Conformal Margin Risk Minimization: An Envelope Framework for Robust Learning Under Label Noise

Yuanjie Shi* **Peihong Li*** **Zijian Zhang** **Jana Doppa** **Yan Yan**
School of EECS, Washington State University, Pullman, WA, USA

Abstract

Most methods for learning with noisy labels require privileged knowledge such as noise transition matrices, clean subsets or pretrained feature extractors, resources typically unavailable when robustness is most needed. We propose **C***onformal* **M***argin* **R***isk* **M***inimization (CMRM)*, a plug-and-play envelope framework that improves *any* classification loss under label noise by adding a single quantile-calibrated regularization term, with no privileged knowledge or training pipeline modification. CMRM measures the confidence margin between the observed label and competing labels, and thresholds it with a conformal quantile estimated per batch to focus training on high-margin samples while suppressing likely mislabeled ones. We derive a learning bound for CMRM under arbitrary label noise requiring only mild regularity of the margin distribution. Across five base methods and six benchmarks with synthetic and real-world noise, CMRM consistently improves accuracy (up to +3.39%), reduces conformal prediction set size (up to −20.44%) and does not hurt under 0% noise, showing that CMRM captures a method-agnostic uncertainty signal that existing mechanisms did not exploit.

1 INTRODUCTION

Deep neural networks have achieved remarkable success across domains such as vision (Zhao et al., 2024; Zhang et al., 2024), language (Naveed et al., 2025; Teubner et al., 2023), and healthcare (Thirunavukarasu et al., 2023; Celard et al., 2023), but their performance typically depends on clean labels (Arpit et al., 2017). However, in real-world scenarios, labels are often corrupted by human annotation errors (e.g., crowd-sourcing) or automated data collection pipelines (Song et al., 2022; Gupta and Gupta, 2019). Learning with label noise (LNL) formalizes this setting, where corrupted labels distort empirical risk and lead to overfitting (Natarajan et al., 2013; Zhang et al., 2021; Arpit et al., 2017). This issue is particularly severe in high-stakes domains such as medical imaging (Shi et al., 2024a) and autonomous driving (Li et al., 2021a), making robustness to label noise an important problem in modern machine learning (Song et al., 2022; Frénay and Verleysen, 2013).

Learning under label noise has been extensively studied (Song et al., 2022), with existing approaches broadly falling into the following categories. Noise modeling approaches assume a specific corruption process (e.g., symmetric, class-conditional, or instance-dependent) and often require estimating or knowing the noise transition matrix (Natarajan et al., 2013; Patrini et al., 2017; Li et al., 2021b; Cheng et al., 2020), which is typically unobservable and difficult to model under complex noise (Yao et al., 2020). Loss correction methods adjust training objectives to counteract label corruption but typically depend on auxiliary information, such as small trusted clean subsets or accurate noise rate estimates (Hendrycks et al., 2018; Xia et al., 2019), both of which are often unavailable or unreliable in practice. Auxiliary methods leverage additional sources, such as peer networks, semi-supervised learning (SSL) pipelines, or external expert models such as large language models (LLMs), which introduce strong data and supervision requirements (Han et al., 2018; Li et al., 2020; Wang et al., 2024). Table 1 summarizes representative methods and their key assumptions (see Section 3 for details).

Despite this progress, these approaches share a common limitation: they all require some form of privileged knowledge about the noise process or access to aux-

*Equal contribution.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

Methods	Assumptions
LNL (Natarajan et al., 2013)	Symmetric noise
Forward (Patrini et al., 2017)	Known label noise transition matrix
GLC (Hendrycks et al., 2018)	Small clean data subsets
Co-teaching (Han et al., 2018)	Two peer nets
CORES (Cheng et al., 2020)	Instance-dependent noise
VolMinNet (Li et al., 2021b)	Class-conditional noise
FINE (Kim et al., 2021)	Eigen-structure noise
ELR+ (Liu et al., 2020)	Clean samples learned first
NCFW (Zhang and Agarwal, 2024)	Known class posteriors
CSGN (Lin et al., 2024)	Latent causal graph transition
NI-ERM (Zhu et al., 2024)	Strong pretrained feature extractor
NoiseGPT (Wang et al., 2024)	External LLM
CMRM (ours)	Smooth CDF + positive density

Table 1: **Classical and recent methods for learning from noisy labels and their assumptions**. CMRM only assumes mild regularity conditions, a smooth CDF with positive density at the target quantile. Details in Section 3. CMRM's assumptions are automatically satisfied in standard training (see Figure 2(d)).

iliary resources, such as a noise transition matrix, a clean data subset, a peer network or a pretrained feature extractor, which is typically unavailable in the settings where robustness to label noise is most needed, especially under severe or heterogeneous noise (Arazo et al., 2019; Song et al., 2022). This raises the main research question of this paper: *Can we design a flexible envelope framework that enhances the robustness of existing methods under arbitrary label noise, requiring only standard mathematical regularity conditions, instead of any knowledge of the noise process or access to auxiliary supervision?*

To answer this question, we propose **C***onformal* **M***argin* **R***isk* **M***inimization (CMRM)*, an envelope framework to improve the robustness and accuracy of prior methods. CMRM does not rely on noise models, clean subsets or auxiliary supervision. CMRM achieves the goal through an uncertainty-aware training objective that integrates confidence margins and conformal quantiles. Confidence margin, defined as the gap between the confidence of the observed label and competing labels, provides a principled signal to distinguish reliable from uncertain training samples (Cui et al., 2019; Pleiss et al., 2020). Conformal quantiles offer distribution-free, statistically valid thresholds for these margins (Angelopoulos and Bates, 2021; Lei et al., 2018).

This algorithm design is motivated by two prior observations: (1) label noise creates a mismatch between corrupted labels and the evolving confidence structure of the model (Pleiss et al., 2020), and (2) ignoring this uncertainty often leads to overfitting on noisy samples and prevents algorithms from leveraging informative signals (Zhang et al., 2021; Arpit et al., 2017). Building on this motivation, CMRM formulates a conformal margin risk that directly incorporates uncertainty into the training objective. Specifically, CMRM computes per-example confidence margins, estimates a conformal quantile to set an adaptive threshold, and minimizes a conformal risk defined as the average negative margin above that threshold. Importantly, CMRM requires no architectural modification, no additional networks and no clean validation data, as it adds only a single quantile-calibrated regularization term to any existing training objective. We further establish a learning bound for CMRM under arbitrary noise. Our extensive experiments on both binary and multi-class classification benchmarks demonstrate that, without requiring prior knowledge of noise, CMRM consistently improves methods with fundamentally different LNL design principles, indicating that conformal margin calibration captures a method-agnostic uncertainty signal that existing mechanisms fail to fully exploit.

Contributions. Our key contributions include:

- We propose CMRM, an envelope framework whose robustness requires only mild regularity of the margin distribution that is automatically satisfied in standard training, instead of assumptions on noise models, clean data or auxiliary resources (see Table 1).
- We derive a learning bound under arbitrary label noise that decomposes the gap between the noisy surrogate and clean conditional margin risk into function-class complexity, quantile estimation error, and distribution shift.
- We demonstrate that CMRM, requiring only a single regularization term, improves accuracy across all base method and dataset combinations tested and reduces prediction set size in nearly all cases, across five base methods (CE, Focal, LDAM, GCE, NI-ERM) and six benchmarks with synthetic and real-world noise, and incurs no accuracy penalty when labels are clean. The CMRM code is available at https://github.com/YuanjieSh/CMRM.

2 RELATED WORK

Learning with Label Noise (LNL) studies how to train models when labels are corrupted by annotation errors, heuristic labeling, or unreliable sources (Frénay and Verleysen, 2013; Song et al., 2022; Johnson and Khoshgoftaar, 2022). Many approaches have been proposed, often under specific assumptions on the noise mechanism or requiring auxiliary supervision (see Table 1 and Section 3). These include loss or posterior correction with known noise models (Natarajan et al., 2013; Patrini et al., 2017; Li et al., 2021b), methods using clean side information or peer networks (Hendrycks et al., 2018; Han et al., 2018; Li et al., 2020; Liu et al., 2020), approaches for instance-dependent or structured

noise (Cheng et al., 2020; Kim et al., 2021; Lin et al., 2024), and methods leveraging external knowledge such as pretrained features or LLMs (Zhang and Agarwal, 2024; Zhu et al., 2024; Wang et al., 2024). In contrast, our envelope approach requires only mild regularity conditions and is compatible with most prior methods.

Conformal Prediction (CP) is a distribution-free uncertainty quantification framework that constructs prediction sets with guaranteed finite-sample coverage (Vovk et al., 2005; Lei et al., 2018; Angelopoulos and Bates, 2021; Fontana et al., 2023; Ghosh et al., 2023a,b; Shi et al., 2024b; Shahrokhi et al., 2025). Recent conformal training methods incorporate these principles into the training process, focusing on minimizing prediction set size (Stutz et al., 2021; Shi et al., 2025) or refining coverage calibration (Einbinder et al., 2022; Kiyani et al., 2024), but assume clean training data. In contrast, CMRM leverages conformal principles to enhance discriminative ability under noisy supervision. Importantly, CMRM is not a conformal prediction method in the standard sense; it does not construct prediction sets or provide coverage guarantees at inference time.

3 PROBLEM SETUP AND MOTIVATION

Notations. Suppose $X \in \mathcal{X}$ is an input from $\mathcal{X}$, and $Y \in \mathcal{Y} = \{0, 1, \cdots, K-1\}$ is the ground-truth label, where K is the number of candidate classes. Let the underlying data distribution be $\mathcal{P}(X, Y)$, which characterizes the relationship between inputs and class labels. We denote by $\mathcal{P}(X)$ the marginal distribution of inputs, and by $\mathcal{P}(Y|X)$ the conditional label distribution given inputs. Let $f : \mathcal{X} \to \mathbb{R}^K$ denote the logit vector produced by soft classifier $f \in \mathcal{F}$, where $\mathcal{F}$ denotes a hypothesis class. We also define $P_f(X) = \sigma \circ f(X) : \mathcal{X} \to \Delta_+^K$ as the corresponding confidence score, where Δ_+^K is the (K-1)-dimensional probability simplex, and σ is the Sigmoid function in binary and Softmax function in multi-class setting, respectively. Let $P_f(X)_y$ denote the confidence score of class y. Define $\mathbb{1}[\cdot]$ as an indicator function. Denote $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{test}}$ as the training and test sets. Let $\mathcal{B}$ be a randomly sampled batch of training data of size s. We denote by $W_1(\mathcal{P}, \mathcal{Q}) = \inf_{\pi \in \Pi(\mathcal{P}, \mathcal{Q})} \int_{\mathcal{X} \times \mathcal{X}} d(x, x') \, \mathrm{d}\pi(x, x')$ as the Wasserstein-1 distance between two distributions $\mathcal{P}$ and $\mathcal{Q}$ on metric space $(\mathcal{X}, d)$, where $\Pi(\mathcal{P}, \mathcal{Q})$ is the set of all joint distributions of $\mathcal{P}$ and $\mathcal{Q}$.

Learning with Noisy Labels (LNL). In many real-world scenarios, the training data are corrupted by noisy labels. Instead of the clean label Y, we only observe a potentially corrupted label $\widetilde{Y} \in \mathcal{Y}$ generated from Y through an unknown label noise transition matrix $T(\widetilde{Y} \mid X, Y)$ (Zhu et al., 2024). For example, symmetric noise corresponds to $T(\widetilde{Y} \mid X, Y)$ being uniform over all incorrect labels, while class-conditional noise assumes dependence only on Y. We highlight that *our proposed CMRM framework does not rely on such assumptions and permits T to be arbitrary.*

Such noisy labels setting arises typically due to annotator mistakes, inconsistent labeling criteria, or spurious labels introduced by large-scale data collection pipelines (Natarajan et al., 2013; Patrini et al., 2017). Formally, we define the noisy training data distribution as:

$$\mathcal{P}_{\text{noisy}}(X, \widetilde{Y}) \;=\; \sum_{y \in \mathcal{Y}} \mathcal{P}(X, y)\, T(\widetilde{Y} \mid X, y).$$

Accordingly, the training set $\mathcal{D}_{\text{tr}} = \{(X_i, \widetilde{Y}_i)\}_{i=1}^n$ of size n is drawn from $\mathcal{P}_{\text{noisy}}$, while the clean test set $\mathcal{D}_{\text{test}}$ is drawn from $\mathcal{P}$.

Limitations of Existing Approaches. Despite substantial progress on learning from noisy labels, most existing methods remain fundamentally constrained by strong assumptions. As summarized in Table 1, these assumptions fall into three major categories:

First, many approaches impose explicit noise-model assumptions, from symmetric noise or known transition matrices (Natarajan et al., 2013; Patrini et al., 2017) to class-conditional (Li et al., 2021b), instance-dependent (Cheng et al., 2020), or eigen-structure noise (Kim et al., 2021), but these still rarely align with real-world corruption.

Second, a range of methods rely on auxiliary information, such as clean subsets (Hendrycks et al., 2018), noise transition matrices (Patrini et al., 2017), strong pretrained feature extractors (Zhu et al., 2024), or external LLMs (Wang et al., 2024), which are costly and unavailable at scale.

Third, some approaches exploit architectural assumptions or model-specific heuristics, such as peer networks (Han et al., 2018; Li et al., 2020), early-learning assumptions (Liu et al., 2020), or latent causal structures (Lin et al., 2024).

Overall, these assumptions limit existing methods, especially under severe or heterogeneous noise (Arazo et al., 2019; Song et al., 2022). Therefore, an envelope framework that provides robustness guarantees under only mild regularity conditions is needed.

4 CMRM FRAMEWORK

In this section, we first describe the general *Conformal Margin Risk Minimization (CMRM)* envelope framework and its variant for binary classification. Next, we develop a practical optimization algorithm for CMRM and theoretically analyze its learning bound.

4.1 General Framework

Confidence Margin. The confidence margin quantifies the separation between the confidence assigned to the observed label and the highest confidence among other candidate labels:

$$M_f(X_i, \widetilde{Y}_i) = P_f(X_i)_{\widetilde{Y}_i} - \max_{y \in \mathcal{Y} \setminus \{\widetilde{Y}_i\}} P_f(X_i)_y. \quad (1)$$

This notion of confidence margin has been widely used in the machine learning literature (Cui et al., 2019; Bartlett and Mendelson, 2002). Large margins indicate a strong preference for the observed label, whereas small or negative margins reflect uncertainty (Cui et al., 2019). Margins are widely used to characterize decision boundary distance (Bartlett and Mendelson, 2002; Neyshabur et al., 2017) and to measure predictive uncertainty (Elsayed et al., 2018; Jiang et al., 2018; Lakshminarayanan et al., 2017).

Conformal Quantile. Conformal prediction (CP) (Vovk et al., 2005; Angelopoulos and Bates, 2021) provides a distribution-free mechanism for defining quantile thresholds. Let V be a real-valued random variable, and let $\{V_i\}_{i=1}^m$ denote an i.i.d. sample from V with size m. Given a target level $\alpha \in (0, 1)$, the empirical conformal quantile is defined as $\widehat{\tau}_\alpha = Q\big(\alpha, \{V_i\}_{i=1}^m\big)$, where Q selects the $\lceil \alpha(m+1) \rceil$-th largest value in $\{V_i\}_{i=1}^m$. Hence, at most an α-fraction of the sample values lie below $\widehat{\tau}_\alpha$, i.e., the empirical α-quantile. Through CP, this quantile enjoys distribution-free validity guarantee, making it a principled tool for UQ.

Conformal Margin Risk. While conformal quantiles provide principled thresholds, their effectiveness depends on using a score that reflects label reliability under noise. Confidence margin serves this role: clean samples typically exhibit large margins, whereas noisy labels tend to have small margins (Zhang and Sabuncu, 2018; Liu et al., 2020; Zhang et al., 2021). This is intuitive: correct labels align with dominant class evidence, while corrupted labels conflict with the input features and are outscored by alternative labels. Figure 1 illustrates this separation on CIFAR-100 under different types of noise (class-conditional and human annotation noise). However, such observations have mainly been used for diagnostic analysis or heuristic filtering, without formal risk formulations.

CMRM closes this gap by calibrating confidence margins with conformal quantiles to yield an assumption-light and uncertainty-aware risk. Specifically, we instantiate V with confidence margin $M_f(X, \widetilde{Y})$, and define the conformal quantile threshold on $\{M_f(X_i, \widetilde{Y}_i)\}_{i=1}^n$:

$$\widehat{\tau}_\alpha(f) = Q\Big(\alpha, \{M_f(X_i, \widetilde{Y}_i)\}_{i=1}^n\Big), \quad (2)$$

so that at most an α-fraction of samples fall below

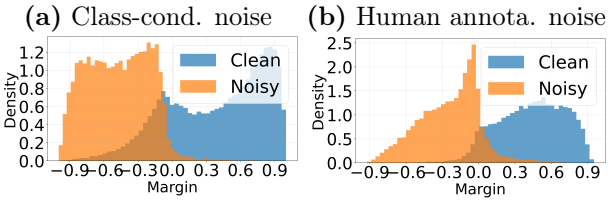

Figure 1: **Confidence margin distributions for clean (blue) and noisy (orange) samples on CIFAR-100.** **(a)** Class-conditional noise at 20% and **(b)** human annotation noise at 40%. In both cases, clean samples concentrate on positive margins while noisy samples shift to negative, showing that confidence margins can distinguish clean from noisy labels without assumptions on the noise process.

this threshold, a property that always holds without assumptions on the noise process. α is a hyper-parameter tuned on validation set. Larger α filters more aggressively, improving noise robustness but risking loss of clean samples, while smaller α retains more signal but with corruption. α is a hyper-parameter selected through validation data (see sensitivity study in Section 5.1).

Next, we define the conformal margin risk $\widehat{\mathcal{L}}_{\mathrm{cr}}(f)$ as:

$$\widehat{\mathcal{L}}_{\mathrm{cr}}(f) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\ell}_{\mathrm{cr}}(f, X_i, \widetilde{Y}_i), \text{ such that} \quad (3)$$

$$\widehat{\ell}_{\mathrm{cr}}(f, X_i, \widetilde{Y}_i) = -M_f(X_i, \widetilde{Y}_i) \cdot \widetilde{\mathbb{1}}[M_f(X_i, \widetilde{Y}_i) \geq \widehat{\tau}_\alpha(f)],$$

where $\widehat{\ell}_{\mathrm{cr}}(f, X_i, \widetilde{Y}_i)$ is the per-sample conformal margin loss of $(X_i, \widetilde{Y}_i)$, and $\widetilde{\mathbb{1}}[x \geq y] = 1/(1 + \exp(-(x - y)/\texttt{temp}))$ is the smoothed indicator by Sigmoid function with temperature parameter `temp`. This formulation assigns a soft weight in $(0, 1)$ to each sample rather than hard 0/1 filtering, smoothly downweighting low-margin samples during training, similar to soft reweighting strategies in Tjandra and Wiens (2023).

By construction, $\widehat{\mathcal{L}}_{\mathrm{cr}}(f)$ is the average negative margin over the selected samples, i.e., those with margins above the threshold. Minimizing $\widehat{\mathcal{L}}_{\mathrm{cr}}(f)$ increases the separation on selected high-confidence samples while discarding low-confidence samples. Thus, this principle aligns training with predictive uncertainty and filters out potentially corrupted labels (empirically verified in Figure 2(b)), without assumptions on the noise model or auxiliary supervision.

Finally, we incorporate CMRM as a plug-and-play regularizer for any classification loss $\widehat{\mathcal{L}}_{\mathrm{cl}}(f)$ (e.g., cross-entropy), including prior LNL loss functions, as demonstrated in our empirical evaluation:

$$\widehat{\mathcal{L}}(f) = \widehat{\mathcal{L}}_{\mathrm{cl}}(f) + \lambda \cdot \widehat{\mathcal{L}}_{\mathrm{cr}}(f), \quad (4)$$

where $\lambda \geq 0$ controls the strength of the CMRM term relative to $\widehat{\mathcal{L}}_{\mathrm{cl}}(f)$. The above formulation is general

for multi-class classification problems with standard accuracy metrics. However, binary classification tasks require optimizing more specific metrics. Hence, we provide an instantiation for binary classification below.

CMRM for Binary Classification. Binary classification often relies on class-conditional performance measures such as false positive and false negative rates (FPR and FNR), which allow users to impose class-conditional tolerances (Zhou and Liu, 2006). Since CMRM's quantile-based formulation naturally extends to class-conditional settings, we introduce a binary variant that adapts its margin thresholding to each class separately, enabling asymmetric error control while preserving its core structure.

First, we define class-conditional quantile thresholds for tolerances α^+ and α^-:

$$\widehat{\tau}^-(f) = \min\Big\{t : \frac{1}{n_0}\sum_{i:\widetilde{Y}_i=0} \mathbb{1}\big[P_f(X_i)_1 \geq t\big] \leq \frac{\lceil \alpha^-(n_0+1)\rceil}{n_0}\Big\},$$

$$\widehat{\tau}^+(f) = \max\Big\{t : \frac{1}{n_1}\sum_{i:\widetilde{Y}_i=1} \mathbb{1}\big[P_f(X_i)_1 \leq t\big] \leq \frac{\lceil \alpha^+(n_1+1)\rceil}{n_1}\Big\},$$

where n_0 and n_1 are the numbers of observed negative and positive samples, respectively. These thresholds control the upper tail of the negative class (potential false positives) and the lower tail of the positive class (potential false negatives).

Next, we define a two-sided hinge formulation relative to these thresholds:

$$\widehat{\mathcal{L}}^{\text{bin}}_{\text{cr}}(f) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\ell}^{\text{bin}}_{\text{cr}}(f, X_i, \widetilde{Y}_i), \text{ such that} \tag{5}$$

$$\widehat{\ell}^{\text{bin}}_{\text{cr}}(f, X_i, \widetilde{Y}_i) = -\lambda^- \widetilde{\mathbb{1}}[\widetilde{Y}_i = 0] \cdot \big(P_f(X_i)_1 - \widehat{\tau}^-(f)\big)^+ - \lambda^+ \widetilde{\mathbb{1}}[\widetilde{Y}_i = 1] \cdot \big(\widehat{\tau}^+(f) - P_f(X_i)_1\big)^+,$$

where $(z)^+ = \max\{0, z\}$ and $\lambda^+, \lambda^- > 0$ control the relative strength of the two penalties.

This binary classification variant preserves the two core ingredients of CMRM: conformal quantile-based filtering and margin maximization on retained samples. By separately controlling the class-conditional distribution, it aligns naturally with standard binary performance metrics such as FPR and FNR.

4.2 Optimization and Theoretical Analysis

Optimization Algorithm. A key challenge in CMRM is that its objective depends on set-level quantiles of confidence margins over the full training data distribution, requiring $O(n \log n)$ sorting per iteration and thus prohibitive for large-scale training. We address this by replacing set-level quantiles with batch-level quantiles $\widehat{\tau}^s_\alpha(f)$: each iteration estimates the quantile from a batch of size s, reducing the cost to $O(s \log s)$. This yields a tractable surrogate objective $\widehat{\mathcal{L}}^s_{\text{cr}}(f)$ by using $\widehat{\tau}^s_\alpha(f)$:

$$\widehat{\mathcal{L}}^s_{\text{cr}}(f) = \mathop{\mathbb{E}}_{\widehat{\tau}^s_\alpha(f) \sim \mathcal{T}_\alpha(f)} \Bigg[\sum_{i=1}^{n} - M_f(X_i, \widetilde{Y}_i) \cdot \widetilde{\mathbb{1}}[M_f(X_i, \widetilde{Y}_i) \geq \widehat{\tau}^s_\alpha(f)]\Bigg], \tag{6}$$

where $\mathcal{T}_\alpha(f)$ is the underlying distribution of $\widehat{\tau}^s_\alpha(f)$. Although this introduces an approximation gap relative to population quantiles, our theoretical analysis shows that the gap is bounded.

Algorithm 1 Conformal Margin Risk Minimization

1: **Input**: training dataset $\mathcal{D}_{\text{tr}}$, regularization parameter λ, batch size s, learning-rate $\eta > 0$, exclusion rate α
2: Randomly initialize the deep neural network f_0
3: **for** $t \leftarrow 0 : T-1$ **do**
4: Randomly sample batch $\mathcal{B}_t \subset \mathcal{D}_{\text{tr}}$
5: Compute $M_{f_t}(X_i, \widetilde{Y}_i)$ on $\mathcal{B}_t$
6: Compute batch-wise quantile $\widehat{\tau}^s_\alpha(f_t)$ on $\mathcal{B}_t$
7: Compute classification loss $\widehat{\mathcal{L}}_{\text{cl}}(f_t)$ on $\mathcal{B}_t$
8: Compute conformal margin risk $\widehat{\mathcal{L}}_{\text{cr}}(f_t)$ on $\mathcal{B}_t$
9: $f_{t+1} \leftarrow f_t - \eta \nabla_f \big(\widehat{\mathcal{L}}_{\text{cl}}(f_t) + \lambda \widehat{\mathcal{L}}_{\text{cr}}(f_t)\big)$
10: **end for**
11: **Output**: the trained model f_T

Algorithm 1 summarizes the CMRM training procedure. The algorithm follows a standard stochastic optimization loop. For each iteration t, we first compute the confidence margins on batch $\mathcal{B}_t$ (Line 5) and then estimate the batch-wise conformal quantile threshold (Line 6). Next, we compute both the classification loss (Line 7) and the conformal margin risk (Line 8). The final update (Line 9) jointly minimizes both terms. Importantly, CMRM requires no change to model architecture or optimization, introducing only a quantile-calibrated regularizer that is compatible with arbitrary classifiers and loss functions.

Theoretical Analysis. Building on the optimization procedure, we analyze the effect of using batch-level estimates and the corresponding learning bound. Proposition 1 shows that the batch-level quantile in $\widehat{\mathcal{L}}^s_{\text{cr}}(f)$ concentrates around the population quantile at rate $\tilde{O}(1/\sqrt{s})$ where s is the batch size. Theorem 1 further bounds the gap between empirical margin risk on noisy data and population risk on clean data. Together, these results characterize the robustness of CMRM under arbitrary label noise with minimal assumptions.

We start with the definition of population quantile

$\tau_\alpha(f)$ on the noisy data distribution:

$$\tau_\alpha(f) = \min\Big\{t : \mathbb{P}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}[M_f(X,\widetilde{Y}) \le t] \ge \alpha\Big\}.$$

Next, we analyze the gap between the batch-level quantile and population one.

Proposition 1 (Gap between τ_α and $\widehat{\tau}^s_\alpha$)**.** *Denote by $G(t)$ the cumulative distribution function (CDF) of $M_f(X,\widetilde{Y})$ under the noisy distribution $\mathcal{P}_{noisy}$. Assume that $G(t)$ is continuously differentiable in a neighborhood of $\tau_\alpha(f)$ with density $g(t)$ and $g(\tau_\alpha(f)) > 0$. Then, for any $\delta \in (0,1)$, we have:*

$$\mathbb{P}\left(|\tau_\alpha(f) - \widehat{\tau}^s_\alpha(f)| \le \tilde{O}\Big(\frac{1}{\sqrt{s}}\Big)\right) \ge 1-\delta,$$

where $\tilde{O}$ hides the logarithmic factors.

Remark 1. Proposition 1 shows that the batch-level quantile $\widehat{\tau}^s_\alpha(f)$ closely approximates the population quantile $\tau_\alpha(f)$, with an error of order $\tilde{O}(1/\sqrt{s})$, which quantifies the statistical accuracy of the threshold estimation step in CMRM. This relies on a mild regularity assumption on the CDF (smoothness and positive density), standard in quantile estimation theory (e.g., Serfling (2009); Van der Vaart (2000)). We empirically verify these assumptions in Figure 2(c).

Next, we analyze the learning bound. Instead of evaluating the unconditional classification risk, CMRM focuses on the retained high-margin region where the model is confident. This reflects the mechanism of the method: low-margin samples—more likely to be corrupted under label noise—are suppressed during training, while high-margin samples dominate the learning signal. Accordingly, the appropriate clean-distribution objective is the conditional margin risk restricted to this retained region, analogous in spirit to Conditional Value at Risk (CVaR)-style tail-risk analyses in the robustness literature.

To this end, we first define the conformal margin risk on the clean distribution as:

$$\mathcal{L}_{\text{cr}}(f) = -\mathbb{E}_{(X,Y)\sim\mathcal{P}}\left[M_f(X,Y) \mid M_f(X,Y) \ge \tau_\alpha(f)\right].$$

We define its empirical Rademacher complexity as:

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) := \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\, f(X_i,\widetilde{Y}_i)\right],$$

where $\{\sigma_i\}_{i=1}^n$ are Rademacher variables. The following theorem analyzes the learning bound of CMRM.

Theorem 1 (Learning bound)**.** *Suppose the assumptions in Proposition 1 hold. Define δ_w as the average Wasserstein-1 distance between the noisy and clean label distributions conditional on X, where $\delta_w = \mathbb{E}_{X\sim\mathcal{P}(X)}\Big[W_1\big(\mathcal{P}_{noisy}(\cdot \mid X),\ \mathcal{P}(\cdot \mid X)\big)\Big]$. Then the learning bound of CMRM is:*

$$\mathcal{L}_{cr}(f) - \widehat{\mathcal{L}}^s_{cr}(f) \le \tilde{O}\Big(\widehat{\mathfrak{R}}_n(\mathcal{F}) + \frac{1}{\sqrt{s}} + \delta_w + \alpha + \texttt{temp}\Big).$$

Remark 2. Theorem 1 characterizes how the surrogate objective optimized by CMRM on noisy data relates to the clean retained-region margin risk, and clearly decomposes the generalization gap of CMRM into three terms reflecting its core components: quantile estimation error $\tilde{O}(1/\sqrt{s})$, function class complexity $\widehat{\mathfrak{R}}_n(\mathcal{F})$, and distribution shift δ_w. These correspond to the use of batchwise quantiles, empirical-to-population approximation on noisy data, and the mismatch between noisy and clean label distributions, respectively. Together, they provide an assumption-light characterization of CMRM's robustness under arbitrary label noise.

5 EXPERIMENTS AND RESULTS

5.1 Multi-class Classification Experiments

Datasets. We evaluate CMRM under both synthetic and real-world noisy supervision. For *synthetic noise*, we conduct experiments on CIFAR-100 (Krizhevsky et al., 2009), mini-ImageNet (Vinyals et al., 2016), and Food-101 (Bossard et al., 2014). CIFAR-100 is corrupted with class-conditional label noise, where each label is randomly replaced with another label within the same coarse superclass, following the protocol of (Yao et al., 2023). For mini-ImageNet and Food-101, we apply symmetric label flips following (Jiang et al., 2020). We vary the noise rate from $\{0\%, 5\%, 10\%, 20\%, 30\%, 40\%\}$, where 0% corresponds to the clean-label setting; additional implementation details are provided in Appendix B.1. For *real-world noise*, we evaluate on CIFAR-10N and its four variants, as well as CIFAR-100N (Wei et al., 2022), whose labels were collected through human annotation. The noise rates of these datasets range from 9.03% to 40.20%, covering a broad spectrum of labeling quality.

Baselines and ML Models. For synthetic noise experiments, we compare CMRM against four representative families of methods: **(i)** cross-entropy (CE), the standard objective; **(ii)** Focal loss (Lin et al., 2017), a standard robustness-oriented variant; **(iii)** LDAM (Cao et al., 2019), a margin-based objective; and **(iv)** GCE (Zhang and Sabuncu, 2018), a general-purpose noise-robust loss that remains a common baseline in the recent learning from noisy labels literature (Englesson and Azizpour, 2024; Nguyen et al.,

Detailed statistics and descriptions of CIFAR-10N and CIFAR-100N are summarized at http://noisylabels.com.

Dataset	Metric	CE		Focal		LDAM		GCE	
		Base	+CMRM	Base	+CMRM	Base	+CMRM	Base	+CMRM
CIFAR-100	ACC (%) ↑	65.16	**66.32** (+1.16)	64.42	**65.39** (+0.97)	59.63	**61.12** (+1.49)	62.17	**63.65** (+1.48)
	M.APSS ↓	6.67	**6.52** (−2.25%)	6.89	**6.61** (−4.06%)	17.85	**17.67** (−0.78%)	7.70	**6.28** (−18.44%)
mini-ImageNet	ACC (%) ↑	57.42	**59.41** (+1.99)	55.54	**57.93** (+2.39)	56.60	**56.62** (+0.02)	55.16	**55.51** (+0.35)
	M.APSS ↓	7.40	**7.04** (−4.86%)	7.67	**7.28** (−5.08%)	13.07	**12.88** (−1.45%)	9.90	**9.59** (−3.13%)
Food-101	ACC (%) ↑	56.21	**58.48** (+2.27)	56.35	**58.92** (+2.57)	55.49	**56.42** (+0.93)	55.66	**59.05** (+3.39)
	M.APSS ↓	7.93	**6.96** (−12.23%)	7.76	**6.89** (−11.21%)	11.94	**11.53** (−3.43%)	8.41	**6.93** (−20.44%)

Table 2: **Top-1 accuracy (%) and marginal average prediction set size (M.APSS ↓) on multi-class datasets corrupted by synthetic noise with noise rate** 20%**.** Each Base objective is paired with its +CMRM counterpart; the better value within each pair is in **bold**. Numbers in parentheses indicate the relative change (%): + denotes accuracy improvement, and − denotes M.APSS reduction compared to the corresponding Base objective. On average across all datasets and objectives, CMRM improves accuracy by 1.58 and reduces M.APSS by 7.28%.

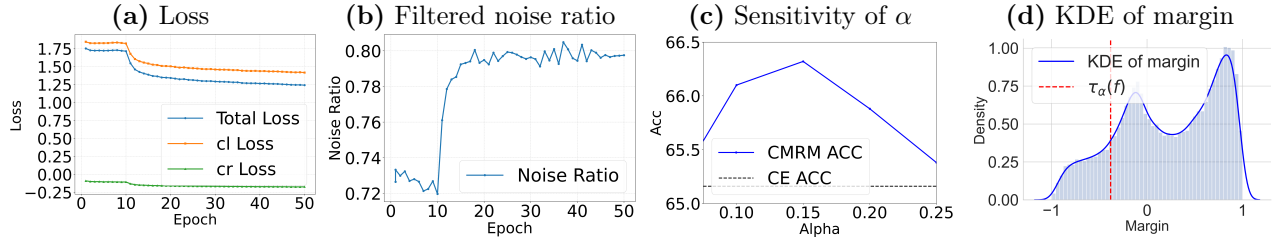

Figure 2: **Justification experiments for multi-class classification** on CIFAR-100 with 20% synthetic label noise. Subfigure **(a)** shows the training dynamics of total loss (Total), classification loss (cl), and CMRM loss (cr) over epochs. CMRM exhibits stable and monotonic convergence alongside standard loss components. Subfigure **(b)** reports the ratio of noisy samples among those filtered (with soft weight < 0.5) out by CMRM ($\alpha = 0.15$) at each epoch, demonstrating that CMRM consistently suppresses noisy examples by excluding low-margin samples during training. Subfigure **(c)** examines the sensitivity of α, showing that CMRM maintains higher accuracy than CE across a range of α values, indicating robustness to hyperparameter α. Subfigure **(d)** depicts the kernel density estimate (KDE) of the margin distribution, with the vertical dashed line indicating the estimated $\tau_\alpha(f)$ with $\alpha = 0.15$. The density curve is smooth and strictly positive around $\tau_\alpha(f)$, supporting the differentiability and positive-density assumption in Proposition 1.

Method	CIFAR-10N (worst)		CIFAR-100N	
	ACC(%)	M.APSS	ACC(%)	M.APSS
NI-ERM	95.71	0.93	83.17	1.49
NI-ERM+CMRM	**97.19** (+1.48)	**0.91** (−2.15%)	**83.95** (+0.78)	**1.29** (−13.42%)

Table 3: **Top-1 accuracy (%) and marginal average prediction set size (M.APSS ↓) on CIFAR-10N and CIFAR-100N corrupted by human annotation noise.** Numbers in parentheses indicate the relative change: + denotes accuracy improvement and − denotes M.APSS reduction. CMRM consistently improves accuracy and reduces uncertainty across CIFAR-N variants, with the largest gains observed on CIFAR-10N Worst and CIFAR-100N. Full results on all CIFAR-10N variants (Aggre, Rand1–3, Worst) and CIFAR-100N are provided in Appendix B.

2024). All methods use a ResNet-20 (He et al., 2016) backbone with standard data augmentation. For real-world noise experiments, we benchmark CMRM against NI-ERM (Zhu et al., 2024), a recent state-of-the-art approach that achieves strong performance on CIFAR-10N and CIFAR-100N by training a linear classifier on frozen DINOv2 (Oquab et al., 2023) features. All methods share the same protocol. For CMRM, We set `temp` $= 1.0$ in all experiments and select $\lambda \in \{0.05, \cdots, 0.25\}$ and $\alpha \in \{0.05, \cdots, 0.25\}$ via grid search; See Appendix B.1 for full details.

Evaluation Metrics. We evaluate models using marginal Top-1 accuracy and marginal average prediction set size (M.APSS). To assess predictive uncertainty, we adopt CP and define prediction sets and average prediction set size (APSS) (Romano et al., 2020; Angelopoulos et al., 2022). For each input, CP outputs a prediction set containing the true label with high probability, controlled by a target coverage rate of 0.9. For multi-class settings, M.APSS denotes the marginal variant of APSS, i.e., averaged uniformly over all test examples. It also complements top-1 accuracy by capturing how sharply the model distinguishes plausible labels under noisy supervision. We also report class-conditional variants (PC APSS, NC APSS) for binary classification.

CMRM improves accuracy and reduces uncertainty under different types of noise. Table 2 reports results on CIFAR-100, mini-ImageNet, and Food-101 corrupted by synthetic label noise. For each objective (CE, Focal, LDAM, GCE), we compare the

Method	Evaluation Metric							
	AUROC (↑)	AUPRC (↑)	FNR (↓)	FPR (↓)	ACC (↑)	M.APSS (↓)	PC APSS (↓)	NC APSS (↓)
LR	0.784	0.885	**0.073**	0.571	0.802	1.223	1.154	1.432
LR + CMRM	**0.852** (+0.068)	**0.925** (+0.04)	0.082 (+0.009)	**0.422** (−0.149)	**0.833** (0.031)	**1.209** (−1.15%)	**1.109** (−3.9%)	**1.308** (−8.66%)
Focal	0.809	0.890	0.136	0.388	0.801	1.257	1.224	1.356
Focal + CMRM	**0.872** (+0.063)	**0.942** (+0.052)	**0.128** (−0.008)	**0.324** (−0.064)	**0.823** (0.022)	**1.221** −2.87%	**1.148** −6.21%	**1.295** −4.5%
SVM	0.808	0.925	**0.029**	0.807	0.776	1.276	1.370	1.512
SVM + CMRM	**0.847** (+3.9%)	**0.937** (+1.2%)	0.048 (+1.9%)	**0.585** (−22.2%)	**0.817** (+4.1%)	**1.199** (−6.03%)	**1.322** (−3.5%)	**1.343** (−11.17%)
GCE	0.819	0.904	**0.119**	0.424	**0.804**	1.286	**1.176**	1.396
GCE + CMRM	**0.846** (+0.027)	**0.928** (+0.024)	0.172 (+0.053)	**0.286** (−0.138)	0.800 (−0.004)	**1.273** (−1.01%)	1.207 (+2.63%)	**1.340** (−4.01%)

Table 4: **Results for binary classification on Adult** with 20% label noise. We report ranking (AUROC, AUPRC), error rates (FNR, FPR), accuracy (ACC), and uncertainty (M.APSS, PC APSS, NC APSS). ↑ indicates higher is better; ↓ lower is better. Best results for each base method (LR, Focal, SVM, GCE) are in **bold**. Numbers in parentheses indicate absolute and relative changes (%) of CMRM, where + denotes increase and − denotes decrease. On average, CMRM improves AUROC by 0.049, AUPRC by 0.032, and ACC by 0.023, while reducing FPR by 0.143 and slightly increasing FNR by 0.018. For uncertainty-aware metrics, CMRM reduces M.APSS, PC APSS, and NC APSS by 2.77%, 2.75%, and 7.09% on average, respectively. Results on Email and Credit show similar trends and are provided in Appendix C.2.

Base model with its +CMRM counterpart. On average across all datasets and objectives, CMRM improves accuracy by 1.58 and reduces M.APSS by 7.28%. The improvements are most pronounced for CE, Focal, and GCE, with accuracy gains of up to 3.39 and substantial uncertainty reduction. Even when combined with LDAM, which already encourages margin separation, CMRM consistently yields additional accuracy improvements without increasing uncertainty. Table 3 summarizes results on CIFAR-10N and CIFAR-100N with human-annotated label noise. CMRM consistently outperforms NI-ERM across all CIFAR-N variants, achieving the largest gains on the most challenging settings (CIFAR-10N Worst and CIFAR-100N). Complete results on all CIFAR-10N variants (Aggre, Rand1–3, Worst) are provided in Appendix B. These findings clearly demonstrate that CMRM improves accuracy and reduces uncertainty under both synthetic and real-world noisy supervision.

CMRM loss convergence results. Figure 2(a) shows the training dynamics of the classification loss and the CMRM regularization loss. Both components decrease steadily and stabilize as training progresses, indicating smooth joint optimization. The CMRM term integrates with standard objectives and does not introduce instability or slow down of convergence, demonstrating that CMRM can be efficiently optimized.

CMRM filters out noisy samples during training. Figure 2(b) shows the fraction of noisy samples among those excluded by CMRM (with soft weight < 0.5) at each epoch. This proportion rapidly increases during the early training phase and stabilizes above 78%, indicating that CMRM consistently identifies and filters out mislabeled examples via its margin-based thresholding mechanism.

CMRM is robust to the choice of hyperparameter α. Figure 2(c) examines the sensitivity of CMRM to the hyperparameter α. Across a range of α values, CMRM consistently achieves higher accuracy than CE, indicating that its performance is robust to the choice of α and does not rely on careful hyperparameter tuning. Notably, CMRM performs well even when α does not match the true noise rate, suggesting that exact noise rate knowledge is unnecessary. In practice, $\alpha \in [0.1, 0.2]$ consistently yields strong results across all settings tested.

Assumptions in Proposition 1 are empirically valid. Figure 2(d) presents the kernel density estimate (KDE) of the margin distribution, with the vertical dashed line indicating the estimated $\tau_\alpha(f)$. The density curve is smooth and strictly positive in the neighborhood of $\tau_\alpha(f)$, supporting the differentiability and positive-density assumption in Proposition 1.

CMRM incurs no penalty on clean labels. Notably, CMRM also improves or matches accuracy at 0% noise across all objectives (Table 6 in Appendix B.2), confirming that the regularizer incurs no penalty even when labels are clean.

5.2 Binary Classification Experiments

Experiment Setup. We also evaluate the binary variant of CMRM on three datasets: Email (Hopkins et al., 1999), Credit (Quinlan, 1987), and Adult (Becker and Kohavi, 1996). To simulate label noise, we randomly flip 20% of the training labels while keeping the test labels clean. Additional implementation details are provided in Appendix C.1. Baselines include logistic regression (LR), focal loss, support vector machines (SVM) with hinge loss as a margin-based method, and GCE. LR, Focal, and GCE models use a two-layer MLP, while SVM employs a linear kernel with default regularization. We report AUC-ROC and PR-AUC to assess ranking quality, FPR and FNR to capture class-specific error tendencies, and Accuracy as a general indicator. We also measure the predictive uncertainty using

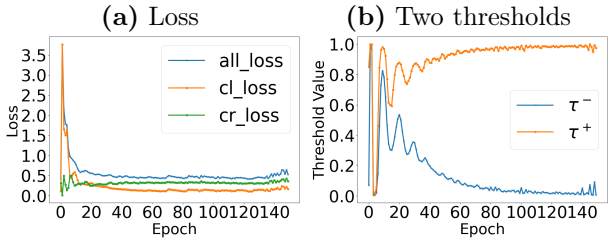

Figure 3: **Training dynamics of LR+CMRM for binary classification** on the Email dataset. Subfigure **(a)** Training dynamics of total loss (all loss), classification loss (cl loss), and CMRM loss (cr loss) over epochs. CMRM exhibits stable and monotonic convergence alongside standard loss components. Subfigure **(b)** τ^- (negative class threshold) and τ^+ (positive class threshold) of LR+CMRM during training. The separation between the thresholds increases, indicating that CMRM actively maximizes the margin between positive and negative classes.

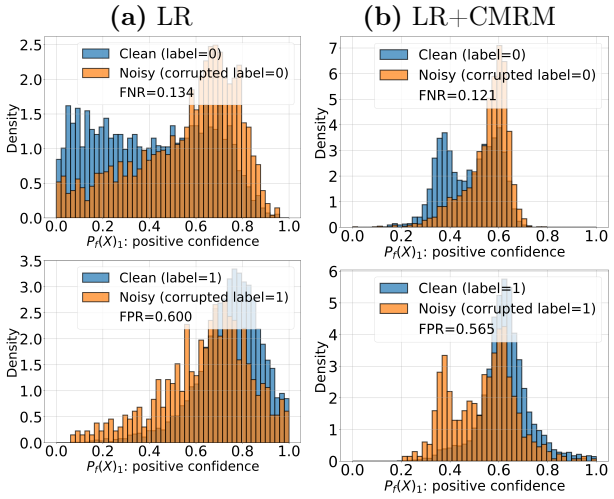

Figure 4: **Histograms of positive confidence distributions for clean (blue) and noisy (orange) samples** on the Credit dataset with 20% label noise. The top and bottom rows correspond to samples with observed labels $\widetilde{Y} = 0$ (negative) and $\widetilde{Y} = 1$ (positive), respectively. Distributions are obtained using LR (left) and LR+CMRM (right). CMRM induces a clearer separation between clean and noisy confidence distributions for both classes.

APSS, including marginal (M.APSS), positive-class (PC APSS), and negative-class (NC APSS) variants. Larger values of AUC-ROC, PR-AUC, and Accuracy indicate better performance (↑), whereas smaller values of FPR, FNR, and APSS metrics (M.APSS, PC APSS, NC APSS) are preferred (↓).

CMRM improves robustness for binary classification. Table 4 reports results on the Adult dataset with 20% label noise. CMRM consistently improves ranking performance, with AUROC and AUPRC increasing by 0.049 and 0.032 on average, respectively, indicating that it enables models to better separate classes under noisy supervision. These gains are accompanied by improvements in classification metrics, with Accuracy increasing by 0.023 on average and FPR decreasing by 0.143, at the cost of a modest increase in FNR (+0.018). Finally, CMRM reduces predictive uncertainty, as reflected by lower M.APSS, PC APSS, and NC APSS (average reduction of 2.77%, 2.75%, and 7.1%, respectively), indicating sharper and more discriminative predictions. Overall, these improvements demonstrate that CMRM enhances robustness in binary classification under label noise. Similar trends are observed on the Email and Credit datasets (see Appendix C.2 for complete results).

CMRM optimization dynamics in binary classification. Figure 3 examines the training behavior of LR + CMRM on the Email dataset with 20% label noise. Subfigure (a) shows that both the classification and CMRM regularization losses decrease steadily and stabilize, indicating that the joint objective can be optimized smoothly on binary classification data. Subfigure (b) tracks the evolution of the class-conditional thresholds τ^+ and τ^- during training. τ^- steadily increases while τ^+ decreases, leading to a widening gap between them. This growing separation reflects that CMRM could actively enlarge the margin between positive and negative classes.

CMRM separates clean and noisy supervision. Figure 4 shows the effect of CMRM on positive-class confidence distributions under label noise. With LR (Figure 4a), the distributions of clean and noisy samples overlap substantially for both observed positive and negative label groups, indicating limited ability to distinguish reliable labels from corrupted labels. In contrast, LR+CMRM (Figure 4b) yields a clearer separation: clean samples concentrate in high-confidence regions for their true class, while noisy samples are pushed toward lower-confidence regions, improving their distinguishability in the confidence space.

6 CONCLUSION

We propose CMRM, a simple envelope framework for robust learning under label noise, requiring only mild regularity of the margin distribution rather than noise models, clean data or auxiliary resources. CMRM uses a batch-wise conformal quantile on confidence margins to focus training on reliable samples while suppressing likely corrupted ones. Theoretically, we establish a learning bound under arbitrary label noise. Empirically, CMRM integrates smoothly into standard pipelines and consistently improves accuracy and robustness across binary and multi-class benchmarks. As a single regularization term with no architectural cost, CMRM is a natural default when training with noisy labels.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the in part support by the USDA-NIFA funded AgAID Institute award 2021-67021-35344, and NSF grants CNS-2312125, IIS-2443828, DUE-2519063. The views expressed are those of the authors and do not reflect the official policy or position of the USDA-NIFA and NSF.

References

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR, 2022.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Devansh Arpit, Stanislaw Jastrzebski, 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 *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Pedro Celard, Eva Lorenzo Iglesias, José Manuel Sorribes-Fdez, Rubén Romero, A Seara Vieira, and Lourdes Borrajo. A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, 35(3):2291–2323, 2023.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

Richard M Dudley. *Real analysis and probability*. Chapman and Hall/CRC, 2018.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in neural information processing systems*, 35:22380–22395, 2022.

Gamaleldin Fathy Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Erik Englesson and Hossein Azizpour. Robust classification via regression for learning with noisy labels. In *ICLR 2024-The Twelfth International Conference on Learning Representations, Messe Wien Exhibition and Congress Center, Vienna, Austria, May 7-11t, 2024*, 2024.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25 (5):845–869, 2013.

Subhankar Ghosh, Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Improving Uncertainty Quantification of Deep Classifiers via Neighborhood Conformal Prediction: Novel Algorithm and Theoretical Analysis. In *Proc. of AAAI Conf.*, pages 7722–7730, 2023a.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically Robust Conformal Prediction. In *UAI Conf.*, volume 216 of *Proc. of Machine Learning Research*, pages 681–690. PMLR, 2023b.

Shivani Gupta and Atul Gupta. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, 161:466–474, 2019.

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. *Advances in neural information processing systems*, 31, 2018.

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.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018.

Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. Spambase. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C53G6X.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya R. Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International conference on machine learning*, pages 4804–4815. PMLR, 2020.

Justin M Johnson and Taghi M Khoshgoftaar. A survey on classifying big data with label noise. *ACM Journal of Data and Information Quality*, 14(4):1–43, 2022.

Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24137–24149, 2021.

Shayan Kiyani, George J Pappas, and Hamed Hassani. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37: 99519–99563, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Sijia Li, Yong Fan, Yue Ma, and Ya Pan. Evaluation of dataset distribution and label quality for autonomous driving system. In *2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 196–200. IEEE, 2021a.

Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *International conference on machine learning*, pages 6403–6413. PMLR, 2021b.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Yexiong Lin, Yu Yao, and Tongliang Liu. Learning the latent causal structure for modeling label noise. *Advances in Neural Information Processing Systems*, 37:120549–120577, 2024.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Tri Nguyen, Shahana Ibrahim, and Xiao Fu. Noisy label learning with instance-dependent outliers: Identifiability via crowd wisdom. *Advances in Neural Information Processing Systems*, 37:97261–97298, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020.

J. R. Quinlan. Credit Approval. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5FS30.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33: 3581–3591, 2020.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Hooman Shahrokhi, Devjeet Raj Roy, Yan Yan, Venera Arnaoudova, and Janaradhan Rao Doppa. Conformal Prediction Sets for Deep Generative Models via Reduction to Conformal Regression, 2025. URL https://arxiv.org/abs/2503.10512.

Jialin Shi, Kailai Zhang, Chenyi Guo, Youquan Yang, Yali Xu, and Ji Wu. A survey of label-noise deep learning for medical image analysis. *Medical image analysis*, 95:103166, 2024a.

Yuanjie Shi, Subhankar Ghosh, Taha Belkhouja, Jana Doppa, and Yan Yan. Conformal Prediction for Classwise Coverage via Augmented Label Rank Calibration. In *Advances in Neural Information Processing Sys. (NeurIPS)*, 2024b.

Yuanjie Shi, Hooman Shahrokhi, Xuesong Jia, Xiongzhi Chen, Janardhan Rao Doppa, and Yan Yan. Direct prediction set minimization via bilevel conformal classifier training. *arXiv preprint arXiv:2506.06599*, 2025.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34 (11):8135–8153, 2022.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil Van Der Aalst, and Oliver Hinz. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101, 2023.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Donna Tjandra and Jenna Wiens. Leveraging an alignment set in tackling instance-dependent label noise. In *Conference on Health, Inference, and Learning*, pages 477–497. PMLR, 2023.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Haoyu Wang, Zhuo Huang, Zhiwei Lin, and Tongliang Liu. Noisegpt: Label noise detection and rectification through probability curvature. *Advances in Neural Information Processing Systems*, 37:120159–120183, 2024.

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. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32, 2019.

Jiangchao Yao, Bo Han, Zhihan Zhou, Ya Zhang, and Ivor W Tsang. Latent class-conditional noise model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9964–9980, 2023.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33:7260–7271, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *Proceedings of the National Academy of Sciences*, 118(3), 2021.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

Mingyuan Zhang and Shivani Agarwal. Multiclass learning from noisy labels for non-decomposable performance measures. In *International Conference on Artificial Intelligence and Statistics*, pages 2170–2178. PMLR, 2024.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99, 2024.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77, 2006.

Yilun Zhu, Jianxin Zhang, Aditya Gangrade, and Clay Scott. Label noise: Ignorance is bliss. *Advances in Neural Information Processing Systems*, 37:116575–116616, 2024.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] We provide it in Section 4.2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We provide it in Section 4.2.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] We provide it in Section 4.2.
   (b) Complete proofs of all theoretical results. [Yes] We provide it in Appendix A.
   (c) Clear explanations of any assumptions. [Yes] We provide it in Section 4.2.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] we provide it in the supplemental material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] we provide it in Appendix B and C.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] we provide it in Appendix B and C.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] we cite it in 5.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] we provide it in the supplemental material.
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable]

A TECHNICAL PROOFS

A.1 Technical Proofs for Proposition 1

Proposition (Re) 1 (Gap between τ_α and $\widehat{\tau}_\alpha^s$)**.** *Denote by $G(t)$ the cumulative distribution function (CDF) of $M_f(X,\widetilde{Y})$ under the noisy distribution $\mathcal{P}_{noisy}$. Assume that $G(t)$ is continuously differentiable in a neighborhood of $\tau_\alpha(f)$ with density $g(t)$ and $g(\tau_\alpha(f)) > 0$. Then, for any $\delta \in (0,1)$, we have:*

$$\mathbb{P}\Big(|\tau_\alpha(f) - \widehat{\tau}_\alpha^s(f)| \le \tilde{O}\Big(\frac{1}{\sqrt{s}}\Big)\Big) \ge 1-\delta,$$

where $\tilde{O}$ hides the logarithmic factors.

Proof. of Proposition 1.

Before proving Proposition 1, we first define $\widehat{G}^s(t) = \frac{1}{s}\sum \mathbb{1}[M_f(X_i,\widetilde{Y}_i) \le t]$ as the empirical CDF of $M_f(X,\widetilde{Y})$ with s samples.

By the definition, $\tau_\alpha(f) = G^{-1}(\alpha)$, where $\widehat{\tau}_\alpha^s(f) := \inf\{t : \widehat{G}^s(t) \ge \alpha\}$.

Step 1. By the definition of $\widehat{\tau}_\alpha^s$, we have $\widehat{G}^s(\widehat{\tau}_\alpha^s) \ge \alpha$. Moreover, since $\widehat{G}^s$ is a step function with jumps of size $1/s$, the value at the first crossing satisfies

$$\alpha \le \widehat{G}^s(\widehat{\tau}_\alpha^s) \le \alpha + \frac{1}{s}. \tag{7}$$

Equivalently, $\widehat{\tau}_\alpha^s$ is an order statistic and the ECDF at an order statistic lies in an interval of width $1/s$.

Step 2. By the DKW inequality, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{t\in\mathbb{R}}\big|\widehat{G}^s(t) - G(t)\big| \le \varepsilon_s(\delta) := \sqrt{\frac{\log(2/\delta)}{2s}}. \tag{8}$$

On the event (8), evaluating at the *same* point $t = \widehat{\tau}_\alpha^s$ gives

$$\big|G(\widehat{\tau}_\alpha^s) - \widehat{G}^s(\widehat{\tau}_\alpha^s)\big| \le \varepsilon_s(\delta).$$

Combining with (7), we obtain the sandwich

$$\alpha - \varepsilon_s(\delta) \le G(\widehat{\tau}_\alpha^s) \le \alpha + \frac{1}{s} + \varepsilon_s(\delta). \tag{9}$$

Step 3. Let $U = (\tau_\alpha - \eta, \tau_\alpha + \eta)$ and assume $\inf_{t\in U} g(t) \ge g_0 > 0$. Then G is strictly increasing on U, hence invertible on $G(U)$, and its inverse is $(1/g_0)$-Lipschitz on $G(U)$: for any $u_1, u_2 \in G(U)$,

$$|G^{-1}(u_1) - G^{-1}(u_2)| \le \frac{1}{g_0}|u_1 - u_2|. \tag{10}$$

Now we show that on the event (8) and for sufficiently large s (or more explicitly, whenever $\varepsilon_s(\delta) + 1/s \le g_0\eta$), the probability interval in (9) lies inside $G(U)$, which implies $\widehat{\tau}_\alpha^s \in U$ and legitimizes applying (10). Indeed, since $G(\tau_\alpha) = \alpha$ and $g \ge g_0$ on U, we have

$$G(\tau_\alpha + \eta) - \alpha \ge g_0\eta, \qquad \alpha - G(\tau_\alpha - \eta) \ge g_0\eta.$$

Thus if $\varepsilon_s(\delta) + 1/s \le g_0\eta$, then

$$\alpha - \varepsilon_s(\delta) \ge G(\tau_\alpha - \eta), \qquad \alpha + \frac{1}{s} + \varepsilon_s(\delta) \le G(\tau_\alpha + \eta),$$

so (9) implies $G(\widehat{\tau}_\alpha^s) \in G(U)$ and hence $\widehat{\tau}_\alpha^s \in U$.

Step 4. On the event (8) and under $\varepsilon_s(\delta) + 1/s \le g_0\eta$, we can apply (10) with $u_1 = G(\widehat{\tau}_\alpha^s)$ and $u_2 = \alpha = G(\tau_\alpha)$:

$$|\widehat{\tau}_\alpha^s - \tau_\alpha| = \left|G^{-1}(G(\widehat{\tau}_\alpha^s)) - G^{-1}(\alpha)\right| \le \frac{1}{g_0}\,|G(\widehat{\tau}_\alpha^s) - \alpha|.$$

Using (9), we obtain

$$|G(\widehat{\tau}_\alpha^s) - \alpha| \le \varepsilon_s(\delta) + \frac{1}{s},$$

hence

$$|\widehat{\tau}_\alpha^s - \tau_\alpha| \le \frac{1}{g_0}\Big(\varepsilon_s(\delta) + \frac{1}{s}\Big) = \frac{1}{g_0}\Big(\sqrt{\frac{\log(2/\delta)}{2s}} + \frac{1}{s}\Big).$$

Finally, since (8) holds with probability at least $1-\delta$, the desired high-probability bound follows. □

A.2 Technical Proofs for Theorem 1

Theorem (Re) 1 (Learning bound)**.** *(Theorem 1 restated.) Suppose the assumptions in Proposition 1 hold. Define δ_w as the average Wasserstein-1 distance between the noisy and clean label distributions conditional on X, where $\delta_w = \mathbb{E}_{X\sim\mathcal{P}(X)}\Big[W_1\big(\mathcal{P}_{noisy}(\cdot \mid X),\ \mathcal{P}(\cdot \mid X)\big)\Big]$. Then the learning bound of CMRM is:*

$$\mathcal{L}_{cr}(f) - \widehat{\mathcal{L}}_{cr}^s(f) \le \tilde{O}\Big(\widehat{\mathfrak{R}}_n(\mathcal{F}) + \frac{1}{\sqrt{s}} + \delta_w + \alpha + \texttt{temp}\Big).$$

Proof. of Theorem 1.

Before proving Theorem 1, we first define conformal margin risk on the noisy distribution as:

$$\mathcal{L}_{\mathrm{cr}}^{\mathrm{noisy}}(f) = -\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\mathrm{noisy}}}\Big[M_f(X,\widetilde{Y})\cdot\mathbb{1}[M_f(X,\widetilde{Y}) \ge \tau_\alpha(f)]\Big].$$

and its surrogate risk on the noisy distribution, where the indicator function $\mathbb{1}[x \ge y]$ is replaced by the Sigmoid function $\widetilde{\mathbb{1}}[x \ge y] = 1/(1+\exp(-(x-y)/\texttt{temp}))$, as:

$$\widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f) = -\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\mathrm{noisy}}}\Big[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \ge \tau_\alpha(f)]\Big].$$

Recall that:

$$\widehat{\mathcal{L}}_{\mathrm{cr}}(f) = \frac{1}{n}\sum_{i=1}^{n} -M_f(X_i,\widetilde{Y}_i)\cdot\widetilde{\mathbb{1}}[M_f(X_i,\widetilde{Y}_i) \ge \widehat{\tau}_\alpha(f)],$$

$$\widehat{\mathcal{L}}_{\mathrm{cr}}^s(f) = \mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[\sum_{i=1}^{n} - M_f(X_i,\widetilde{Y}_i)\cdot\widetilde{\mathbb{1}}[M_f(X_i,\widetilde{Y}_i) \ge \widehat{\tau}_\alpha^s(f)]\right],$$

and

$$\mathcal{L}_{\mathrm{cr}}(f) = -\mathbb{E}_{(X,Y)\sim\mathcal{P}}\left[M_f(X,Y) \mid M_f(X,Y) \ge \tau_\alpha(f)\right].$$

Then we show the following technical lemmas:

Lemma 1. *(Immediate results from Theorem 3.3 in (Mohri et al., 2018)) For any $f \in \mathcal{F}$, the following inequality holds with high probability:*

$$\mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{noisy}}\Big[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{1}}\big[M_f(X,\widetilde{Y}) \ge \widehat{\tau}_\alpha^s(f)\big]\Big]\right] \le \widetilde{\mathcal{L}}_{cr}^s(f) + \tilde{O}\Big(\widehat{\mathfrak{R}}_n(\mathcal{F}) + \frac{1}{\sqrt{n}}\Big).$$

Lemma 2. *Suppose the assumptions in Proposition 1 hold. Then, the following inequality holds:*

$$\left|\widetilde{\mathcal{L}}_{cr}^{noisy}(f) - \mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{noisy}}\left[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{I}}\big[M_f(X,\widetilde{Y})\geq\widehat{\tau}_\alpha^s(f)\big]\right]\right]\right| \leq \tilde{O}(1/\sqrt{s}).$$

Lemma 3. *Suppose the assumptions in Proposition 1 hold. Then, the following inequality holds:*

$$|\mathcal{L}_{cr}^{noisy}(f) - \widetilde{\mathcal{L}}_{cr}^{noisy}(f)| \leq O(\texttt{temp}).$$

Lemma 4. *Define δ_w as the average Wasserstein-1 distance between posteriors under the training and test domains, where $\delta_w = \mathbb{E}_{X\sim\mathcal{P}(X)}\Big[W_1\big(\mathbb{P}_{\mathcal{P}_{noisy}}(\cdot\mid X),\ \mathbb{P}_{\mathcal{P}}(\cdot\mid X)\big)\Big]$. Then the following inequality holds:*

$$\left|\mathcal{L}_{cr}(f) - \mathcal{L}_{cr}^{noisy}(f)\right| \ \leq\ 2\delta_w + \alpha.$$

Now we begin to prove Theorem 1:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{cr}}(f) - \widehat{\mathcal{L}}_{\mathrm{cr}}^s(f)\\
=&\underbrace{\mathcal{L}_{\mathrm{cr}}(f) - \mathcal{L}_{\mathrm{cr}}^{\mathrm{noisy}}(f)}_{2\delta_w+\alpha,\text{ Lemma 4}} + \underbrace{\mathcal{L}_{\mathrm{cr}}^{\mathrm{noisy}}(f) - \widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f)}_{O(\texttt{temp}),\text{ Lemma 3}}\\
&+\underbrace{\widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f) - \mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\mathrm{noisy}}}\left[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{I}}\big[M_f(X,\widetilde{Y})\geq\widehat{\tau}_\alpha^s(f)\big]\right]\right]}_{\tilde{O}(1/\sqrt{s},\text{ Lemma 2}}\\
&+\underbrace{\mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\mathrm{noisy}}}\left[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{I}}\big[M_f(X,\widetilde{Y})\geq\widehat{\tau}_\alpha^s(f)\big]\right]\right] - \widehat{\mathcal{L}}_{\mathrm{cr}}^s(f)}_{\tilde{O}\left(\widehat{\mathfrak{R}}_n(\mathcal{F})+\frac{1}{\sqrt{n}}\right),\text{ Lemma 1}}\\
\leq&2\delta_w + \alpha + O(\texttt{temp}) + \tilde{O}(1/\sqrt{s}) + \tilde{O}\Big(\widehat{\mathfrak{R}}_n(\mathcal{F}) + \frac{1}{\sqrt{n}}\Big)\\
\leq&\tilde{O}\Big(\widehat{\mathfrak{R}}_n(\mathcal{F}) + \frac{1}{\sqrt{s}} + \delta_w + \alpha + \texttt{temp}\Big),
\end{aligned}$$

where the first inequality is due to Lemma 1, 2, 3, 4, and the second inequality is due to $s \leq n$.

Therefore, we have:

$$\mathcal{L}_{\mathrm{cr}}(f) - \widehat{\mathcal{L}}_{\mathrm{cr}}^s(f) \leq \tilde{O}\Big(\widehat{\mathfrak{R}}_n(\mathcal{F}) + \frac{1}{\sqrt{s}} + \delta_w + \alpha + \texttt{temp}\Big).$$

□

A.3 Proofs for Technical Lemmas

A.3.1 Proofs for Lemma 2

Lemma (Re) 1. *(Lemma 2 restated.) Suppose the assumptions in Proposition 1 hold. Then, the following inequality holds:*

$$\left|\widetilde{\mathcal{L}}_{cr}^{noisy}(f) - \mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{noisy}}\left[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{I}}\big[M_f(X,\widetilde{Y})\geq\widehat{\tau}_\alpha^s(f)\big]\right]\right]\right| \leq \tilde{O}(1/\sqrt{s}).$$

Proof. of Lemma 2.

Before proving Lemma 2, we first show the following technical lemma:

Lemma 5. *(Lipschitz continuity of $\widetilde{\mathcal{L}}_{cr}^{noisy}$ in τ) Define the smoothed conformal margin risk on the noisy distribution as a function of the threshold τ:*

$$\widetilde{\mathcal{L}}_{cr}^{noisy}(f;\tau) := -\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{noisy}}[M_f(X,\widetilde{Y})\cdot\tilde{\mathbb{1}}[M_f(X,\widetilde{Y})\geq\tau]],$$

where the softened indicator is the Sigmoid function $\tilde{\mathbb{1}}(M_f\geq\tau) := \sigma(\frac{M_f(X,\widetilde{Y})-\tau}{\texttt{temp}})$, and $\sigma(u)=\frac{1}{1+\exp(-u)}$. Assume $|M_f(X,\widetilde{Y})|\leq 1$ almost surely and $\texttt{temp}>0$. Then, for any fixed f, the function $\tau\to\widetilde{\mathcal{L}}_{cr}^{noisy}(f;\tau)$ is L-Lipscthiz continuous with $L=1/(4\texttt{temp})$.

The proof of Lemma 5 is deferred to the end of this proof.

Now we begin to prove Lemma 2.

$$\begin{aligned}
&\left|\widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}(f)-\mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\left[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{1}}\big[M_f(X,\widetilde{Y})\geq\widehat{\tau}_\alpha^s(f)\big]\right]\right]\right|\\
=&\left|-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{1}}[M_f(X,\widetilde{Y})\geq\tau_\alpha(f)]-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\left[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{1}}\big[M_f(X,\widetilde{Y})\geq\widehat{\tau}_\alpha^s(f)\big]\right]\right|\\
\leq&\frac{1}{4\texttt{temp}}\cdot\left|\tau_\alpha(f)-\mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[\widehat{\tau}_\alpha^s(f)\right]\right|\\
\leq&\frac{1}{4\texttt{temp}}\cdot\tilde{O}(1/\sqrt{s})\\
\leq&\tilde{O}(1/\sqrt{s}),
\end{aligned}$$

where the first inequality is due to Lemma 5, the second inequality is due to Proposition 1, and the last inequality is due to the setting $\texttt{temp}=1.0$.

Therefore, we have:

$$\left|\widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}(f)-\mathop{\mathbb{E}}_{\widehat{\tau}_\alpha^s(f)\sim\mathcal{T}_\alpha(f)}\left[-\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\left[M_f(X,\widetilde{Y})\cdot\widetilde{\mathbb{1}}\big[M_f(X,\widetilde{Y})\geq\widehat{\tau}_\alpha^s(f)\big]\right]\right]\right|\leq\tilde{O}(1/\sqrt{s}).$$

□

Now we begin to prove Lemma 5.

Proof. (of Lemma 5)

Let $Z := M_f(X,\widetilde{Y}), u := \frac{Z-\tau}{\texttt{temp}}$.

We can rewrite the interested function as $\widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}(f;\tau)=-\mathbb{E}[Z\cdot\sigma(u)]$, where the expectation is taken over $(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}$.

First, we develop the differentiation of $\widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}$ w.r.t. τ by applying the chain rule as follows

$$\frac{\partial\widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}(f;\tau)}{\partial\tau}=-\mathbb{E}[Z\cdot\frac{\partial\sigma(u)}{\partial\tau}]. \tag{11}$$

Since $u=(Z-\tau)/\texttt{temp}$, we have $\frac{\partial u}{\partial\tau}=-1/\texttt{temp}$.

The derivative of the Sigmoid function is

$$\frac{\partial\sigma(u)}{\partial u}=\frac{\exp(-u)}{(1+\exp(-u))^2}.$$

Thus, by using the chain rule, we have

$$\frac{\partial\sigma(u)}{\partial\tau}=\frac{\partial\sigma(u)}{\partial u}\cdot\frac{\partial u}{\partial\tau}=-\frac{1}{\texttt{temp}}\cdot\frac{\exp(-u)}{(1+\exp(-u))^2}.$$

Substituting the above equality back into Equation (11), we have

$$\frac{\partial \widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}}{\partial \tau}(f;\tau) = -\mathbb{E}[Z \cdot (-\frac{1}{\texttt{temp}} \cdot \frac{\exp(-u)}{(1+\exp(-u))^2})] = \frac{1}{\texttt{temp}} \cdot \mathbb{E}[Z \frac{\exp(-u)}{(1+\exp(-u))^2}]. \tag{12}$$

Then, we would like to bound the derivative. Taking absolute value of (12) and using $|Z| \leq 1$, we have

$$|\frac{\partial \widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f;\tau)}{\partial \tau}| \leq \frac{1}{\texttt{temp}} \cdot \mathbb{E}[\frac{\exp(-u)}{(1+\exp(-u))^2}].$$

Recall that $h(u) := \frac{\exp(-u)}{(1+\exp(-u))^2} = \sigma(u) \cdot (1-\sigma(u))$, which is the standard logistic hat function. It satisfies

$$0 \leq h(u) \leq 1/4, \text{ with } h(0) = 1/4.$$

Therefore, we have $\mathbb{E}[h(u)] \leq 1/4$. Hence, we conclude $|\frac{\partial \widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f;\tau)}{\partial \tau}| \leq 1/(4\texttt{temp})$.

Finally, to determine the Lipschitz parameter, for τ_1, τ_2, we have

$$|\widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f;\tau_1) - \widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f;\tau_2)| \leq \sup_\tau |\frac{\partial \widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f;\tau)}{\partial \tau}| \cdot |\tau_1 - \tau_2| \leq \frac{1}{4\texttt{temp}} |\tau_1 - \tau_2|.$$

Thus, it shows $\widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f;\tau)$ is Lipschitz in τ with constant $L = 1/(4\texttt{temp})$. □

A.3.2 Proof for Lemma 3

Lemma (Re) 2. *(Lemma 3 restated.) Suppose the assumptions in Proposition 1 hold. Then, for any f, there exists a constant $C > 0$ independent from f and $\texttt{temp}$ such that:*

$$|\mathcal{L}_{cr}^{noisy}(f) - \widetilde{\mathcal{L}}_{cr}^{noisy}(f)| \leq C\,\texttt{temp}.$$

In particular, with Big-O notation, we have

$$|\mathcal{L}_{cr}^{noisy}(f) - \widetilde{\mathcal{L}}_{cr}^{noisy}(f)| \leq O(\texttt{temp}).$$

Proof. (of Lemma 3.)

Before proving Lemma 3, we first present the following technical lemma:

Lemma 6. *(Upper bounding the gap between hard and soft indicator functions) Let*

$$\Gamma(\Delta) := \mathbb{1}[\Delta \geq 0] - \tilde{\mathbb{1}}[\Delta \geq 0], \text{ and } \tilde{\mathbb{1}}[\Delta \geq 0] = 1/(1+\exp(-\Delta/\texttt{temp})),$$

where $\texttt{temp} > 0$ is the temperature parameter. Then, for every $\Delta \in \mathbb{R}$, the following inequality holds:

$$|\Gamma(\Delta)| \leq \exp(-|\Delta|/\texttt{temp}).$$

Now we begin to prove Lemma 3.

Recall the definitions of the noisy conformal margin risk and its smoothed variant:

$$\mathcal{L}_{\mathrm{cr}}^{\mathrm{noisy}}(f) = -\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\mathrm{noisy}}}[M_f(X,\widetilde{Y}) \cdot \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]],$$
$$\widetilde{\mathcal{L}}_{\mathrm{cr}}^{\mathrm{noisy}}(f) = -\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\mathrm{noisy}}}[M_f(X,\widetilde{Y}) \cdot \widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]].$$

where the soft smoothed indicator function is $\widetilde{\mathbb{1}}[M \geq \tau] = \frac{1}{1+\exp(-(M-\tau)/\texttt{temp})}$.

For a fixed f, define the margin–threshold difference as

$$\Delta_f(X,\widetilde{Y}) := M_f(X,\widetilde{Y}) - \tau_\alpha(f),$$

and define the difference between the hard and soft smoothed indicators as

$$\Gamma(\Delta_f(X.\widetilde{Y})) := \mathbb{1}[\Delta_f(X,\widetilde{Y}) \geq 0] - \widetilde{\mathbb{1}}[\Delta_f(X,\widetilde{Y}) \geq 0].$$

By assumption, $M_f(X,\widetilde{Y}) \in (-1,1)$. Since $\tau_\alpha(f)$ is the α-quantile of $M_f(X,\widetilde{Y})$ under $\mathcal{P}_{\text{noisy}}$, we also have $\tau_\alpha(f) \in (-1,1)$. Hence we have

$$\Delta_f(X,\widetilde{Y}) \in (-2,2) \text{ almost surely.}$$

In what follows, we use three steps to prove the desired result.

Step 1: Reducing to bounding $\mathbb{E}[|\Gamma(\Delta_f)|]$. We first write the difference between the two risks explicitly as follows:

$$\begin{aligned}\mathcal{L}_{\text{cr}}^{\text{noisy}}(f) - \widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}(f) =& -\mathbb{E}_{\text{noisy}}[M_f(X,\widetilde{Y}) \cdot \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]] + \mathbb{E}_{\text{noisy}}[M_f(X,\widetilde{Y}) \cdot \widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]] \\ =& \mathbb{E}_{\text{noisy}}[M_f(X,\widetilde{Y}) \cdot (\widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)] - \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)])] \\ =& -E_{\text{noisy}}[M_f(X,\widetilde{Y}) \cdot \Gamma(\Delta_f(X,\widetilde{Y}))].\end{aligned}$$

Taking the absolute values and using $|M_f(X,\widetilde{Y})| \leq B$, we have

$$\begin{aligned}|\mathcal{L}_{\text{cr}}^{\text{noisy}}(f) - \widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}(f)| =& |\mathbb{E}_{\text{noisy}}[M_f(X,\widetilde{Y}) \cdot \Gamma(\Delta_f(X,\widetilde{Y}))]| \leq \mathbb{E}_{\text{noisy}}[|M_f(X,\widetilde{Y})| \cdot |\Gamma(\Delta_f(X,\widetilde{Y}))|] \\ \leq& \mathbb{E}_{\text{noisy}}[B|\Gamma(\Delta_f(X,\widetilde{Y}))|].\end{aligned}$$

Therefore, it suffices to upper bound $\mathbb{E}_{\text{noisy}}[|\Gamma(\Delta_f)|]$.

Step 2: Use the density of Δ_f and Lemma 6. Let $G(t)$ and $g(t) = G'(t)$ denote the CDF and the density of $M_f(X,\widetilde{Y})$ under $\mathcal{P}_{\text{noisy}}$, respectively. By the assumptions in Proposition 1, g is continuous and bounded, i.e.,

$$g_\infty := \sup_{t\in\mathbb{R}} g(t) < \infty$$

For fixed f, $\tau_\alpha(f)$ is a constant, so the random variable $\Delta_f = M_f - \tau_\alpha(f)$ has density which is also continuous and bounded with $\|g_{\Delta_f}\|_\infty \leq g_\infty$

Therefore, we can write

$$\mathbb{E}_{\text{noisy}}[|\Gamma(\Delta_f)|] = \int_{-2}^{2} |\Gamma(\delta)| \cdot g_{\Delta_f}(\delta) d\delta,$$

where we use the fact that $\Delta_f \in (-2,2)$ almost surely.

By Lemma 6, for all $\delta \in \mathbb{R}$, we have

$$|\Gamma(\delta)| \leq \exp(-|\delta|/\texttt{temp}).$$

Combining this with the bound on g_{Δ_f}, we derive

$$\mathbb{E}_{\text{noisy}}[|\Gamma(\Delta_f)|] = \int_{-2}^{2} |\Gamma(\delta)| g_{\delta_f}(\delta) d\delta \leq g_\infty \cdot \int_{-2}^{2} \exp(-|\delta|/\texttt{temp}) d\delta.$$

The integral can be computed explicitly:

$$\int_{-2}^{2} \exp(-\frac{|\delta|}{\texttt{temp}}) d\delta = 2\int_{0}^{2} \exp(-\delta/\gamma) d\delta = 2\texttt{temp}(1 - \exp(-2/\texttt{temp})) \leq 2\texttt{temp},$$

where we used $1 - \exp(2/\texttt{temp}) \leq 1$.

Thus, we have $\mathbb{E}_{\text{noisy}}[|\Gamma(\Delta_f)|] \leq 2g_\infty\texttt{temp}$.

Step 3: Conclude the upper bound. Putting everything together, we derive

$$|| \leq \mathbb{E}_{\text{noisy}}[|\Gamma(\Delta_f)|] \leq 2g_\infty \texttt{temp}.$$

Hence the difference between the hard formal margin risk and its smoothed counter is $O(\texttt{temp})$.

Recall that the $M_f(X, \widetilde{Y})$ is the confidence gap between the confidence score of the observed label and the highest confidence among other candidate labels. Thus, $M_f(X, \widetilde{Y}) \in (-1, 1)$. As $\tau_\alpha(f)$ is the α-quantile of $M_f(X, \widetilde{Y})$, $\tau_\alpha(f) \in (-1, 1)$ and $\Delta_f(X, Y) \in (-2, 2)$.

We also recall that $G(t)$ is the CDF of $M_f(X, \widetilde{Y})$ under $\mathcal{P}_{\text{noisy}}$, and $g(t) = G'(t)$ is its density, which is assumed to be continuous and bounded. Denote $g_\infty = \sup_t g(t) < +\infty$. Since $\tau_\alpha(f)$ is a constant for fixed f, the density of Δ_f is $g_{\Delta_f}(t) = g(t + \tau_\alpha(f))$, which satisfies $\|g_{\Delta_f}\|_\infty = g_\infty$.

Then, we have:

$$\begin{aligned}
&|\mathcal{L}_{\text{cr}}^{\text{noisy}}(f) - \widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}(f)| \\
=&\Big|\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\big[M_f(X,\widetilde{Y}) \cdot \widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]\big] - \mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\big[M_f(X,\widetilde{Y}) \cdot \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]\big]\Big| \\
=&\Big|\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\Big[M_f(X,\widetilde{Y}) \cdot \big(\widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)] - \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]\big)\Big]\Big| \\
\leq&|\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\big[M_f(X,\widetilde{Y})\big]| \cdot \Big|\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\Big[\big(\widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)] - \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]\big)\Big]\Big| \\
\leq&\Big|\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\Big[\big(\widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)] - \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]\big)\Big]\Big| \\
\leq&\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\Big|\widetilde{\mathbb{1}}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)] - \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]\Big| \\
=&\mathbb{E}_{(X,\widetilde{Y})\sim\mathcal{P}_{\text{noisy}}}\Big|\widetilde{\mathbb{1}}[\Delta_f(X,Y) \geq 0] - \mathbb{1}[\Delta_f(X,Y) \geq 0]\Big| \\
=&\mathbb{E}[\Gamma(\Delta_f(X,Y))] \\
=&\int_{-2}^{2} \Gamma(\Delta_f(X,Y))\partial_{\Delta_f(X,Y)}\Gamma(\Delta_f(X,Y))d\Delta_f(X,Y) \\
\leq&\int_{-2}^{2} \exp\Big(-\frac{|\Delta_f(X,Y)|}{\texttt{temp}}\Big)\partial_{\Delta_f(X,Y)}\Gamma(\Delta_f(X,Y))d\Delta_f(X,Y) \\
\leq&g_\infty \cdot \int_{-2}^{2} \exp\Big(-\frac{|\Delta_f(X,Y)|}{\texttt{temp}}\Big)d\Delta_f(X,Y) \\
=&g_\infty \cdot 2\texttt{temp}(1 - \exp(-2/\texttt{temp})) \\
\leq&O(\texttt{temp}),
\end{aligned}$$

where the first inequality is due to the Hölder's inequality, the second inequality is due to the Jansen's inequality and expectation is convex function, the third inequality is due to $M_f(X, \widetilde{Y}) \in (-1, 1)$, the fourth inequality is due to Lemma 6, the fifth inequality is due to the upper bound for the gradient of $\Delta_f(X, Y)$ is g_∞, and the last inequality is due to $1 - \exp(-2/\texttt{temp}) \leq 1$.

Therefore, we have:

$$|\mathcal{L}_{\text{cr}}^{\text{noisy}}(f) - \widetilde{\mathcal{L}}_{\text{cr}}^{\text{noisy}}(f)| \leq O(\texttt{temp}).$$

□

Now we begin to prove Lemma 6.

Proof. (of Lemma 6)

We analyze the two cases, i.e., $\Delta \geq 0$ and $\Delta < 0$ separately below.

Case 1: $\Delta \geq 0$

In this case, by using the condition $\Delta \geq 0$, we first reformulate $\Gamma(\Delta)$ as follows

$$\Gamma(\Delta) = 1 - \frac{1}{1+\exp(-\Delta/\texttt{temp})} = \frac{\exp(-\Delta/\texttt{temp})}{1+\exp(-\Delta/\texttt{temp})}.$$

Since the denominator is at least 1 (due to $\exp(\cdot) > 0$), we have the following inequalities

$$0 < \Gamma(\Delta) = |\Gamma(\Delta)| \leq \exp(-\Delta/\texttt{temp}) = \exp(-|\Delta|/\texttt{temp}), \tag{13}$$

where the last equality is due to $\Delta \geq 0$, the condition for this **case 1**. Thus, the desired inequality holds for all $\Delta > 0$.

Case 2: $\Delta < 0$

In this case, by using the condition $\Delta < 0$, we reformulate $\Gamma(\Delta)$ as follows

$$\Gamma(\Delta) = 0 - \frac{1}{1+\exp(-\Delta/\texttt{temp})} = -\frac{1}{1+\exp(-\Delta/\texttt{temp})} < 0.$$

Then, with the condition $\Delta < 0$, we have $|\Delta| = -\Delta$. To upper bound $|\Gamma(\Delta)|$ in this case, it suffixes to show:

$$|\Gamma(\Delta)| = -\Gamma(\Delta) = \frac{1}{1+\exp(-\Delta/\texttt{temp})} \leq \exp(\Delta/\texttt{temp}) = \exp(-|\Delta|/\texttt{temp}), \tag{14}$$

where the last inequality is due to $\exp(a) \leq 1 + \exp(a)$ for any $a \in \mathbb{R}$.

Finally, combining (13) and (14) implies that for any $\Delta \in \mathbb{R}$, the following inequality holds:

$$|\Gamma(\Delta)| \leq \exp(-|\Delta|/\texttt{temp}).$$

□

A.3.3 Proof for Lemma 4

Lemma (Re) 3. *(Lemma 4 restated.) Define δ_w as the average Wasserstein-1 distance between posteriors under the training and test domains, where $\delta_w = \mathbb{E}_{X\sim\mathcal{P}(X)}\Big[W_1\big(\mathbb{P}_{\mathcal{P}_{noisy}}(\cdot \mid X),\ \mathbb{P}_{\mathcal{P}}(\cdot \mid X)\big)\Big]$. Then the following inequality holds:*

$$\big|\mathcal{L}_{cr}(f) - \mathcal{L}_{cr}^{noisy}(f)\big| \ \leq\ 2\delta_w + \alpha.$$

Proof. of Lemma 4.

Before proving Lemma 4, we first show the following technical lemmas:

Lemma 7 (Range of Probability)**.** *Define δ_w as the average Wasserstein-1 distance between posteriors under the training and test domains, where $\delta_w = \mathbb{E}_{X\sim\mathcal{P}(X)}\Big[W_1\big(\mathbb{P}_{\mathcal{P}_{noisy}}(\cdot \mid X),\ \mathbb{P}_{\mathcal{P}}(\cdot \mid X)\big)\Big]$. Then the probability $\mathbb{P}_{\mathcal{P}}\big(M_f(X,y) \geq \tau_\alpha(f)\big) \in [1-\alpha-\delta_w, 1-\alpha+\delta_w]$.*

Lemma 8 (Wasserstein-1 distance bound of expected difference (Dudley, 2018; Arjovsky et al., 2017))**.** *For any bounded measurable function $g : X \to \mathbb{R}$, we have:*

$$\Big|\mathbb{E}_{\mathcal{P}_{noisy}}[g] - \mathbb{E}_{\mathcal{P}}[g]\Big| \leq 2\|g\|_\infty W_1\big(\mathbb{P}_{\mathcal{P}_{noisy}}, \mathbb{P}_{\mathcal{P}}\big).$$

The proof of Lemma 7 is deferred to the end of this proof. Now we begin to prove Lemma 4.

Recall that $M_f(X, \widetilde{Y})$ as the gap of confidence scores, where $|M_f(X,y)| \leq 1$ for any $(X, \widetilde{Y})$. We also recall that $\mathcal{L}_{\text{cr}}(f) = -\mathbb{E}_{X\sim\mathcal{P}}\left[M_f(X,Y) \mid M_f(X,Y) \geq \tau_\alpha(f)\right]$.

Define $P = \mathbb{P}_{\mathcal{P}}\big(M_f(X,Y) \geq \tau_\alpha(f)\big)$. Thus, we have:

$$
\begin{aligned}
&\big|\mathcal{L}_{\mathrm{cr}}(f) - \mathcal{L}_{\mathrm{cr}}^{\mathrm{noisy}}(f)\big| \\
=&\big|\mathbb{E}_{\mathcal{P}}\left[M_f(X,Y) \mid M_f(X,Y) \geq \tau_\alpha(f)\right] - \mathbb{E}_{\mathcal{P}_{\mathrm{noisy}}}\Big[M_f(X,\widetilde{Y}) \cdot \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]\Big]\,\big| \\
=&\left|\frac{\mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]\Big]}{\mathbb{P}_{\mathcal{P}}\big(M_f(X,Y) \geq \tau_\alpha(f)\big)} - \mathbb{E}_{\mathcal{P}_{\mathrm{noisy}}}\Big[M_f(X,\widetilde{Y}) \cdot \mathbb{1}[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)]\Big]\right| \\
=&\left|\frac{1}{P}\mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]\Big] - \mathbb{E}_{\mathcal{P}_{\mathrm{noisy}}}\Big[M_f(X,\widetilde{Y})\mathbb{1}\big[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)\big]\Big]\right| \\
=&\left|\Big(\frac{1}{P} - 1\Big)\mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]\Big] + \mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]\Big]\right. \\
&\left. - \mathbb{E}_{\mathcal{P}_{\mathrm{noisy}}}\Big[M_f(X,\widetilde{Y})\mathbb{1}\big[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)\big]\Big]\right| \\
\leq&\underbrace{\left|\Big(\frac{1}{P} - 1\Big)\mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]\Big]\right|}_{A} + \\
&\underbrace{\left|\mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]\Big] - \mathbb{E}_{\mathcal{P}_{\mathrm{noisy}}}\Big[M_f(X,\widetilde{Y})\mathbb{1}\big[M_f(X,\widetilde{Y}) \geq \tau_\alpha(f)\big]\Big]\right|}_{B},
\end{aligned}
\tag{15}
$$

where the first inequality is due to the triangle inequality.

Then we analyze the upper bound of A and B, respectively.

Part I: Upper bound of A

$$
\begin{aligned}
&\left|\Big(\frac{1}{P} - 1\Big)\mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]\Big]\right| \\
\leq&\left|\Big(\frac{1}{P} - 1\Big)\right|\left|\mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]\Big]\right| \\
=&\left|\Big(\frac{1}{P} - 1\Big)\right|\left|\mathbb{E}_{\mathcal{P}}\big[M_f(X,Y)\big]\right|\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big] \\
\leq&\left|\Big(\frac{1}{P} - 1\Big)\right|\mathbb{E}_{\mathcal{P}}\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big] \\
=&\left|\Big(\frac{1}{P} - 1\Big)\right|P \\
=&\frac{\Big|1 - P\Big|}{P}P \\
\leq&\alpha + \delta_w,
\end{aligned}
\tag{16}
$$

where the first inequality is due to the triangle inequality, the second inequality is due to $|M_f(X,Y)| \leq 1$, and the last inequality is due to the Lemma 7.

Part II: Upper bound of B

For any fixed X, we define $\phi_X(Y) = M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\big]$. Then, we try to bound $\|\phi_X(y)\|_\infty$ as:

$$
\|\phi_X(Y)\|_\infty = \|M_f(X,Y)\mathbb{1}\big[M_f(X,Y) \geq \tau_\alpha(f)\|_\infty \leq \|M_f(X,Y)\|_\infty \leq 1,
$$

where the first inequality is due to $\mathbb{1}[\cdot] \in [0,1]$, and the second inequality is due to $|M_f(X,Y)| \leq 1$.

Then, we rewrite B as:

$$\begin{aligned}
&\left|\mathbb{E}_{\mathcal{P}}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y)\geq\tau_\alpha(f)\big]\Big]-\mathbb{E}_{\mathcal{P}_{\text{noisy}}}\Big[M_f(X,\widetilde{Y})\mathbb{1}\big[M_f(X,\widetilde{Y})\geq\tau_\alpha(f)\big]\Big]\right|\\
=&\left|\mathbb{E}_X\left[\mathbb{E}_{y\sim\mathbb{P}(\cdot|X)}\Big[M_f(X,Y)\mathbb{1}\big[M_f(X,Y)\geq\tau_\alpha(f)\big]\Big]-E_{y\sim\mathbb{P}_{\text{noisy}}(\cdot|X)}\Big[M_f(X,\widetilde{Y})\mathbb{1}\big[M_f(X,\widetilde{Y})\geq\tau_\alpha(f)\big]\Big]\right]\right|\\
=&\left|\mathbb{E}_X\left[\mathbb{E}_{y\sim\mathbb{P}(\cdot|X)}\Big[\phi_X(Y)\Big]-E_{y\sim\mathbb{P}_{\text{noisy}}(\cdot|X)}\Big[\phi_X(\widetilde{Y})\Big]\right]\right|\\
\leq&\left|\mathbb{E}_X\left[2W_1\big(\mathbb{P}_{\mathcal{P}_{\text{noisy}}}(\cdot\mid X),\ \mathbb{P}_{\mathcal{P}}(\cdot\mid X)\big)\right]\right|\\
=&2\left|\mathbb{E}_X\left[W_1\big(\mathbb{P}_{\mathcal{P}_{\text{noisy}}}(\cdot\mid X),\ \mathbb{P}_{\mathcal{P}}(\cdot\mid X)\big)\right]\right|\\
=&2\delta_w,
\end{aligned}\tag{17}$$

where the first inequality is due to Lemma 8.

Combining inequalities (15),(16), and (17), we have that:

$$\big|\mathcal{L}_{\text{cr}}(f)-\mathcal{L}_{\text{cr}}^{\text{noisy}}(f)\big|\leq 2\delta_w+\alpha.$$

□

Proof. (of Lemma 7)

Recall that $\mathcal{Y}$ denotes the label space equipped with the 0–1 distance $d(y_1,y_2)=\mathbb{I}\{y_1\neq y_2\}$. For each fixed X, define $\Phi_X(Y)=\mathbb{1}\big(M_f(X,Y)\geq\tau_\alpha(f)\big)$. Under this metric, $\Phi_X(Y)$ is 1-Lipschitz because $|\Phi_X(y_1)-\Phi_X(y_2)|\leq d(y_1,y_2)$.

Denote by $\mu_X=\mathbb{P}_{\mathcal{P}_{\text{noisy}}}(\cdot\mid X)$ and $\nu_X=\mathbb{P}_{\mathcal{P}}(\cdot\mid X)$ the label posteriors under the noisy and clean domains, respectively. Then, we have:

$$\begin{aligned}
&\left|\mathbb{P}_{\mathcal{P}_{\text{noisy}}}\big(M_f(X,\widetilde{Y})\geq\tau_\alpha(f)\big)-\mathbb{P}_{\mathcal{P}}\big(M_f(X,Y)\geq\tau_\alpha(f)\big)\right|\\
=&\left|\mathbb{E}_{X\sim\mathcal{P}(X)}\Big[\mathbb{E}_{\widetilde{Y}\sim\mu_X}\Phi_X(\widetilde{Y})-\mathbb{E}_{Y\sim\nu_X}\Phi_X(Y)\Big]\right|\\
\leq&\mathbb{E}_{X\sim\mathcal{P}(X)}\Big[\Big|\mathbb{E}_{\widetilde{Y}\sim\mu_X}\Phi_X(\widetilde{Y})-\mathbb{E}_{Y\sim\nu_X}\Phi_X(Y)\Big|\Big]\\
=&\mathbb{E}_{X\sim\mathcal{P}(X)}\Big[\Big|\sum_{y\in\mathcal{Y}}\Phi_X(y)\big(\mu_X(y)-\nu_X(y)\big)\Big|\Big]\\
\leq&\mathbb{E}_{X\sim\mathcal{P}_X}\Big[W_1\big(\mu_X,\ \nu_X\big)\Big]\\
=&\mathbb{E}_{X\sim\mathcal{P}_X}\Big[W_1\big(\mathbb{P}_{\mathcal{P}_{\text{noisy}}}(\cdot\mid X),\ \mathbb{P}_{\mathcal{P}}(\cdot\mid X)\big)\Big]=\delta_w,
\end{aligned}$$

where the first inequality is due to Jensen's inequality, and the last inequality is due to Kantorovich–Rubinstein duality for Wasserstein-1.

Due to $\mathbb{P}_{\mathcal{P}_{\text{noisy}}}\big(M_f(X,\widetilde{Y})\geq\tau_\alpha(f)\big)=1-\alpha$, we have: $\Big|\mathbb{P}_{\mathcal{P}}\big(M_f(X,Y)\geq\tau_\alpha(f)\big)-(1-\alpha)\Big|=\delta_w$. □

B ADDITIONAL EXPERIMENTS FOR MULTI-CLASS CLASSIFICATION

B.1 Additional Experimental Setup Details

Datasets. To evaluate model robustness under noisy supervision, we construct asymmetric label noise on both CIFAR-100 and mini-ImageNet, simulating realistic mislabeling patterns.

CIFAR-100 consists of 100 fine-grained classes organized into 20 coarse-grained superclasses (e.g., aquatic mammals, large carnivores). To inject structured label noise, we first build a mapping from each fine label to its corresponding coarse superclass. For noise rates $\rho \in \{0\%, 5\%, 10\%, 20\%, 30\%, 40\%\}$, we randomly select ρn training samples, where n is the dataset size. Each selected label is replaced by a randomly chosen different label from the same superclass, ensuring that the corrupted label remains semantically similar to the original. The indices of noisy samples are recorded to facilitate evaluation and ablation.

mini-ImageNet is a 100-class image classification dataset. To inject asymmetric noise, we randomly select a fraction $\rho = 20\%$ of training samples and replace each label Y with $(Y + 1) \mod 100$, introducing a deterministic and minimal perturbation.

FOOD101 is a 101-class image classification dataset. To inject asymmetric noise, we randomly select a fraction $\rho = 20\%$ of training samples and replace each label Y with $(Y + 1) \mod 100$, introducing a deterministic and minimal perturbation.

This form of circular asymmetric noise preserves the class index structure and ensures the new label is different from the original. The transformation is applied only to training labels, leaving the validation and test sets clean for evaluation.

Hyperparameters for training. We set datasets, base loss, batch size, training epochs, training parameters (learning rate, learning schedule, momentum, gamma, and weight decay), λ and α as hyperparameter choices. We search for hyperparameters on batch size $\in \{64, 128, 256, 512\}$, epochs $\in \{50, 100\}$, learning rate $(\eta) \in \{0.1, 0.07, 0.05, 0.03, 0.01\}$, learning rate schedule $\in \{[10], [10, 30], [60], [60, 80]\}$, Momentum $= 0.9$, weight decay $= 0.0002$, $\gamma = 0.01$, $\lambda = \{0.05, 0.1, 0.15, 0.2, 0.25\}$, and $\alpha = \{0.05, 0.1, 0.15, 0.2, 0.25\}$ to select the best combination of hyperparameters of each methods. For GCE, we additionally scale λ by a factor of 0.1, resulting in $\lambda \in \{0.005, 0.01, 0.015, 0.02, 0.025\}$. The hyerparameters employed to get the results presented in the main paper are summarized in Table 5.

Data	Loss	Batch size	Epochs	η	lr schedule	Momentum	γ	weight decay	λ	α
CIFAR-100	CE+CMRM	128	50	0.05	[10]	0.9	0.01	0.0002	0.1	0.15
	Focal+CMRM	128	50	0.05	[10]	0.9	0.01	0.0002	0.15	0.1
	LDAM+CMRM	128	100	0.05	[60, 80]	0.9	0.01	0.0002	0.1	0.15
	GCE+CMRM	128	50	0.05	[10]	0.9	0.01	0.0002	0.005	0.05
mini-ImageNet	CE+CMRM	512	50	0.05	[10]	0.9	0.01	0.0002	0.15	0.2
	Focal+CMRM	512	50	0.05	[10]	0.9	0.01	0.0002	0.15	0.15
	LDAM+CMRM	512	100	0.05	[60, 80]	0.9	0.01	0.0002	0.2	0.1
	GCE+CMRM	512	50	0.05	[10]	0.9	0.01	0.0002	0.0005	0.15
FOOD101	CE+CMRM	512	50	0.05	[10]	0.9	0.01	0.0002	0.15	0.15
	Focal+CMRM	512	50	0.05	[10]	0.9	0.01	0.0002	0.15	0.1
	LDAM+CMRM	512	100	0.05	[60, 80]	0.9	0.01	0.0002	0.05	0.15
	GCE+CMRM	512	50	0.05	[10]	0.9	0.01	0.0002	0.0001	0.1

Table 5: **The details we used to train our models for multi-class classification corrupted by synthetic noise** with noise rate 20%. We reported the hyperparameters that give the best accuracy. We employed SGD optimizer for all training unless specified.

B.2 Additional Experimental Results

Result: CMRM improves accuracy and reduces uncertainty under different types of noise.

Table 6 summarizes results on CIFAR-100 with synthetic label noise at rates $\{0\%, 5\%, 10\%, 20\%, 30\%, 40\%\}$, where 0% corresponds to the clean-label setting. For each objective (CE, Focal, LDAM, and GCE), we compare the Base model with its +CMRM variant. On average across all objectives and noise levels, CMRM improves accuracy by 1.34 and reduces M.APSS by 3.89%. The most notable gains occur under moderate to high noise. For example, CE and Focal achieve up to +2.14 and 2.22 accuracy improvements, while GCE shows the largest uncertainty reduction (up to -18.44% in M.APSS). Even when combined with LDAM, which already promotes margin separation, CMRM consistently provides additional accuracy improvements without increasing predictive uncertainty.

Table 7 reports the mean ± standard deviation of Top-1 accuracy across multiple random seeds on CIFAR-100 with 20% synthetic label noise. For each objective (CE, Focal, LDAM, and GCE), we compare the Base model

Noise Rates	Metric	CE		Focal		LDAM		GCE	
		Base	+CMRM	Base	+CMRM	Base	+CMRM	Base	+CMRM
0%	ACC (%) ↑	68.82	**69.38** (+0.56)	68.32	**69.27** (+0.95)	62.94	**64.15** (+1.19)	**62.80**	**62.80** (+0)
	M.APSS ↓	**3.22**	3.37 (+4.52%)	**3.04**	3.26 (+7.24%)	**10.89**	11.06 (+1.56%)	**5.95**	**5.95** (+0%)
5%	ACC (%) ↑	67.48	**68.02** (+0.54)	66.76	**67.66** (+0.90)	61.63	**63.00** (+1.37)	63.88	**65.29** (+1.41)
	M.APSS ↓	**3.79**	4.13 (+8.97%)	**3.66**	3.82 (+4.37%)	**13.03**	13.28 (+1.92%)	**6.27**	6.30 (+0.48%)
10%	ACC (%) ↑	66.29	**67.32** (+1.03)	66.52	**66.82** (+0.30)	61.06	**62.28** (+1.22)	62.80	**64.73** (+1.93)
	M.APSS ↓	**4.31**	4.57 (+6.03%)	**4.28**	4.53 (+5.84%)	15.57	**15.32** (−1.61%)	6.55	**5.58** (−14.81%)
20%	ACC (%) ↑	65.16	**66.32** (+1.16)	64.42	**65.39** (+0.97)	59.63	**61.12** (+1.49)	62.17	**63.65** (+1.48)
	M.APSS ↓	6.67	**6.52** (−2.25%)	6.89	**6.61** (−4.06%)	17.85	**17.67** (−0.78%)	7.70	**6.28** (−18.44%)
30%	ACC (%) ↑	64.02	**65.22** (+1.2)	63.46	**65.33** (+1.87)	58.27	**59.56** (+1.29)	61.49	**63.23** (+1.74)
	M.APSS ↓	8.81	**7.88** (−10.56%)	9.12	**7.66** (−16.01%)	20.44	**18.63** (−8.86%)	6.77	**6.18** (−8.71%)
40%	ACC (%) ↑	62.63	**64.77** (+2.14)	62.65	**64.87** (+2.22)	58.25	**59.27** (+1.02)	59.34	**60.85** (+1.51)
	M.APSS ↓	10.31	**8.99** (−12.80%)	10.55	**9.12** (−13.55%)	**19.54**	20.95 (+7.22%)	7.76	**7.74** (−0.26%)

Table 6: **Top-1 accuracy (%) and marginal average prediction set size (M.APSS ↓) on CIFAR-100 datasets corrupted by synthetic noise with noise rates** $\{0\%, 5\%, 10\%, 20\%, 30\%, 40\%\}$**.** Each Base objective is paired with its +CMRM counterpart; the better value within each pair is in **bold**. Numbers in parentheses indicate the relative change (%): + denotes accuracy improvement, and − denotes M.APSS reduction compared to the corresponding Base objective. On average across all datasets and objectives, CMRM improves accuracy by 1.34 and reduces M.APSS by 3.89%.

Metric	CE		Focal		LDAM		GCE	
	Base	+CMRM	Base	+CMRM	Base	+CMRM	Base	+CMRM
ACC (%) ↑	64.74 ± 0.50	**65.95 ± 0.30 (+1.21)**	64.33 ± 0.15	**65.26 ± 0.21 (+0.93)**	59.64 ± 0.34	**60.76 ± 0.26 (+1.12)**	62.46 ± 0.70	**63.60 ± 0.24 (+1.14)**

Table 7: **Mean ± standard deviation (std) of Top-1 accuracy (%) across multiple random seeds on CIFAR-100 with synthetic label noise (noise rate** 20%**).** Each base objective is paired with its +CMRM variant. The better value within each pair is shown in **bold**. Numbers in parentheses denote the absolute accuracy improvement over the corresponding base objective.

Method	CIFAR-10N (Aggre)		CIFAR-10N (Rand1)		CIFAR-10N (Rand2)		CIFAR-10N (Rand3)		CIFAR-10N (Worst)		CIFAR-100N	
	ACC(%)	M.APSS	ACC(%)	M.APSS	ACC(%)	M.APSS	ACC(%)	M.APSS	ACC(%)	M.APSS	ACC(%)	M.APSS
NI-ERM	98.69	0.904	98.80	0.902	98.65	**0.903**	98.67	0.904	95.71	0.93	83.17	1.49
NI-ERM+CMRM	**98.81**	**0.903**	**99.03**	**0.901**	**98.95**	**0.903**	**98.88**	**0.899**	**97.19**	**0.91**	**83.95**	**1.29**
	(+0.12)	(−0.11%)	(+0.23)	(−0.11%)	(+0.30)	(0%)	(+0.21)	(−0.55%)	(+1.48)	(−2.15%)	(+0.78)	(−13.42%)

Table 8: **Top-1 accuracy (%) and marginal average prediction set size (M.APSS ↓) on CIFAR-10N and CIFAR-100N corrupted by human annotation noise.** Numbers in parentheses indicate the relative change: + denotes accuracy improvement and −% denotes M.APSS reduction. CMRM consistently improves accuracy and reduces uncertainty across CIFAR-N variants, with the largest gains observed on CIFAR-10N and CIFAR-100N. On average across all datasets, CMRM improves accuracy by 0.52 and reduces M.APSS by 2.72%.

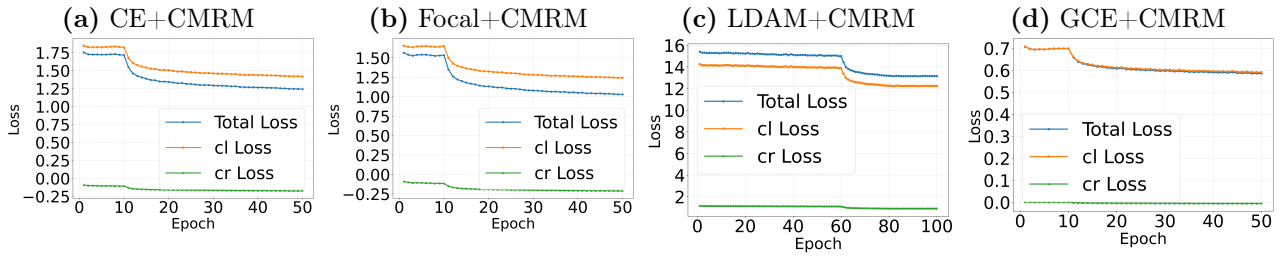

Figure 5: **training dynamics of total loss (Total), classification loss (cl), and CMRM loss (cr) for all base losses** over epochs on CIFAR-100 with 20% synthetical label noise. Subfigure **(a)** CE+CERM; Subfigure **(b)** Focal+CERM; Subfigure **(c)** LDAM+CERM; Subfigure **(d)** GCE+CERM; CMRM exhibits stable and monotonic convergence alongside standard loss components.

with its +CMRM variant. Across all objectives, CMRM consistently improves accuracy under the multi-seed evaluation. In particular, CMRM yields accuracy gains of +1.21, +0.93, +1.12, and +1.14 for CE, Focal, LDAM, and GCE, respectively. These results demonstrate that the performance improvements brought by CMRM remain stable across different random seeds.

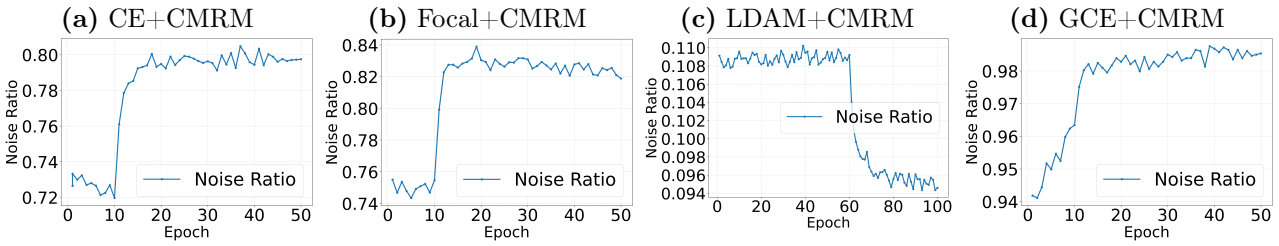

Figure 6: **The ratio of noisy samples among those filtered out by CMRM at each epoch** on CIFAR-100 with 20% synthetical label noise. Subfigure **(a)** CE+CMRM ($\alpha = 0.15$); Subfigure **(b)** Focal+CMRM ($\alpha = 0.1$); Subfigure **(c)** LDAM+CMRM ($\alpha = 0.15$); Subfigure **(d)** GCE+CMRM ($\alpha = 0.05$); CMRM consistently suppresses noisy examples by excluding low-margin samples during training.

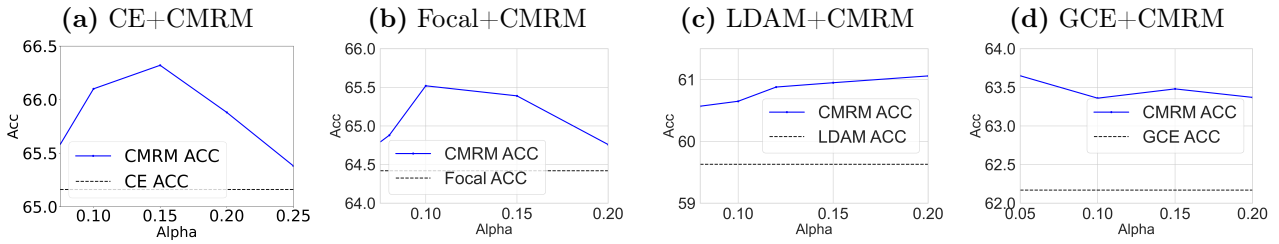

Figure 7: **The sensitivity of α of different base losses** on CIFAR-100 with 20% synthetical label noise. Subfigure **(a)** CE+CERM; Subfigure **(b)** Focal+CERM; Subfigure **(c)** LDAM+CERM; Subfigure **(d)** GCE+CERM; CMRM maintains higher accuracy than CE across a range of α values, indicating robustness to hyperparameter α.

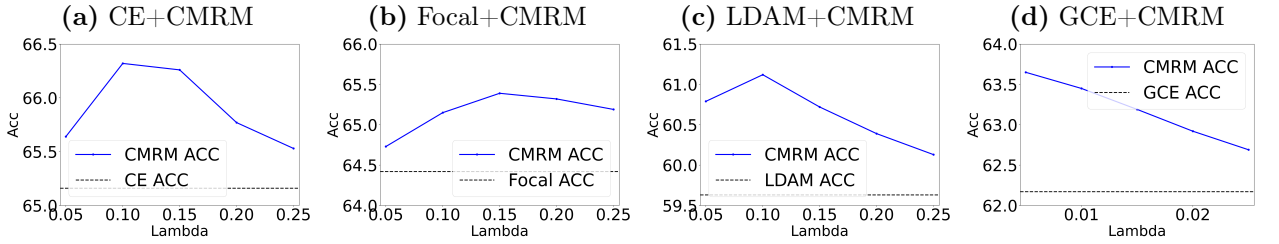

Figure 8: **The sensitivity of λ of different base losses** on CIFAR-100 with 20% synthetical label noise. Subfigure **(a)** CE+CERM; Subfigure **(b)** Focal+CERM; Subfigure **(c)** LDAM+CERM; Subfigure **(d)** GCE+CERM; CMRM maintains higher accuracy than CE across a range of λ values, indicating robustness to hyperparameter λ.

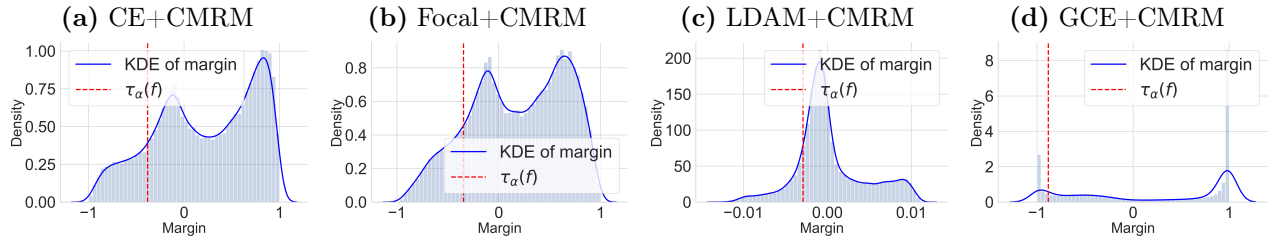

Figure 9: **the kernel density estimate (KDE) of the margin distribution of different base losses** on CIFAR-100 with 20% synthetical label noise. Subfigure **(a)** CE+CMRM ($\alpha = 0.15$); Subfigure **(b)** Focal+CMRM ($\alpha = 0.1$); Subfigure **(c)** LDAM+CMRM ($\alpha = 0.15$); Subfigure **(d)** GCE+CMRM ($\alpha = 0.05$); The vertical dashed line indicating the estimated $\tau_\alpha(f)$. The density curve is smooth and strictly positive around $\tau_\alpha(f)$, supporting the differentiability and positive-density assumption in Proposition 1.

Table 8 summarizes results on CIFAR-10N and CIFAR-100N with human-annotated label noise. Across all datasets, CMRM consistently improves both accuracy and uncertainty compared to NI-ERM. On average, CMRM increases accuracy by 0.52 and reduces M.APSS by 2.72%, indicating more confident and reliable predictions under noisy supervision. The improvements are particularly pronounced on the most challenging settings, i.e., CIFAR-10N (Worst) and CIFAR-100N, where accuracy rises by up to 1.48 and predictive uncertainty decreases

by as much as 13.42%. These consistent gains demonstrate that CMRM effectively enhances both robustness and calibration under real-world label noise.

Result: CMRM loss convergences. Figure 5 shows the training dynamics of the classification loss and the CMRM regularization loss. Both components decrease steadily and stabilize as training progresses, indicating smooth joint optimization. The CMRM term integrates with standard objectives and does not introduce instability or slowing down of convergence, demonstrating that CMRM can be efficiently optimized.

Result: CMRM filters out noisy samples during training. Figure 6 shows the fraction of noisy samples among those excluded by CMRM at each epoch. This proportion rapidly increases during the early training phase and stabilizes above 78%, indicating that CMRM consistently identifies and filters out mislabeled examples via its margin-based thresholding mechanism.

Result: CMRM is robust to the choices of hyperparameter α and λ. Figure 7 and 8 examine the sensitivity of CMRM to the hyperparameter α and λ, respectively. Across a range of α and λ values, CMRM consistently achieves higher accuracy than CE, indicating that its performance is robust to the choice of α and λ, and does not rely on careful hyperparameter tuning.

Result: Assumptions in Proposition 1 are empirically valid. Figure 9 presents the kernel density estimate (KDE) of the margin distribution, with the vertical dashed line indicating the estimated $\tau_\alpha(f)$. The density curve is smooth and strictly positive in the neighborhood of $\tau_\alpha(f)$, supporting the differentiability and positive-density assumption in Proposition 1.

C ADDITIONAL EXPERIMENTS FOR BINARY CLASSIFICATION

C.1 Additional Experimental Setup Details

Datasets. We evaluate our methods on four standard binary classification benchmarks, three of which are sourced from the UCI Machine Learning Repository:

- **EMAIL**: This dataset contains features extracted from emails and aims to classify whether an email is spam or not. We treat spam emails as the positive class (label 1) and non-spam emails as the negative class (label 0). Following prior work, we invert the original labels in the dataset to conform to this definition.
- **ADULT**: This dataset involves predicting whether an individual's income exceeds $50K$ based on demographic and employment features. We define the positive class as individuals earning $\leq 50K$ (label 1) and the negative class as those earning $> 50K$ (label 0), effectively focusing on identifying lower-income individuals.
- **CREDIT**: This dataset contains information on credit card clients and whether they defaulted on their payment in the following month. We define the positive class as non-defaulting clients (label 1) and the negative class as clients who defaulted (label 0), inverting the original label to emphasize reliable borrowers.

Hyperparameters for training. We set datasets, base losses, base models, learning rate, λ^+, λ^-, α^+, and α^- as hyperparameter choices. We search for hyperparameters on learning rate (η) $\in \{0.001, 0.005, 0.01\}$, $\lambda^+ \in \{0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5\}$, $\lambda^- \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5\}$, $\alpha^+ \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and $\alpha^- \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ to select the best combination of hyperparameters of each methods. The hyperparameters employed to get the results presented in the main paper are summarized in Table 10.

C.2 Additional Experimental Results

Result: CMRM improves robustness for binary classification. Table 9 reports results on the Adult, Email, and Credit datasets under a 20% label noise setting. Across all base models (LR, Focal, SVM, GCE), CMRM consistently improves ranking (AUROC, AUPRC) and classification (ACC, FPR) performance, while maintaining comparable FNR. It also reduces predictive uncertainty, as shown by lower M.APSS, PC APSS, and NC APSS values. These results demonstrate that CMRM enhances model robustness and uncertainty estimation under noisy supervision without requiring any noise priors.

Dataset	Method	Measure							
		AUROC (↑)	AUPRC (↑)	FNR (↓)	FPR (↓)	ACC (↑)	M.APSS (↓)	PC APSS (↓)	NC APSS (↓)
Adult	LR	0.784	0.885	**0.073**	0.571	0.802	1.223	1.154	1.432
	LR + CMRM	**0.852**	**0.925**	0.082	**0.422**	**0.833**	**1.209**	**1.109**	**1.308**
	Focal	0.809	0.890	0.136	0.388	0.801	1.257	1.224	1.356
	Focal + CMRM	**0.872**	**0.942**	**0.128**	**0.324**	**0.823**	**1.221**	**1.148**	**1.295**
	SVM	0.808	0.925	**0.029**	0.807	0.776	1.276	1.370	1.512
	SVM + CMRM	**0.847**	**0.937**	0.048	**0.585**	**0.817**	**1.199**	**1.322**	**1.343**
	GCE	0.819	0.904	**0.119**	0.424	**0.804**	1.286	**1.176**	1.396
	GCE + CMRM	**0.846**	**0.928**	0.172	**0.286**	0.800	**1.273**	1.207	**1.340**
Email	LR	0.831	0.869	**0.246**	0.200	0.773	1.405	1.415	1.392
	LR + CMRM	**0.875**	**0.910**	0.259	**0.129**	**0.793**	**1.335**	**1.247**	**1.291**
	Focal	0.833	0.858	**0.228**	0.208	0.780	1.404	1.449	1.345
	Focal + CMRM	**0.907**	**0.916**	0.246	**0.080**	**0.822**	**1.235**	**1.152**	**1.202**
	SVM	0.952	0.964	**0.043**	0.219	0.885	1.029	1.003	1.001
	SVM + CMRM	**0.954**	**0.967**	0.060	**0.125**	**0.913**	**0.975**	**0.996**	**0.993**
	GCE	0.931	0.925	**0.069**	0.101	0.918	**0.976**	**0.977**	**0.975**
	GCE + CMRM	**0.938**	**0.933**	0.074	**0.082**	**0.922**	0.984	0.982	0.986
Credit	LR	0.690	0.866	0.134	0.600	0.764	1.468	1.414	1.522
	LR + CMRM	**0.714**	**0.877**	**0.121**	**0.565**	**0.782**	**1.400**	**1.365**	**1.436**
	Focal	0.673	0.862	**0.169**	0.588	0.739	**1.507**	**1.467**	1.547
	Focal + CMRM	**0.677**	**0.861**	0.176	**0.547**	**0.743**	1.527	1.543	**1.511**
	SVM	0.688	0.856	**0.079**	0.618	0.803	**1.387**	**1.335**	1.438
	SVM + CMRM	**0.711**	**0.869**	0.084	**0.553**	**0.813**	1.421	1.437	**1.405**
	GCE	0.671	0.861	**0.134**	0.845	0.751	1.446	1.366	1.526
	GCE + CMRM	**0.701**	**0.875**	**0.115**	**0.627**	**0.773**	**1.417**	1.339	**1.495**

Table 9: **Performance comparison on binary classification.** We evaluate three datasets (Adult, Email, and Credit) under a 20% label noise setting. Models are assessed across accuracy (ACC), ranking (AUROC, AUPRC), calibration (FNR, FPR), and uncertainty metrics: marginal average prediction set size (M. APSS), positive-class prediction set size (PC APSS), and negative-class prediction set size (NC APSS). ↑ means higher is better; ↓ lower is better. Best results for each base model (LR, Focal, SVM, GCE) are in **bold**. Our method consistently improves uncertainty estimation (M.APSS, PC APSS, NC APSS), ranking metrics (AUROC, AUPRC), and accuracy (ACC, FPR), while slightly increasing FNR.

Data	loss	Architecture	Epochs	η	λ^+	λ^-	α^+	α^-
Adult	LR+CMRM	MLP	150	0.01	0.4	0.5	0.3	0.1
	Focal+CMRM	MLP	150	0.01	0.1	0.4	0.3	0.5
	SVM+CMRM	SVM	150	0.01	0.15	1.5	0.2	0.5
	GCE+CMRM	MLP	150	0.001	0.0	0.2	0.1	0.5
Email	LR+CMRM	MLP	150	0.01	0.7	0.2	0.2	0.2
	Focal+CMRM	MLP	150	0.01	0.4	0.6	0.1	0.1
	SVM+CMRM	SVM	150	0.01	0.3	1.0	0.3	0.3
	GCE+CMRM	MLP	150	0.01	0.9	0.7	0.4	0.4
Credit	LR+CMRM	MLP	150	0.005	0.0	0.3	0.1	0.1
	Focal+CMRM	MLP	150	0.001	0.05	0.3	0.3	0.1
	SVM+CMRM	SVM	150	0.001	0.0	1.0	0.1	0.5
	GCE+CMRM	MLP	150	0.0005	0.0	0.3	0.1	0.1

Table 10: **The details we used to train our models for binary classification**. We reported the hyperparameters that give the best combination os metrics.

CMRM Loss convergences. To verify the stability and trainability of our method, we monitor the learning dynamics of CMRM during optimization. As shown in the top row of Figure 10, both the classification loss and CMRM loss steadily decrease and stabilize, indicating that the joint objective converges smoothly. The CMRM regularization term integrates seamlessly with standard classification training and does not introduce optimization instability or slow down convergence. This confirms that CMRM can be efficiently optimized using standard gradient-based methods and is compatible with commonly used loss functions.

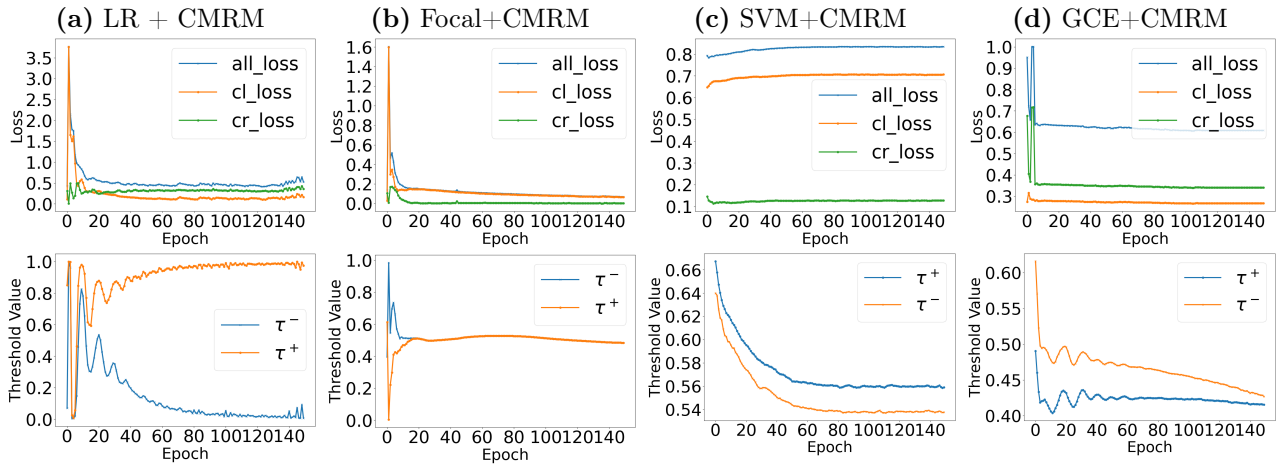

Figure 10: **Dynamics of losses (Top Row) and two thresholds (τ^+ and τ^-, Bottom Row) for binary classification of all base loss (LR, Focal, SVM, GCE) with CMRM** on Email datasets. **Top row** shows the training dynamics of total loss (all loss), classification loss (cl loss), and CMRM regularization loss (CMRM loss) over epochs. CMRM exhibits stable and monotonic convergence alongside standard loss components. **Bottom row** shows τ^+ (negative class threshold) and τ^- (positive class threshold) of LR+CMRM during training. The separation between the thresholds increases, indicating that CMRM actively maximizes the margin between favorable and unfavorable classes.

Margin $\tau^- - \tau^+$ of CMRM grows over Time. The bottom row of Figure 10 illustrates the evolution of the class-conditional thresholds τ^+ and τ^- over training. These thresholds define the CMRM margin region, with τ^- indicating the lower bound for confident positives and τ^+ the upper bound for confident negatives. As training progresses, we observe that τ^- increases while τ^+ decreases, leading to an expanding margin between the two thresholds. This demonstrates that CMRM successfully enforces separation between favorable and unfavorable classes in high-confidence regions, which is critical for robustness under posterior shift.