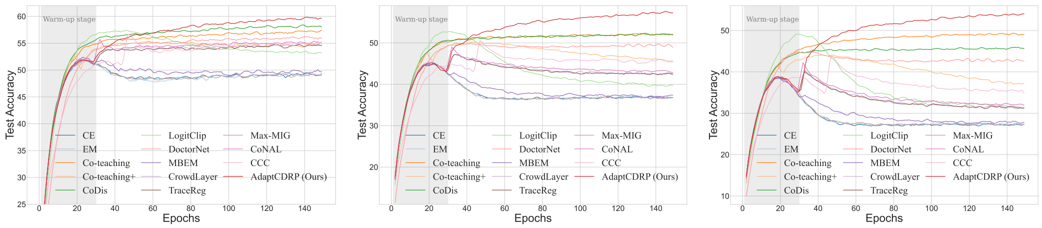
(a) IDN-LOW

(b) IDN-MID

(c) IDN-HIGH

Figure 6: Average test accuracy on learning the CIFAR-10 dataset ($R = 5$) during the training process.

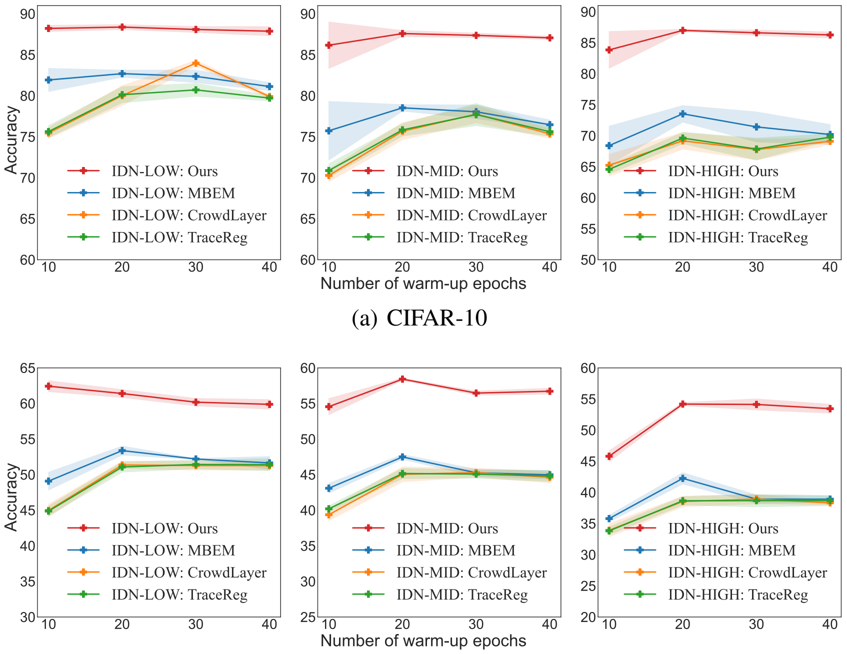
https://doi.org/10.52202/079017-2627



(a) IDN-LOW (b) IDN-MID (c) IDN-HIGH

Figure 7: Average test accuracy on learning the CIFAR-100 dataset ($R = 5$) during the training process.

Impact of the number of warm-up epochs. Following [10, 46], we use 30 warm-up epochs for the CIFAR-10 and CIFAR-100 datasets in our experiments. To rigorously assess the impact of the warm-up stage, we conduct additional experiments with varying numbers of warm-up epochs $(10, 20, 30, 40)$ on both our method and baseline approaches that also incorporate warm-up. The results, presented in Figure 8, illustrate how different warm-up durations affect performance.



(a) CIFAR-10

(b) CIFAR-100

Figure 8: Average test accuracies for learning the CIFAR-10 and CIFAR-100 datasets with varying numbers of warm-up epochs. The error bars representing standard deviation are shaded.

Different transition matrix estimation methods. Our work does not focus on precise estimation of the noise transition matrix; instead, we use a simple frequency-counting method for noise transition estimation in our experiments. Nevertheless, our approach is versatile and can be integrated with various methods for estimating the noise transition matrix or the true label posterior. Additional experiments using advanced transition matrix estimation methods are presented in Table 6. As demonstrated, integrating these methods with AdaptCDRP significantly improves test accuracies compared to directly using the estimated noise transition matrices. Furthermore, applying advanced noise transition estimation methods enhances the performance of our method on real datasets. These results highlight the robustness and adaptability of our method.

https://doi.org/10.52202/079017-2627

Table 6: Average test accuracies (with associated standard errors expressed after the ± signs) of learning the CIFAR-10 and real datasets with different transition matrix estimation methods.

Method	CIFAR-10			Real datasets			
	IDN-LOW	IDN-MID	IDN-HIGH	CIFAR-10N	CIFAR-100N	Animal10N	LabelMe
TraceReg [12]	$80.72_{\pm 0.79}$	$77.71_{\pm 1.36}$	$67.86_{\pm 1.77}$	$82.94_{\pm 0.27}$	$47.71_{\pm 0.70}$	$80.34_{\pm 0.66}$	$83.10_{\pm 0.15}$
TraceReg+Ours	$87.74_{\pm 0.26}$	$86.76_{\pm 0.07}$	$85.83_{\pm 0.37}$	$88.38_{\pm 0.35}$	$52.16_{\pm 1.02}$	$83.05_{\pm 0.26}$	$83.80_{\pm 0.44}$
GeoCrowdNet (F) [47]	$84.73_{\pm 0.39}$	$81.44_{\pm 1.00}$	$77.29_{\pm 1.23}$	$87.49_{\pm 0.45}$	$47.74_{\pm 1.17}$	$81.07_{\pm 0.45}$	$84.59_{\pm 0.19}$
GeoCrowdNet (F) + Ours	$88.06_{\pm 0.33}$	$87.43_{\pm 0.29}$	$86.69_{\pm 0.13}$	$88.30_{\pm 0.13}$	$51.07_{\pm 0.57}$	$83.12_{\pm 0.42}$	$86.20_{\pm 0.48}$
GeoCrowdNet (W) [47]	$83.82_{\pm 0.53}$	$75.72_{\pm 1.10}$	$64.64_{\pm 2.23}$	$87.36_{\pm 0.24}$	$47.49_{\pm 0.91}$	$80.19_{\pm 0.33}$	$81.63_{\pm 1.49}$
GeoCrowdNet (W) + Ours	$87.94_{\pm 0.35}$	$87.21_{\pm 0.33}$	$83.48_{\pm 5.69}$	$87.81_{\pm 0.12}$	$52.03_{\pm 0.45}$	$82.41_{\pm 0.04}$	$83.32_{\pm 0.51}$
BayesianIDNT [10]	$86.46_{\pm 1.07}$	$85.14_{\pm 0.96}$	$82.49_{\pm 2.86}$	$87.83_{\pm 0.53}$	$51.06_{\pm 0.72}$	$81.22_{\pm 0.59}$	$83.01_{\pm 0.32}$
BayesianIDNT + Ours	$87.66_{\pm 0.85}$	$86.44_{\pm 0.57}$	$84.38_{\pm 0.10}$	$88.31_{\pm 0.20}$	$53.13_{\pm 0.74}$	$83.80_{\pm 0.44}$	$84.09_{\pm 0.53}$

Impact of sparse annotation. To further address the issue of annotation sparsity, we increase the total number of annotators, R, to 200, and manually corrupt the datasets according to the following annotator groups:

R=200:

IDN-LOW. *70 IDN-10% annotators, 70 IDN-20% annotators, 60 IDN-30% annotators;*

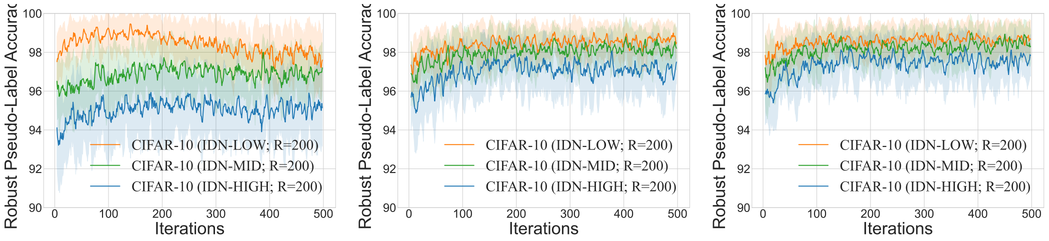
IDN-MID. *70 IDN-30% annotators, 70 IDN-40% annotators, 60 IDN-50% annotators;*

IDN-HIGH. *70 IDN-50% annotators, 70 IDN-60% annotators, 60 IDN-70% annotators.*

The three groups of annotators, labeled as IDN-LOW, IDN-MID, and IDN-HIGH, have average labeling error rates of approximately 26%, 34%, and 42%, respectively. In this setup, we incorporate regularization techniques - specifically, GeoCrowdNet (F) and GeoCrowdNet (W) penalties [47] - into our method. We then compare the results against those obtained using the traditional frequency-counting approach for estimating the noise transition matrices. Table 7 presents the performance of our proposed method on the CIFAR10 ($R = 200$) dataset, where different approaches are used to estimate the noise transition matrices. In addition, Figure 9 displays the average accuracies of the robust pseudo-labels generated by our method during the training process. These pseudo-labels play a crucial role in constructing the pseudo-empirical distribution.

Table 7: Average test accuracies (with associated standard errors expressed after the ± signs) of learning the CIFAR-10 dataset ($R = 200$) with different transition matrix estimation methods.

Mthod	IDN-LOW	IDN-MID	IDN-HIGH
Ours + frequency-counting	$86.01_{\pm 0.67}$	$85.48_{\pm 0.58}$	$85.07_{\pm 0.59}$
Ours + GeoCrowdNet (F) penalty [47]	$90.89_{\pm 0.21}$	$90.27_{\pm 0.46}$	$89.25_{\pm 0.63}$
Ours + GeoCrowdNet (W) penalty [47]	$90.99_{\pm 0.42}$	$90.23_{\pm 0.27}$	$89.42_{\pm 0.29}$



(a) Ours + Frequency Counting (b) Ours + GeoCrowdNet (F) (c) Ours + GeoCrowdNet (W)

Figure 9: Average accuracy of robust pseudo-labels on the CIFAR-10 dataset ($R = 200$) using different transition matrix estimation methods during the training process.

https://doi.org/10.52202/079017-2627

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect our contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work performed by the authors in the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-2627

Answer: [Yes]

Justification: We provide the assumptions in the main text, and the proofs can be found in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all the experimental details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We upload our code and use public datasets.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the details and the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide the error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

https://doi.org/10.52202/079017-2627

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the training time on different datasets.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: It's not appropriate for the scope and focus of our paper, and we don't see any direct negative social impacts of our paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-2627

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our model doesn't have a high risk for misuse or dual-use.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-2627

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the data generation details.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2627