
Direct Prediction Set Minimization via Bilevel Conformal Classifier Training

Yuanjie Shi^{*1} Hooman Shahrokhi^{*1} Xuesong Jia¹ Xiongzhi Chen² Janardhan Rao Doppa¹ Yan Yan¹

Abstract

Conformal prediction (CP) is a promising uncertainty quantification framework which works as a wrapper around a black-box classifier to construct prediction sets (i.e., subset of candidate classes) with provable guarantees. However, standard calibration methods for CP tend to produce large prediction sets which makes them less useful in practice. This paper considers the problem of integrating conformal principles into the training process of deep classifiers to directly minimize the size of prediction sets. We formulate conformal training as a bilevel optimization problem and propose the *Direct Prediction Set Minimization (DPSM)* algorithm to solve it. The key insight behind DPSM is to minimize a measure of the prediction set size (upper level) that is conditioned on the learned quantile of conformity scores (lower level). We analyze that DPSM has a learning bound of $O(1/\sqrt{n})$ (with n training samples), while prior conformal training methods based on stochastic approximation for the quantile has a bound of $\Omega(1/s)$ (with batch size s and typically $s \ll \sqrt{n}$). Experiments on various benchmark datasets and deep models show that DPSM significantly outperforms the best prior conformal training baseline with 20.46% ↓ in the prediction set size and validates our theory.

1. Introduction

Deep neural networks have achieved high prediction accuracy and enabled numerous applications in diverse domains including computer vision (He et al., 2016; 2017) and natural language processing (Vaswani et al., 2017; Kumar et al.,

^{*}Equal contribution ¹School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington, USA ²Department of Mathematics and Statistics, Washington State University, Pullman, Washington, USA. Correspondence to: Yan Yan <yan.yan1@wsu.edu>, Janardhan Rao Doppa <jana.doppa@wsu.edu>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

2016). However, to safely deploy deep machine learning (ML) classifiers in high-stakes applications (e.g., medical diagnosis (Begoli et al., 2019; Yang & Fevens, 2021)) and to build efficient human-ML collaborative systems (e.g., a classifier produces a small set of decisions and a human selects from them (Cresswell et al., 2024; Straitouri et al., 2023; Babbar et al., 2022)), we need reliable uncertainty quantification (UQ) (Abdar et al., 2021). UQ could take the form of a prediction set (a subset of classes) for classification tasks. For example, in medical diagnosis (Begoli et al., 2019; Yang & Fevens, 2021), such prediction sets allow a doctor to rule out harmful diagnoses such as stomach cancer, even if the most likely diagnosis is a stomach ulcer.

Conformal prediction (CP) is a general framework that provides finite-sample guarantees of *valid* prediction sets to include the correct output, which are agnostic to the ML model and data distribution (Vovk et al., 2005; Angelopoulos & Bates, 2021). As a result, CP is increasingly used for UQ in real-world problems (Cresswell et al., 2024; Lu et al., 2022b;a). The general principle behind CP is to convert the output of an ML model on a given input example (e.g., softmax scores of a deep classifier) into a prediction set (Romano et al., 2020) that contains the true label with a user-specified probability (e.g., 90% probability) referred to as *coverage* via a calibration step. One of the main limitations of standard CP is that it tends to produce large prediction sets, which may not be useful in practice (Vovk et al., 2016; Babbar et al., 2022; Straitouri et al., 2023). Recent prior work focused on how to improve the predictive efficiency of CP (Fisch et al., 2021a;b; Angelopoulos et al., 2021; Huang et al., 2023a; Ding et al., 2024; Shi et al., 2024). However, their focus is on the calibration step only by considering the underlying ML model as a black-box.

Recent work has considered integrating CP into classifier training to trade-off between prediction accuracy and *conformal alignment* (e.g., minimizing the size of prediction sets) in an end-to-end fashion (Stutz et al., 2021; Einbinder et al., 2022; Yan et al., 2024). For instance, ConfTr (Stutz et al., 2021) incorporates a stochastic approximation (SA) of the average prediction set size into the training objective function based on empirical quantiles over mini-batches. To achieve accurate conditional coverage, conformal uncertainty-aware training (CUT) (Einbinder et al., 2022) employs a similar SA approach to acquire the empiri-

cal quantiles of non-conformity scores on mini-batches and penalizes their deviation from the uniformity. However, a primary challenge for these SA-based conformal training methods is the large error between the empirical batch-level quantiles and the true quantiles, resulting in a learning bound of up to $O(1/\sqrt{s})$ with batch size s (Einbinder et al., 2022). This bound is significantly larger than the standard learning bound of $O(1/\sqrt{n})$ with n training samples in empirical risk minimization (ERM) (Mohri et al., 2018; Shalev-Shwartz & Ben-David, 2014). Consequently, it notably degenerates the overall trade-off between the prediction accuracy and conformal alignment of the trained classifier. This motivates the main question of this paper: *can we develop a conformal classifier training method to calibrate the conformal uncertainty with a standard $O(1/\sqrt{n})$ learning bound?*

This paper gives an affirmative answer by developing a novel conformal classifier training algorithm referred to as *Direct Prediction Set Minimization (DPSM)*. The key idea behind DPSM is to formulate conformal training as a bilevel optimization problem (Ghadimi & Wang, 2018; Zhang et al., 2024) that estimates the empirical quantile of non-conformity scores by learning a quantile regression (QR) in the lower-level subproblem. Conditioned on this learned quantile, the upper-level subproblem measures and minimizes the prediction set size of CP. This bilevel formulation implicitly defines a conformal alignment objective and we analyze that minimizing the resulting function leads to a learning bound of $O(1/\sqrt{n})$. In contrast, we show that the learning bound of the prior SA-based conformal training methods is lower bounded by $\Omega(1/s)$ with batch size s (typically $s \ll \sqrt{n}$). Moreover, to optimize the DPSM objective, we develop a simple and practical stochastic first-order algorithm. Our experiments on diverse real-world datasets with different conformity scores and deep models demonstrate that DPSM achieves significant reduction ($\downarrow 20.46\%$) in prediction set sizes compared to the best baseline.

Contributions. The key contributions of this paper include:

- We develop a new conformal training algorithm DPSM as a bilevel formulation, which directly minimizes the prediction set size (upper level) conditioned on the learned quantile of the conformity scores (lower level).
- We analyze the learning bound of DPSM and find that it is bounded above by $O(1/\sqrt{n})$, which significantly improves over the existing conformal training methods whose lower bound is at least $\Omega(1/s)$.
- We develop a simple stochastic first-order algorithm to optimize the DPSM training objective.
- Experiments on multiple benchmark datasets to demonstrate significant improvements of DPSM over the best prior conformal training baseline (20.46% \downarrow prediction set size) and to validate our theoretical results.

The DPSM code is available at https://github.com/YuanjieSh/DPSM_code.

2. Related Work

Conformal Prediction has been extensively studied (Angelopoulos & Bates, 2021) recently. Key contributions include laying the foundational principles in CP (Vovk et al., 2005; 1999; Shafer & Vovk, 2008). There is some focus on cross-validation methods (Vovk, 2015) and the jackknife+ approach (Barber et al., 2021), but split CP is the most common approach (Vovk et al., 2005; Romano et al., 2020; Oliveira et al., 2024), where calibration of a given pre-trained model is performed on a held-out dataset. CP is applied in many tasks including classification (Romano et al., 2020; Angelopoulos et al., 2021; Cauchois et al., 2021), regression (Romano et al., 2019; Gibbs et al., 2023; Gibbs & Candes, 2021; Sesia & Romano, 2021), computer vision (Angelopoulos et al., 2024; 2022; Bates et al., 2021), adversarial attacks (Liu et al., 2024; Ghosh et al., 2023b), online learning (Angelopoulos et al., 2023; Bhatnagar et al., 2023), and generative tasks (Shahrokhi et al., 2025). Valid coverage and predictive efficiency are two critical and often competing evaluation measures in CP (Angelopoulos et al., 2021). Recent CP advances include the development of new conformity scores (Angelopoulos et al., 2021; Huang et al., 2023a) and calibration procedures (Fisch et al., 2021b;a; Guan, 2023; Ghosh et al., 2023a; Ding et al., 2024; Kiyani et al., 2024b; Shi et al., 2024; Zhu et al., 2025) to improve predictive efficiency with valid coverage. These methods follow the general CP workflow by relying on pre-trained models. Consequently, their predictive efficiency critically depends on the underlying ML model. However, since the ML models are not specifically trained for conformal alignment, CP may produce large prediction sets (Bellotti, 2021).

Conformal Training aims at integrating CP into the classifier training process to promote conformal alignment. A typical approach is to introduce a regularization function that integrates conformity property into the training process (Bellotti, 2021; Stutz et al., 2021; Einbinder et al., 2022; Yan et al., 2024). For example, ConfTr (Stutz et al., 2021) integrates CP into the training of deep classifiers through a differentiable approximation for the average prediction set size, allowing for smaller and valid prediction sets using the CP-aware classifier. Conformalized uncertainty-aware training (CUT) (Einbinder et al., 2022) combines the conformity scores with a uncertainty-aware loss function for accurate conditional coverage. It penalizes the deviation of the conformity scores from the uniformity during the training process. However, both ConfTr and CUT employ a stochastic quantile of non-conformity scores from a mini-batch of s samples, resulting in a large $O(1/\sqrt{s})$ error with respect to the true quantile depending on the batch size s .

Bilevel Optimization has been extensively studied due to its critical role in many ML use-cases (Liu et al., 2021), including reinforcement learning (Cheng et al., 2022; Zhou, 2024; Shen et al.), federated learning (Huang et al., 2023b; Yang et al., 2023), and continual learning (Borsos et al., 2020; Zhou et al., 2022; Hao et al., 2024; Borsos et al., 2024). Bilevel optimization methods can be roughly categorized into three groups: (i) implicit gradient methods (Chen et al., 2021; Ji et al., 2021; Dagiéou et al., 2022; Ghadimi & Wang, 2018; Pedregosa, 2016; Domke, 2012), (ii) iterative differentiation methods (Bolte et al., 2022; 2021), and (iii) penalty-based methods (Lu & Mei, 2024; Chen et al., 2024b;a; Shen & Chen, 2023; 2025). Compared to penalty-based methods, implicit gradient methods typically require more restrictive assumptions, such as twice differentiable and strongly-convex lower-level function (Shen & Chen, 2025). On the other hand, iterative differentiation methods typically require an iterative subroutine for the lower-level problem with more computational cost (Shen & Chen, 2025). Instead, penalty-based methods build a single level problem and only use first-order information (Shen & Chen, 2023). To summarize, bilevel optimization algorithms either require restrictive assumptions or have complex update steps. It is non-trivial to find a practical algorithm that is particularly tailored for our proposed conformal training problem.

Our goal is to develop a bilevel conformal training algorithm with (i) a standard $O(1/\sqrt{n})$ learning bound for n training samples to improve over $O(1/\sqrt{s})$ for batch size s from prior conformal training methods, and (ii) a practical stochastic first-order bilevel optimization algorithm.

3. Background and Motivation

Notations. Suppose $X \in \mathcal{X}$ is an input from \mathcal{X} , and $Y \in \mathcal{Y} = \{1, 2, \dots, K\}$ is the ground-truth label, where K is the number of candidate classes. Let (X, Y) be a data sample drawn from an underlying distribution \mathcal{P} defined on $\mathcal{X} \times \mathcal{Y}$. Let $f(X) : \mathcal{X} \rightarrow \Delta_+^K$ denote the confidence prediction by soft classifier $f \in \mathcal{F}$ (e.g., Softmax scores), where Δ_+^K is the $(K-1)$ -dimensional probability simplex, $f(X)_y$ is the confidence score of class y and \mathcal{F} denotes a hypothesis class. Define $\mathcal{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^n$ as the training set of n samples for training classifier f . We denote \mathcal{D}_{cal} and $\mathcal{D}_{\text{test}}$ as two separate calibration and testing datasets, respectively. We assume that there are m calibration samples, i.e., $|\mathcal{D}_{\text{cal}}| = m$. Define $\mathbb{1}[\cdot]$ as an indicator function. Denote $\mathcal{B} = \{(X_i, Y_i) | i \in I_s\}$ as a randomly sampled batch of training data of size s , such that the batch index set $I_s \subset \{1, 2, \dots, n\}$ with $|I_s| = s$. Given a function $h(x, y)$, we denote $\nabla_x h(x, y)$ as the directional derivative of h in x and $\widehat{\nabla}_x h(x, y)$ as its stochastic estimation.

Conformal Prediction. As mentioned above, CP calibrates predictions from a given model based on a *non-conformity*

scoring function. Let $S : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ denote a non-conformity scoring function to measure the difference between a new data and existing ones (Vovk et al., 2005). For simplicity of notation, we denote the non-conformity score of the i -th example as $S_i = S_f(X_i, Y_i)$ for $(X_i, Y_i) \in \mathcal{D}_{\text{tr}}$ and of the j -th smallest value in $\{S_i\}_{i=1}^n$ as $S_{(j)}$. Many non-conformity scoring functions are proposed in prior work (Huang et al., 2023a; Angelopoulos et al., 2021; Shafer & Vovk, 2008). In this paper, we consider Homogeneous Prediction Sets (HPS) (Sadinle et al., 2019), Adaptive Prediction Sets (APS) (Romano et al., 2020), and Regularized Adaptive Prediction Sets (RAPS) (Angelopoulos et al., 2021) (details are in Appendix A). We assume that there is no tie in non-conformity scores (Romano et al., 2020).

Given the mis-coverage parameter α and an underlying model f , on a testing data $(X_{\text{test}}, Y_{\text{test}})$, CP aims to achieve a coverage of the true label with a prediction set $\mathcal{C}_f : \mathcal{X} \rightarrow 2^{|\mathcal{Y}|}$ with probability at least $1 - \alpha$, i.e.,

$$\mathbb{P}_{(X_{\text{test}}, Y_{\text{test}}) \sim \mathcal{P}}\{Y_{\text{test}} \in \mathcal{C}_f(X_{\text{test}})\} \geq 1 - \alpha. \quad (1)$$

CP typically constructs $\mathcal{C}_f(X_{\text{test}})$ via the empirical quantile $\widehat{Q}_f(\alpha)$ on the calibration set \mathcal{D}_{cal} :

$$\mathcal{C}_f(X_{\text{test}}) = \{y \in \mathcal{Y} : S_f(X_{\text{test}}, y) \leq \widehat{Q}_f(\alpha)\}, \quad (2)$$

where $\widehat{Q}_f(\alpha)$ is computed as the $\lceil (1-\alpha)(1+m) \rceil$ th smallest value in $\{S_{f,i}\}_{i=1}^m$ and is an estimation to the population quantile $Q_f(\alpha) = \min\{\tau : \mathbb{P}_{X,Y}[S_f(X, Y) \leq \tau] \geq 1 - \alpha\}$. Throughout this paper, unless otherwise specified, we omit the target mis-coverage α for notation simplicity, e.g., Q_f .

Quantile-based Conformal Training. The goal is to align the classification model f with favorable conformal principles (Bellotti, 2021). A general framework (Stutz et al., 2021; Einbinder et al., 2022) is to establish the trade-off between prediction accuracy and conformal alignment on the training set \mathcal{D}_{tr} with an objective structured as follows:

$$\min_f \mathcal{L}_{\text{cls}}(f) + \lambda \cdot \mathcal{L}_c(f), \quad (3)$$

where $\mathcal{L}_{\text{cls}}(f)$ is the classification loss (e.g., cross entropy), $\mathcal{L}_c(f)$ is the conformal alignment loss based on the quantile Q_f , and λ is the regularization hyper-parameter. Following (Stutz et al., 2021), this paper focuses on a concrete definition of $\mathcal{L}_c(f)$ as the expected differentiable size of prediction sets:

$$\mathcal{L}_c(f) = \ell(f, Q_f), \text{ such that} \quad (4)$$

$$\ell(f, q) \triangleq \mathbb{E}_X \left[\underbrace{\sum_{y \in \mathcal{Y}} \tilde{\mathbb{1}}[S_f(X, y) \leq q]}_{\text{differentiable } \mathbb{E}_X[|\mathcal{C}_f(X)|]} \right],$$

where $\tilde{\mathbb{1}}[\cdot]$ is a smoothed estimator for the indicator function $\mathbb{1}[\cdot]$ and defined by a Sigmoid function (Stutz et al., 2021),

i.e., $\tilde{\mathbb{1}}[S \leq q] = 1/(1 + \exp(-(q - S)/\tau_{\text{Sigmoid}}))$ with a tunable temperature parameter τ_{Sigmoid} . Here, we denote $\ell(f, q)$ as a general conformal loss function that takes any f and q as inputs, and $\mathcal{L}_c(f)$ requires the true quantile Q_f as the input to ℓ .

However, it is challenging to accurately compute $\mathcal{L}_c(f, Q_f)$ when training f , as Q_f is an implicit function of f . On one hand, updating f iteratively also changes Q_f accordingly. On the other hand, after getting an updated f , re-computing an empirical quantile \hat{Q}_f on training data \mathcal{D}_{tr} requires sorting all n non-conformity scores of training data, with a computational complexity of $O(n \log(n))$.

To address the above-mentioned challenge, prior conformal training methods (Stutz et al., 2021; Einbinder et al., 2022) employ a stochastic approximation (SA) method to estimate $\mathcal{L}_c(f)$ in batches. Specifically, at each iteration during training, they use the empirical batch-level quantile from a mini-batch \mathcal{B} sampled from \mathcal{D}_{tr} , denoted by \hat{q}_f , as the input to ℓ , i.e., $\ell(f, \hat{q}_f)$. Since \hat{q}_f depends on the randomly sampled batches, we regard \hat{q}_f as a random variable drawn from its underlying distribution Q_f , i.e., $\hat{q}_f \sim Q_f$ (we defer our analysis for Q_f to Proposition 3.4 stated below). Then the in-effect SA-based conformal loss $\hat{\mathcal{L}}_c^{\text{SA}}(f)$ is given by

$$\begin{aligned} \hat{\mathcal{L}}_c^{\text{SA}}(f) &= \mathbb{E}_{\hat{q}_f \sim Q_f} [\hat{\ell}(f, \hat{q}_f)], \text{ such that} \\ \hat{\ell}(f, q) &\triangleq \frac{1}{n} \sum_{i=1}^n \left[\sum_{y \in \mathcal{Y}} \tilde{\mathbb{1}}[S_f(X_i, y) \leq q] \right], \end{aligned} \quad (5)$$

where $\hat{\ell}(f, q)$ is an empirical version of $\ell(f, q)$ on \mathcal{D}_{tr} .

A key limitation of the SA method is the large error gap between the true and in-effect conformal alignment loss functions, i.e., $|\hat{\mathcal{L}}_c^{\text{SA}}(f) - \mathcal{L}_c(f)|$. Prior work (Einbinder et al., 2022) showed that the error of SA-based conformal alignment loss is upper bounded by $O(1/\sqrt{s})$ with batch size s . We further analyze the learning error for the SA-based conformal alignment loss in this paper, including its lower bound. We start with the following assumptions.

Assumption 3.1. (Bi-Lipschitz continuity of score $S_{(j)}$) $S_{(j)}$ is bi-Lipschitz continuous for normalized order j/n such that $L_1 |\frac{j_1}{n} - \frac{j_2}{n}| \leq |S_{(j_1)} - S_{(j_2)}| \leq L_2 |\frac{j_1}{n} - \frac{j_2}{n}|$ for $L_1, L_2 > 0$.

Assumption 3.2. (μ -strongly concavity of $\hat{\ell}(f, q)$) $\hat{\ell}(f, q)$ is μ -strongly concave locally around $\mathbb{E}[\hat{q}_f]$ for fixed f .

Remark 3.3. We conduct experiments to empirically verify the Assumptions 3.1 and 3.2 (See Fig 4). These assumptions allow us to analyze how the SA-based conformal training method approaches to the true conformal alignment loss.

The following proposition reveals the specification of the distribution Q_f aforementioned in $\hat{\mathcal{L}}_c^{\text{SA}}$ (5) that considers \hat{q}_f as a random variable drawn from the training dataset \mathcal{D}_{tr} .

Proposition 3.4. (Distribution of mini-batch quantiles) Define an event $Z(j)$ as “the j -th smallest score $S_{(j)}$ from $\{S_i\}_{i=1}^n$ is randomly selected as $\hat{q}^s(\alpha)$ on a mini-batch \mathcal{B} ”. Then the probability of $Z(j)$ is:

$$\mathbb{P}(Z(j)) = \frac{\binom{j}{\lceil(1-\alpha)(s+1)\rceil} \binom{n-j-1}{s-\lceil(1-\alpha)(s+1)\rceil-1}}{\binom{n}{s}}.$$

Furthermore, we have the following asymptotic result

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z(j)) = \frac{e}{n} \mathbb{P}_{\text{Beta}}\left(\frac{j}{n}; \lceil(1-\alpha)(s+1)\rceil + 1; s - \lceil(1-\alpha)(s+1)\rceil\right),$$

where $\mathbb{P}_{\text{Beta}}(x; a; b)$ is the PDF of Beta distribution with two shape parameters a, b and e is the Euler's number.

With the above proposition that captures the random sampling process of the quantile \hat{q}_f in each mini-batch, we are ready to present the following main result analyzing the learning bound of the SA-based conformal training method.

Theorem 3.5. (Learning bounds of SA method) Suppose that Assumption 3.1 and 3.2 hold. Assume $(1-\alpha)(s+1)$ is not an integer. If $\lceil(1-\alpha)(s+1)\rceil - (1-\alpha)(s+1) \geq \Omega(1/s)$ and $s \leq \sqrt{n}$, then the following inequality holds with probability at least $1 - \delta$:

$$\Omega(1/s) \leq |\hat{\mathcal{L}}_c^{\text{SA}}(f) - \mathcal{L}_c(f)| \leq \tilde{O}(1/\sqrt{s}).$$

Remark 3.6. Throughout this paper, we use \tilde{O} to suppress the log dependency. The above result shows the upper bound $\tilde{O}(1/\sqrt{s})$ and lower bound $\Omega(1/s)$ of learning error of the SA-based conformal training method. It is worth noting that even the lower bound $\Omega(1/s)$ can be significantly larger than the standard result $O(1/\sqrt{n})$ as in empirical risk minimization (Mohri et al., 2018) under a common setting $s \leq \sqrt{n}$ for training deep models (Masters & Luschi, 2018).

4. Direct Prediction Set Minimization Method

In this section, we first propose **Direct Prediction Set Minimization (DPSM)** algorithm by formulating it as a bilevel optimization problem. Next, we show that the learning bound of DPSM is at most $O(1/\sqrt{n})$ (Theorem 4.1), which improves over the SA-based conformal training methods ($\Omega(1/s)$, Theorem 3.5). Finally, we develop a simple and practical optimization algorithm to solve it.

4.1. Bilevel Problem Formulation and Learning Bound

The key idea to directly minimize the prediction sets is inspired by quantile regression (QR), a well-studied method to learn the quantile of a set of random variables. A common approach to training QR is to minimize an average pinball loss on a set of data (Narayan et al., 2024; Gibbs et al.,

2023; Koenker, 2005; Koenker & Bassett Jr, 1978). For the $(1 - \alpha)$ -quantile, the pinball loss is defined as follows.

$$\rho_\alpha(q, S) = \begin{cases} (1 - \alpha)(S - q), & \text{if } S \geq q, \\ \alpha(q - S), & \text{otherwise,} \end{cases} \quad (6)$$

where S and q represent a real value and quantile prediction, respectively. Minimizing the average pinball loss on a set of data $\{S_i\}_{i=1}^n$ gives the $(1 - \alpha)$ -quantile of n scores $\{S_i\}_{i=1}^n$:

$$q^* \in \arg \min_q \frac{1}{n} \sum_{i=1}^n \rho_\alpha(q, S_i). \quad (7)$$

Conformal Training via Bilevel Optimization. Instead of using a stochastic quantile on each batch of data, we propose to incorporate QR into the conformal training objective as a constraint. Conditioned on solving this QR-based constraint and acquiring the accurate quantile, we directly minimize the average size of prediction sets. Specifically, we formulate the following *bilevel optimization* problem for our DPSM method for conformal training:

$$\begin{aligned} & \min_{f,q} \widehat{\mathcal{L}}_{\text{cls}}(f) + \lambda \cdot \widehat{\mathcal{L}}_c^{\text{DM}}(f, q) \\ & \text{s.t. } q \in \mathcal{U}(f) \triangleq \arg \min_{q'} \widehat{\mathcal{L}}^{\text{QR}}(f, q'), \end{aligned} \quad (8)$$

where $\widehat{\mathcal{L}}^{\text{QR}}(f, q) = \frac{1}{n} \sum_{i=1}^n \rho_\alpha(q, S_f(X_i, Y_i))$ in the lower level is the average pinball loss for the non-conformity scores on n training samples, and the conformal alignment loss $\widehat{\mathcal{L}}_c^{\text{DM}}(f, q)$ is given by

$$\widehat{\mathcal{L}}_c^{\text{DM}}(f, q) = \frac{1}{n} \sum_{i=1}^n \left[\sum_{y \in \mathcal{Y}} \mathbb{1}[S_f(X_i, y) \leq q] \right], \quad (9)$$

which follows the definition of $\widehat{\mathcal{L}}_c^{\text{SA}}(f)$ in (5) except that it allows any variable q as the input quantile, instead of a stochastic quantile $\widehat{q}_f \sim \mathcal{Q}_f$. To analyze DPSM, following the bilevel optimization literature (Chen et al., 2024a; Shen & Chen, 2023), we define the *implicit conformal loss*:

$$\bar{\mathcal{L}}_c^{\text{DM}}(f) \triangleq \min_{q \in \mathcal{U}(f)} \widehat{\mathcal{L}}_c^{\text{DM}}(f, q).$$

The key innovation in the proposed bilevel conformal training problem (8) compared with the SA-based method is to *explicitly parameterize the quantile* with a single score variable q . This approach has been used in the CP literature (Gibbs et al., 2023; Gibbs & Candès, 2024; Kiyani et al., 2024a; Deshpande et al., 2024; Podkopaev et al., 2024), and we employ it for conformal training in our bilevel formulation for the first time. Conditioned on q , regardless of how accurate q is during training relative to the true Q_f , the conformal alignment loss $\widehat{\mathcal{L}}_c^{\text{DM}}(f, q)$ participates in training.

The major benefit of the quantile parameterization idea is that it *decouples the quantile q from the stochastic batches*

during training f , rather than the immediate dependency as in the SA-based conformal training, i.e., $\widehat{q}_f \sim \mathcal{Q}_f$. Specifically, QR in the lower-level subproblem, due to its well-known property, enables the iterative updates for q until attaining the $(1 - \alpha)$ -quantile of non-conformity scores as training f . If an optimization algorithm ensures the simultaneous convergence of both variables in Problem (8) to their optimal solution sets (discussed in Section 4.2), i.e.,

$$q \rightarrow \mathcal{U}(f), \quad f \rightarrow \arg \min_f \widehat{\mathcal{L}}_{\text{cls}}(f) + \lambda \cdot \bar{\mathcal{L}}_c^{\text{DM}}(f),$$

then we can achieve an improved learning bound for the conformal alignment loss $|\bar{\mathcal{L}}_c^{\text{DM}}(f) - \mathcal{L}_c(f)|$ that is independent from batch size s . This is the reason why we named our method as *direct prediction set minimization (DPSM)*.

Improved Learning Bound of DPSM. As in Theorem 3.5, we show that the learning error of SA-based method $|\widehat{\mathcal{L}}_c^{\text{SA}}(f) - \mathcal{L}_c(f)|$ is upper bounded by $O(1/\sqrt{s})$ and lower bounded by $\Omega(1/s)$, respectively. We then show that the learning error of DPSM $|\bar{\mathcal{L}}_c^{\text{DM}}(f) - \mathcal{L}_c(f)|$ can be bounded above by $O(1/\sqrt{n})$ in the following result.

Theorem 4.1. (*Learning bound of DPSM*) Suppose Assumption 3.1 holds. For any classifier model $f \in \mathcal{F}$, with probability at least $1 - \delta$, we have:

$$|\bar{\mathcal{L}}_c^{\text{DM}}(f) - \mathcal{L}_c(f)| \leq \tilde{O}(1/\sqrt{n}).$$

Remark 4.2. The above theorem directly answers the central question that we asked in the introduction, i.e., bilevel conformal training calibrates the conformal uncertainty with a standard $O(1/\sqrt{n})$ learning bound as in ERM (Mohri et al., 2018; Shalev-Shwartz & Ben-David, 2014). In comparison with the learning error bounds of the SA-based conformal training method derived in Theorem 3.5, the DPSM bound improves by dropping the dependency on the batch size s .

The improved learning bound of DPSM over SA-based methods is mainly due to smaller error in estimating the true quantile, i.e., $|q - Q_f|$. Indeed, results shown in Fig 7 (c) and (d) empirically demonstrates this hypothesis. The practical significance of this result is that the smaller estimation error of DPSM results in improved conformal alignment of the classifier leading to smaller prediction set sizes.

4.2. Stochastic Optimization for DPSM

Challenges for DPSM. There is a large literature on methods for bilevel optimization (Liu et al., 2021; Beck et al., 2023). However, existing methods typically require restrictive assumptions to guarantee convergence. Some examples of such assumptions include requiring Lipschitz Hessians (Chen et al., 2024a; 2021; Dagréou et al., 2022; Yang et al., 2021a); twice continuously differentiable (Kwon et al., 2024); requiring solving a subproblem (Liu et al., 2020); strongly convex lower-level subproblem (Gong et al., 2024b;

Algorithm 1 Direct Prediction Set Minimization (DPSM)

```

1: Input: Training dataset  $\mathcal{D}_{\text{tr}}$ , regularization parameter  $\lambda$ ,  

   learning-rate  $\eta, \gamma > 0$ , mis-coverage level  $\alpha$   

2: Randomly initialize the deep neural network  $f_0$  and  

   quantile regression model parameter  $q_0 \in \mathbb{R}$   

3: Randomly split  $\mathcal{D}_{\text{tr}}$  into two disjoint subsets  $\mathcal{D}_1$  and  $\mathcal{D}_2$   

   with the same size such that  $|\mathcal{D}_1| = |\mathcal{D}_2| = \frac{n}{2}$   

4: for  $t \leftarrow 0 : T - 1$  do  

5:   Randomly sample two batches  $\mathcal{B}_t^1 \subset \mathcal{D}_1, \mathcal{B}_t^2 \subset \mathcal{D}_2$   

6:   Compute  $\widehat{\nabla}_{f,t} \leftarrow \widehat{\nabla}_f \widehat{\mathcal{L}}_{\text{cls}}(f_{t-1})$  on batch  $\mathcal{B}_t^1$   

7:   Compute  $\widehat{\nabla}_{q,t}^{\text{QR}} \leftarrow \widehat{\nabla}_q \widehat{\mathcal{L}}^{\text{QR}}(f_{t-1}, q_{t-1})$  on batch  $\mathcal{B}_t^1$   

8:   Compute  $\widehat{\nabla}_{f,t}^{\text{DM}} \leftarrow \widehat{\nabla}_f \widehat{\mathcal{L}}_c^{\text{DM}}(f_{t-1}, q_{t-1})$  on batch  $\mathcal{B}_t^2$   

9:    $f_{t+1} \leftarrow f_t - \eta(\widehat{\nabla}_{f,t} + \lambda \widehat{\nabla}_{f,t}^{\text{DM}})$   

10:   $q_{t+1} \leftarrow q_t - \gamma \widehat{\nabla}_{q,t}^{\text{QR}}$   

11: end for  

12: Output: the trained classification model  $f_T$ 

```

Kwon et al., 2023; Ji et al., 2021; Gong et al., 2024a; Chen et al., 2024c; Hong et al., 2023; Khanduri et al., 2021); convexity in both upper and lower levels (Sabach & Shtern, 2017); unique lower-level solution (Liu et al., 2022); and Lipschitz smoothness (Shen & Chen, 2023) (less restrictive).

In the DPSM problem (8), the standard classification loss $\widehat{\mathcal{L}}_{\text{cls}}(f)$ and the conformal alignment loss $\widehat{\mathcal{L}}_c^{\text{DM}}(f, q)$ are not necessarily convex in f and q . Meanwhile, the QR loss $\widehat{\mathcal{L}}^{\text{QR}}(f, q)$ in the lower level is not continuously differentiable, not strongly convex, and typically has non-unique solution set even for the case without any tie in conformity scores. To the best of our knowledge, there is no prior stochastic optimization algorithm that solves DPSM problem with the assumptions of bilevel optimization literature satisfied and convergence guaranteed. Consequently, it is *non-trivial and an open challenge to develop a simple stochastic gradient optimization algorithm to solve a problem that shares similar conditions with DPSM*. We leave this challenge for future work.

Penalty-based Reformulation. Given that the lower-level function in DPSM is not strongly convex, implicit gradient methods, e.g., (Ghadimi & Wang, 2018), cannot handle it. Penalty-based methods is another large family in the bilevel optimization literature. A recent paper (Chen et al., 2024b) studied a penalty-based method when the lower-level problem satisfies Hölderian error bound, and meanwhile both upper and lower level objectives are non-smooth, which is one of the most relevant setting to the DPSM problem at hand. Hence, we reformulate DPSM as a penalty-based problem and analyze the global solution of the penalty-based DPSM problem with a penalty parameter σ .

$$\begin{aligned} \min_{f,q} & \widehat{\mathcal{L}}_{\text{cls}}(f) + \lambda \cdot \widehat{\mathcal{L}}_c^{\text{DM}}(f, q) \\ & + \sigma(\widehat{\mathcal{L}}^{\text{QR}}(f, q) - \widehat{\mathcal{L}}^{\text{QR}}(f, q^*)), \text{ where } q^* \in \mathcal{U}(f). \end{aligned} \quad (10)$$

Definition 4.3. (Hölderian error bound) A function $h(x)$, where its domain $\text{dom}(h)$ is a closed convex set, satisfies Hölderian error bound if there exists $\nu \geq 1$ and $c > 0$ s.t.

$$\text{dist}(x, X^*)^\nu \leq c(h(x) - \min_{x' \in X^*} h(x')), \forall x \in \text{dom}(h),$$

where $X^* = \arg \min_{x \in \text{dom}(h)} h(x)$ is the optimal solution set for minimizing $h(x)$ and $\text{dist}(x, X^*) = \min_{x'} \|x - x'\|$ denotes the Euclidean distance between x and X^* .

Remark 4.4. Hölderian error bound (HEB) is a well-studied condition in optimization (Pang, 1997; Bolte et al., 2017). It captures the local sharpness of the objective function that helps accelerate the optimization convergence (Roulet & d'Aspremont, 2017; Yang & Lin, 2018). The following result shows that $\widehat{\mathcal{L}}_c^{\text{DM}}(f, q)$ satisfies HEB w.r.t. q .

Lemma 4.5. (HEB for QR loss) Suppose there is no tie in conformity scores $\{S_i\}_{i=1}^n$. Then for a fixed f , $\widehat{\mathcal{L}}^{\text{QR}}(f, q)$ satisfies HEB w.r.t. q for the exponent $\nu = 1$ and $c > 0$.

Corollary 4.6. (Global solution of penalized problem) Suppose $\widehat{\mathcal{L}}_{\text{cls}}(f)$ and $\widehat{\mathcal{L}}_c^{\text{DM}}(f, q)$ is L_{cls} - and L_{DM} -Lipschitz continuous, respectively. For any given $\epsilon > 0$, let $l = (L_{\text{cls}} + L_{\text{DM}})$ and $\sigma = cl$ in (10). Then the ϵ -optimal solution of the penalized problem is an $(\epsilon, \epsilon/l)$ -optimal solution of the original problem.

This is an immediate result from Theorem 2.7 in (Chen et al., 2024b) and connects the solution of the penalized DPSM problem (10) to that of the original DPSM problem (8).

Simple Stochastic Gradient-based Algorithm. We further present a simple stochastic gradient algorithm to optimize Problem (8). We summarize this approach in Algorithm 1. It mainly follows the common schemes in standard stochastic gradient methods to iteratively update the two variables f and q simultaneously. First, it randomly splits the training set \mathcal{D}_{tr} into two disjoint subsets \mathcal{D}_1 and \mathcal{D}_2 with the same size (Line 3). Next, given the fixed model f_{t-1} and quantile variable q_{t-1} , we compute three stochastic gradients over mini-batches in each training iteration: (i) $\widehat{\nabla}_f \widehat{\mathcal{L}}_{\text{cls}}(f_{t-1})$ on mini-batches of \mathcal{D}_1 (Line 6), (ii) $\widehat{\nabla}_f \widehat{\mathcal{L}}_c^{\text{DM}}(f_{t-1}, q_{t-1})$ on mini-batches of \mathcal{D}_2 (Line 8), and (iii) $\widehat{\nabla}_q \widehat{\mathcal{L}}^{\text{QR}}(f_{t-1}, q_{t-1})$ on mini-batches of \mathcal{D}_1 (Line 7). Next, we update the model f_t (Line 9) and quantile q_t (Line 10) accordingly with their learning rates η_t and γ_t .

It is worth noting that the proposed DPSM optimization algorithm is particularly tailored for our bilevel conformal training problem (8). **1)** We employ \mathcal{D}_1 to compute $\widehat{\nabla}_f \widehat{\mathcal{L}}_{\text{cls}}(f)$ and $\widehat{\nabla}_q \widehat{\mathcal{L}}^{\text{QR}}(f, q)$, while evaluate $\widehat{\nabla}_f \widehat{\mathcal{L}}_c^{\text{DM}}(f, q)$ on \mathcal{D}_2 . This has been used in conformal training (Einbinder et al., 2022) and helps to prevent over-fitting. **2)** We rely on simple first-order information to update, rather than hyper-gradient or penalization from the existing bilevel optimization methods. This design keeps the iterative update simple especially for practitioners. Our empirical results in Figure 7 demonstrate

Table 1. HPS → Training, HPS and APS → Calibration/Testing (details in Appendix E): The average prediction set size (APSS) on three datasets with two deep models trained with HPS and calibrated/tested with HPS and APS when $\alpha = 0.1$. \downarrow indicates the percentage improvement in predictive efficiency compared to the best existing method, whereas \uparrow denotes a percentage decrease in predictive efficiency. All results are averaged over 10 different runs, with the mean and standard deviation reported. DPSM significantly outperforms almost all the best baselines with 20.46% reduction in prediction set size on average across all three datasets and two scores.

Model	HPS				APS			
	CE	CUT	ConfTr	DPSM	CE	CUT	ConfTr	DPSM
CalTech-101								
DenseNet	3.50 ± 0.10	1.62 ± 0.030	4.10 ± 0.19	0.90 ± 0.003 (\downarrow 44.44%)	8.44 ± 0.15	3.87 ± 0.11	8.64 ± 0.21	1.58 ± 0.022 (\downarrow 59.17%)
ResNet	1.57 ± 0.018	1.64 ± 0.049	1.52 ± 0.040	0.91 ± 0.005 (\downarrow 44.51%)	4.50 ± 0.059	4.59 ± 0.072	3.61 ± 0.08	1.74 ± 0.031 (\downarrow 51.80%)
CIFAR-100								
DenseNet	2.59 ± 0.053	2.27 ± 0.09	2.28 ± 0.07	2.17 ± 0.086 (\downarrow 4.82%)	3.38 ± 0.12	2.41 ± 0.11	3.08 ± 0.11	2.64 ± 0.086 (\uparrow 8.71%)
ResNet	3.39 ± 0.10	3.01 ± 0.11	3.77 ± 0.14	2.94 ± 0.08 (\downarrow 2.32%)	3.98 ± 0.13	3.81 ± 0.08	4.90 ± 0.18	3.53 ± 0.11 (\downarrow 7.35%)
iNaturalist								
DenseNet	94.58 ± 3.45	77.13 ± 3.72	79.93 ± 3.70	61.22 ± 2.49 (\downarrow 20.63%)	101.97 ± 3.16	88.93 ± 3.06	90.79 ± 3.17	75.98 ± 2.99 (\downarrow 14.56%)
ResNet	99.48 ± 8.95	73.09 ± 2.00	76.73 ± 3.87	70.04 ± 1.99 (\downarrow 4.17%)	95.81 ± 3.80	79.00 ± 2.21	88.70 ± 3.88	79.43 ± 2.39 (\uparrow 0.54%)

stable convergence of both upper and lower level problems of DPSM by employing standard practices for learning rates η and γ .

5. Experiments and Results

This section describes our experimental evaluation of the proposed DPSM algorithm and baselines on real datasets.

5.1. Experimental Setup

Datasets. We utilize the benchmark datasets CIFAR-100 (Krizhevsky et al., 2009), Caltech-101 (Fei-Fei et al., 2004), and iNaturalist (Van Horn et al., 2018), where all details are summarized in Table 2 of Appendix E.

Deep Models. We train two widely used neural network architectures with HPS scoring function: ResNet (He et al., 2016) and DenseNet (Huang et al., 2017). The training hyperparameters are provided in Table 3 of Appendix.

Conformity scoring functions. We consider three non-conformity scoring functions: HPS (Vovk et al., 2005; Lei et al., 2013) APS (Romano et al., 2020), and RAPS (Angelopoulos et al., 2021). The detailed review of these scoring functions is in Appendix A.

Baseline methods. We select three different training methods as baselines: (i) CE, training with standard cross-entropy loss only; (ii) CUT (Einbinder et al., 2022), mitigating the overconfidence of the deep neural network by penalizing the gap between the CDF of the non-conformity scores and the uniform distribution; and (iii) ConfTr (Stutz et al., 2021), aiming to decrease prediction set size by inducing conformal loss as the average differentiable prediction set size. The details of CUT and ConfTr are in Appendix A.

Evaluation metrics. Our first evaluation metric is the marginal coverage (Marg-Cov), defined as MargCov =

$\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} \mathbb{1}[Y_i \in C_f(X_i)]$. Our second evaluation metric is the prediction set size (Avg-Set-Size), defined as Avg-Set-Size = $\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} |C_f(X_i)|$. To make the conformal alignment loss smooth, DPSM and ConfTr use the Sigmoid function to approximate the average prediction set size (APSS) during training, so we use the third evaluation metric as the Soft Set Size (Avg-Soft-Size), defined as Avg-Softset-Size = $\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} \sum_{y \in \mathcal{Y}} \tilde{\mathbb{1}}[S_f(X, y) \leq \hat{q}_f]$, recall that $\tilde{\mathbb{1}}[S \leq q] = 1/(1 + \exp(-(q - S)/\tau_{\text{Sigmoid}}))$ with a hyperparameter τ_{Sigmoid} .

5.2. Results and Discussion

We discuss the results comparing DPSM with baseline methods, fine-grained analysis for DPSM to demonstrate its effectiveness and validate our theoretical results.

DPSM generates smaller prediction sets. Table 1 presents the prediction set sizes for different methods using HPS score for training, HPS and APS scores for calibration and testing phases at $\alpha = 0.1$, as detailed in Appendix E. On average, DPSM reduces the prediction set size by 20.46% across the three datasets relative to the best baseline. Specifically, DPSM provides the best predictive efficiency compared to all other existing methods when calibrated with HPS. It also outperforms the best baselines when calibrated with APS in all cases except \uparrow 8.71% increase on CIFAR-100 with DenseNet and \uparrow 0.54% increase on iNaturalist with ResNet in terms of the prediction set size. The coverage results for overall comparison and additional results for calibrating with RAPS scoring function are in Appendix F.1. We also visualize the coverage rate and APSS with confidence intervals of all methods using DenseNet and HPS score in Figure 1. It clearly confirms that DPSM achieves significantly smaller prediction set size while maintaining the valid coverage.

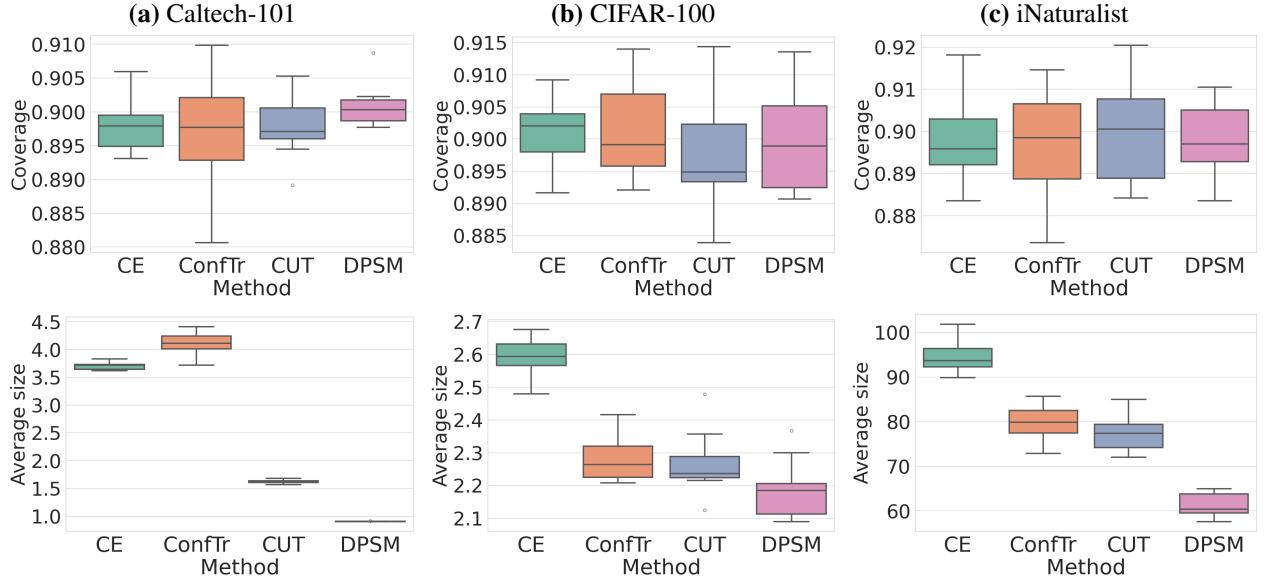


Figure 1. Box plots of coverage (Top row) and APSS (Bottom row) of all methods using DenseNet and HPS score. DPSM achieves significantly smaller prediction set size while maintaining the valid coverage.

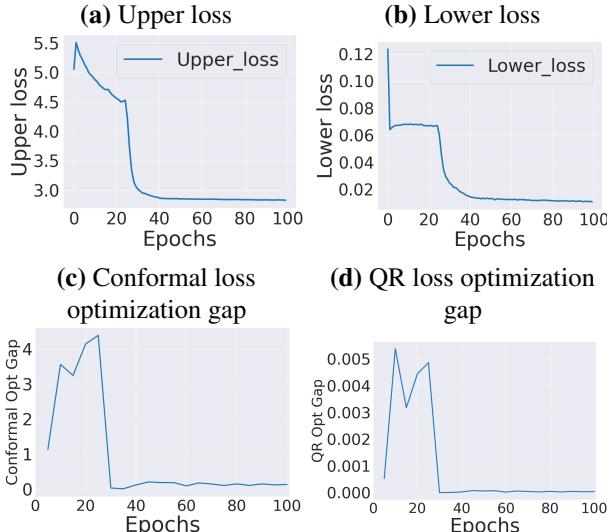


Figure 2. Justification experiments for the convergence of DPSM on CIFAR-100 using DenseNet and HPS score. (a) Upper level loss (i.e., a combination of classification loss and conformal alignment loss); (b) Lower level loss (i.e., QR loss); (c) Optimization gap of conformal loss, defined as the difference between conformal losses using learned batch-level quantiles and dataset-level quantiles on the training set; (d) Optimization gap of the lower-level QR loss, defined similarly as the loss difference between learned batch-level quantiles and dataset-level quantiles.

DPSM converges to stable error for bilevel optimization. To further explore how DPSM effectively generates smaller prediction sets, we analyze the convergence of DPSM by plotting the loss of the upper level function (i.e., a combination of classification loss and conformal alignment loss) and the lower level function (i.e., QR loss) over training epochs. Figure 2 (a) and (b) show the upper-level loss and lower-level loss over 100 epochs, respectively. We also re-

port the results of 40-epoch training regime in Appendix F.1 for reference. As shown, the upper-level loss of DPSM exhibits an initial increase during the first 5 epochs, reaching the peak, then steadily decreases before stabilizing around epoch 35. In contrast, the lower-level loss decreases sharply within the first 5 epochs, followed by a gradual reduction until convergence near the end of training. These results empirically demonstrate that DPSM effectively converges in terms of both upper and lower level training errors, validating its bilevel optimization approach. To investigate how the learned quantiles influence the optimization error, we compute both conformal and QR losses using the learned quantiles and the optimal (dataset-level) quantiles. The corresponding optimization errors—defined as the loss differences between learned and optimal quantiles—are visualized in Figure 2 (c) and (d). Both errors converge to nearly 0, indicating that the learned quantiles effectively approximate the optimal quantiles over training epochs.

DPSM estimates empirical quantiles with small error. To compare the precision of empirical quantiles estimation in ConfTr and DPSM, we plot the estimation error between \hat{Q}_f^n (quantiles evaluated on the whole training dataset) and \hat{q}_f (quantiles evaluated in ConfTr or learned in DPSM on mini-batches). Figure 3 (a) plots this estimation error over training epochs. For the first 25 epochs, the estimation errors for DPSM are significantly larger compared to ConfTr. However, as the training progresses, the estimation errors for DPSM decrease rapidly, converging close to 0 after epoch 35. This result verifies the theoretical result for smaller estimation error in learning bound analysis from Theorem 4.1. The rapid reduction in estimation error also reflects the convergence of the lower loss (i.e., QR loss), highlighting the effectiveness of DPSM in accurately estimating quantiles.

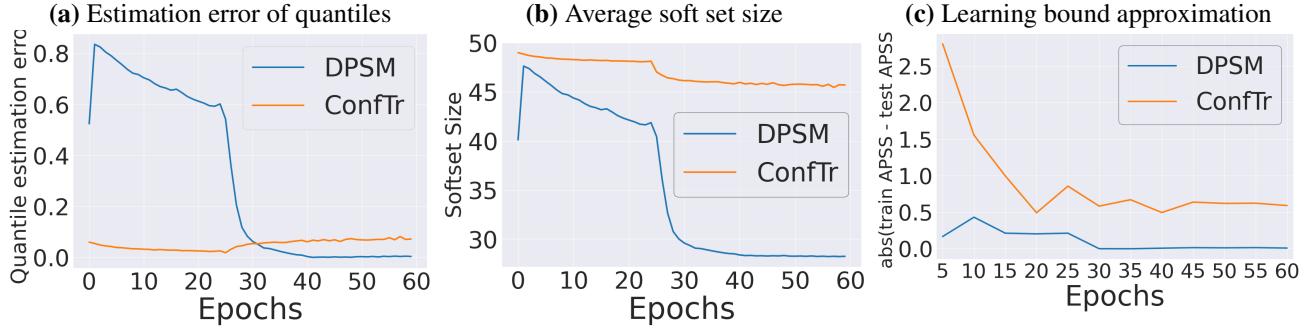


Figure 3. Justification experiments for the learning bound of DPSM on CIFAR-100 using DenseNet and HPS score. (a) Estimation error between the \hat{Q}_f^n (i.e., the dataset-level quantiles on training data) and \hat{q}_f (i.e., either the batch-level quantiles evaluated in ConfTr or the learned ones in DPSM); **(b)** Average soft set size (approximated using Sigmoid function) of DPSM and ConfTr; **(c)** Approximated learning error comparison between DPSM and ConfTr, measured by their gaps between the training and testing APSS.

Learning bound of DPSM is much tighter than ConfTr. To approximately compare the learning bounds of DPSM and ConfTr, we compare the conformal alignment losses of ConfTr and DPSM during training in terms of the average soft set size, as shown in Figure 3 (b). The soft set size of DPSM is consistently smaller than that of ConfTr during training. Combining the empirical results of the smaller estimation error of quantiles from Figure 3 (a), we can conclude that the learning bound of DPSM is much tighter than the learning bound of ConfTr, providing the empirical verification of Theorem 4.1 and 3.5. Although the learning bound cannot be empirically computed, we approximate it using a common strategy in ML literature (Yuan et al., 2019; Yang et al., 2021b), which estimates the generalization error by the absolute gap between training and test errors. For CP, we use APSS evaluated on the train and test sets to approximate the learning errors. Specifically, for DPSM, at each iteration, we: (i) compute APSS on the training set using the learned quantiles as thresholds. It includes optimization error since the learned quantiles are not optimal (true dataset-level quantiles); (ii) compute the APSS on the testing set using the dataset-level quantiles as thresholds. The gap between these two APSS values is employed as an approximation of the learning bound. We apply the same strategy to the SA-based ConfTr, where the training APSS is computed using the quantiles evaluated on mini-batches from the training data, and the test APSS is computed using the dataset-level quantiles from the test data. This comparison is shown in Figure 3 (c), which demonstrates that the approximated learning error is improved by DPSM.

Assumption 3.1 is empirically valid. Figure 4 (a) illustrates the conformity scores against their corresponding normalized order. The x-axis represents the normalized order, while the y-axis represents conformity scores. It is clear that the curve does not remain near the x-axis or y-axis, indicating that the gradient of conformity scores with respect to normalized index is both upper and lower bounded. This observation empirically validates Assumption 3.1.

Assumption 3.2 is empirically valid. Figure 4 (b) visualizes the soft set size of ConfTr, with input as coverage rate $\in [0.02, 0.98]$ with range 0.02. When coverage rate approaches 0.9, the curves of all methods exhibit a concave shape (zoom-in version also shown), providing empirical verification for Assumption 3.2.

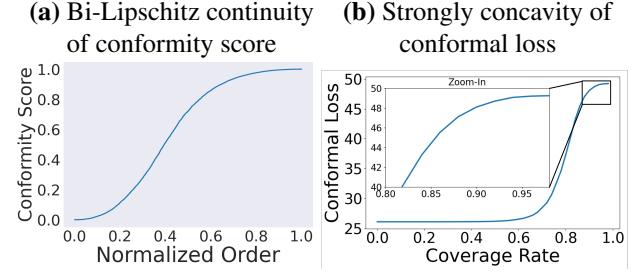


Figure 4. Assumption verification on CIFAR-100 using DenseNet and HPS score on the calibration dataset. **(a)** HPS scores over corresponding normalized index produced by ConfTr. The x-axis is the normalized order, the y-axis is the corresponding conformity score; **(b)** The soft set size measure of ConfTr. The input coverage rate is from $[0.02, 0.98]$ with 0.02 range with its zoom-in version where coverage rate is close to the target coverage 0.9; When coverage rate is close to 0.9, the curve for the soft set size exhibits a concave shape.

6. Conclusion

This paper developed Direct Prediction Set Minimization (DPSM), a novel conformal training algorithm formulated as a bilevel optimization problem. By leveraging quantile regression in the lower-level subproblem, DPSM precisely learns the empirical quantile of non-conformity scores, leading to smaller prediction sets. Theoretical analysis showed that DPSM attains a tight learning bound of $O(1/\sqrt{n})$ with n training samples. Empirical evaluation on real-world datasets confirmed that DPSM significantly outperforms existing conformal training methods, validating both its theoretical advantages and practical effectiveness.

Acknowledgements

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

Impact Statement

This paper focuses on integrating conformal principles into the training procedure of deep neural network classifiers to produce small prediction sets. The insights in this paper could improve the safe deployment of deep classifiers in may high-risk applications. There is not any negative consequences to be specifically highlighted here.

References

- Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.
- Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.
- Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.
- Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pp. 717–730. PMLR, 2022.
- Angelopoulos, A. N., Candes, E., and Tibshirani, R. Conformal pid control for time series prediction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.
- Babbar, V., Bhatt, U., and Weller, A. On the utility of prediction sets in human-ai teams. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022.
- Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.
- Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.
- Beck, Y., Ljubić, I., and Schmidt, M. A survey on bilevel optimization under uncertainty. *European Journal of Operational Research*, 311(2):401–426, 2023.
- Begoli, E., Bhattacharya, T., and Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.
- Bellotti, A. Optimized conformal classification using gradient descent approximation. *arXiv preprint arXiv:2105.11255*, 2021.
- Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 2337–2363. PMLR, 2023.
- Bolte, J., Nguyen, T. P., Peyrouquet, J., and Suter, B. W. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2017.
- Bolte, J., Le, T., Pauwels, E., and Silveti-Falls, T. Nonsmooth implicit differentiation for machine-learning and optimization. *Advances in neural information processing systems*, 34:13537–13549, 2021.
- Bolte, J., Pauwels, E., and Vaiter, S. Automatic differentiation of nonsmooth iterative algorithms. *Advances in Neural Information Processing Systems*, 35:26404–26417, 2022.
- Borsos, Z., Mutny, M., and Krause, A. Coresets via bilevel optimization for continual learning and streaming. *Advances in neural information processing systems*, 33:14879–14890, 2020.
- Borsos, Z., Mutny, M., Tagliasacchi, M., and Krause, A. Data summarization via bilevel optimization. *Journal of Machine Learning Research*, 25(73):1–53, 2024.
- Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22(81):1–42, 2021.
- Chen, L., Xu, J., and Zhang, J. On finding small hypergradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 947–980. PMLR, 2024a.

- Chen, P., Shi, X., Jiang, R., and Wang, J. Penalty-based methods for simple bilevel optimization under hölderian error bounds. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.
- Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.
- Chen, X., Xiao, T., and Balasubramanian, K. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *Journal of Machine Learning Research*, 25(151):1–51, 2024c.
- Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.
- Cresswell, J. C., Sui, Y., Kumar, B., and Vouitsis, N. Conformal prediction sets improve human decision making. In *Forty-first International Conference on Machine Learning*, 2024.
- Dagréou, M., Ablin, P., Vaïter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.
- Deshpande, S., Marx, C., and Kuleshov, V. Online calibrated and conformal prediction improves bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1450–1458. PMLR, 2024.
- Ding, T., Angelopoulos, A., Bates, S., Jordan, M., and Tibshirani, R. J. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024.
- Domke, J. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pp. 318–326. PMLR, 2012.
- Einbinder, B.-S., Romano, Y., Sesia, M., and Zhou, Y. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in Neural Information Processing Systems*, 35:22380–22395, 2022.
- Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004.
- Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Efficient conformal prediction via cascaded inference with expanded admission. In *Proceedings of The Tenth International Conference on Learning Representations*, 2021a.
- Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, pp. 3329–3339. PMLR, 2021b.
- Ghadimi, S. and Wang, M. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.
- Ghosh, S., Belkhoudja, T., Yan, Y., and Doppa, J. R. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: Novel algorithm and theoretical analysis. In *AAAI*, pp. 7722–7730, 2023a.
- Ghosh, S., Shi, Y., Belkhoudja, T., Yan, Y., Doppa, J., and Jones, B. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, pp. 681–690. PMLR, 2023b.
- Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.
- Gibbs, I. and Candès, E. J. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.
- Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.
- Gong, X., Hao, J., and Liu, M. An accelerated algorithm for stochastic bilevel optimization under unbounded smoothness. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.
- Gong, X., Hao, J., and Liu, M. A nearly optimal single loop algorithm for stochastic bilevel optimization under unbounded smoothness. In *ICML*, 2024b.
- Guan, L. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.
- Hao, J., Ji, K., and Liu, M. Bilevel coresets selection in continual learning: A new formulation and algorithm. *Advances in Neural Information Processing Systems*, 36, 2024.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

- He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023. doi: 10.1137/20M1387341.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- Huang, J., Xi, H., Zhang, L., Yao, H., Qiu, Y., and Wei, H. Conformal prediction for deep classifier via label ranking. *arXiv preprint arXiv:2310.06430*, 2023a.
- Huang, M., Zhang, D., and Ji, K. Achieving linear speedup in non-iid federated bilevel learning. In *International Conference on Machine Learning*, pp. 14039–14059. PMLR, 2023b.
- Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.
- Kawaguchi, K. and Lu, H. Ordered sgd: A new stochastic optimization framework for empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 669–679. PMLR, 2020.
- Khanduri, P., Zeng, S., Hong, M., Wai, H. T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.
- Kiyani, S., Pappas, G. J., and Hassani, H. Conformal prediction with learned features. In *Forty-first International Conference on Machine Learning*, 2024a.
- Kiyani, S., Pappas, G. J., and Hassani, H. Length optimization in conformal prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.
- Koenker, R. *Quantile regression*, volume 38. Cambridge university press, 2005.
- Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.
- Kumar, A., Irsay, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pp. 1378–1387. PMLR, 2016.
- Kwon, J., Kwon, D., Wright, S., and Nowak, R. D. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pp. 18083–18113. PMLR, 2023.
- Kwon, J., Kwon, D., Wright, S., and Nowak, R. D. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *The Twelfth International Conference on Learning Representations*, 2024.
- Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.
- Lei, J., Robins, J., and Wasserman, L. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.
- Liu, B., Ye, M., Wright, S., Stone, P., and qiang liu. BOME! bilevel optimization made easy: A simple first-order approach. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.
- Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International conference on machine learning*, pp. 6305–6315. PMLR, 2020.
- Liu, R., Gao, J., Zhang, J., Meng, D., and Lin, Z. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2021.
- Liu, Z., Cui, Y., Yan, Y., Xu, Y., Ji, X., Liu, X., and Chan, A. B. The pitfalls and promise of conformal inference under adversarial attacks. In *International Conference on Machine Learning*, pp. 30908–30928. PMLR, 2024.
- Lu, C., Angelopoulos, A. N., and Pomerantz, S. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 545–554. Springer, 2022a.
- Lu, C., Lemay, A., Chang, K., Höbel, K., and Kalpathy-Cramer, J. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference*

- on Artificial Intelligence*, volume 36, pp. 12008–12016, 2022b.
- Lu, Z. and Mei, S. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.
- Masters, D. and Luschi, C. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.
- Narayan, T., Wang, S. L., Canini, K. R., and Gupta, M. Expected pinball loss for quantile regression and inverse cdf estimation. *Transactions on Machine Learning Research*, 2024.
- Oliveira, R. I., Orenstein, P., Ramos, T., and Romano, J. V. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38, 2024.
- Pang, J.-S. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, 1997.
- Pedregosa, F. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.
- Podkopaev, A., Xu, D., and chih Lee, K. Adaptive conformal inference by betting. In *Forty-first International Conference on Machine Learning*, 2024.
- Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.
- Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.
- Roulet, V. and d’Aspremont, A. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.
- Sabach, S. and Shtern, S. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017. doi: 10.1137/16M105592X.
- Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.
- Sesia, M. and Romano, Y. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.
- Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.
- Shahrokhi, H., Roy, D. R., Yan, Y., Arnaoudova, V., and Doppa, J. R. Conformal prediction sets for deep generative models via reduction to conformal regression. *CoRR*, abs/2503.10512, 2025. doi: 10.48550/ARXIV.2503.10512. URL <https://doi.org/10.48550/arXiv.2503.10512>.
- Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- Shen, H. and Chen, T. On penalty-based bilevel gradient descent method. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 30992–31015, 2023.
- Shen, H. and Chen, T. On penalty-based bilevel gradient descent method. *Mathematical Programming (MAPR)*, to appear, 2025.
- Shen, H., Yang, Z., and Chen, T. Principled penalty-based methods for bilevel reinforcement learning and rlhf. In *Forty-first International Conference on Machine Learning*.
- Shi, Y., GHOSH, S., Belkhouja, T., Doppa, J., and Yan, Y. Conformal prediction for class-wise coverage via augmented label rank calibration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Straitouri, E., Wang, L., Okati, N., and Rodriguez, M. G. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning (ICML)*, 2023.
- Stutz, D., Cemgil, A. T., Doucet, A., et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.
- Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Vovk, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, 2015.
- Vovk, V., Gammerman, A., and Saunders, C. Machine-learning applications of algorithmic randomness. 1999.

- Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.
- Vovk, V., Fedorova, V., Nouretdinov, I., and Gammerman, A. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*, pp. 23–39. Springer, 2016.
- Yan, G., Romano, Y., and Weng, T.-W. Provably robust conformal prediction with improved efficiency. In *The Twelfth International Conference on Learning Representations*, 2024.
- Yang, J., Ji, K., and Liang, Y. Provably faster algorithms for bilevel optimization. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021a.
- Yang, S. and Fevens, T. Uncertainty quantification and estimation in medical image classification. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part III 30*, pp. 671–683. Springer, 2021.
- Yang, T. and Lin, Q. Rsg: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(6):1–33, 2018.
- Yang, Y., Xiao, P., and Ji, K. SimFBO: Towards simple, flexible and communication-efficient federated bilevel learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Yang, Z., Bai, Y., and Mei, S. Exact gap between generalization error and uniform convergence in random feature models. In *International Conference on Machine Learning*, pp. 11704–11715. PMLR, 2021b.
- Yuan, Z., Yan, Y., Jin, R., and Yang, T. Stagewise training accelerates convergence of testing error over sgd. *Advances in Neural Information Processing Systems*, 32, 2019.
- Zhang, Y., Khanduri, P., Tsaknakis, I., Yao, Y., Hong, M., and Liu, S. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1):38–59, 2024.
- Zhou, W. Bi-level offline policy optimization with limited exploration. *Advances in Neural Information Processing Systems*, 36, 2024.
- Zhou, X., Pi, R., Zhang, W., Lin, Y., Chen, Z., and Zhang, T. Probabilistic bilevel coreset selection. In *International Conference on Machine Learning*, pp. 27287–27302. PMLR, 2022.
- Zhu, Y., Hernández, D., He, Y., Ding, Z., Xiong, B., Kharlamov, E., and Staab, S. Predicate-conditional conformalized answer sets for knowledge graph embeddings. *arXiv preprint arXiv:2505.16877*, 2025.

A. Additional background details

Non-conformity scoring functions. The homogeneous prediction sets (HPS) (Sadinle et al., 2019) scoring function is defined as follows:

$$S_f^{\text{HPS}}(X, Y) = 1 - f_\theta(X)_Y. \quad (11)$$

(Romano et al., 2020) has proposed another conformity scoring function, Adaptive Prediction Sets (APS). APS scoring function is based on sorted probabilities. For a given input X , we sort the softmax probabilities for all classes $\{1, \dots, K\}$ such that $1 \geq f_\theta(x)_{(1)} \geq \dots \geq f_\theta(x)_{(K)} \geq 0$, and compute the cumulative confidence as follows:

$$S_f^{\text{APS}}(X, Y) = \sum_{l=1}^{r_f(X, Y)-1} f_\theta(X)_{(l)} + U \cdot f_\theta(X)_{(r_f(X, Y))}, \quad (12)$$

where $U \in [0, 1]$ is a uniform random variable to break ties.

To reduce the probability of including unnecessary labels (i.e., labels with high ranks) and thus improve the predictive efficiency, (Angelopoulos et al., 2021) has proposed Regularized Adaptive Prediction Sets (RAPS) scoring function. The RAPS score is computed as follows:

$$S_f^{\text{RAPS}}(X, Y) = \sum_{l=1}^{r_f(X, Y)-1} f_\theta(X)_{(l)} + U \cdot f_\theta(X)_{(r_f(X, Y))} + \lambda_{\text{RAPS}}(r_f(X, Y) - k_{\text{reg}})^+, \quad (13)$$

where λ_{RAPS} and k_{reg} are two hyper-parameters.

Objective function for conformal training methods. ConfTr (Stutz et al., 2021) estimates a soft measure of the prediction set size, defined as follows:

$$\widehat{\mathcal{L}}_{c, \text{CTr}}^{\text{SA}}(f) = \mathbb{E}_{\widehat{q}_f \sim \mathcal{Q}_f} \left[\frac{1}{n} \sum_{i=1}^n \sum_{y \in \mathcal{Y}} \tilde{\mathbb{1}}[S_f(X_i, y) \leq \widehat{q}_f] \right], \quad (14)$$

where $\tilde{\mathbb{1}}[\cdot]$ is a smoothed estimator for the indicator function $\mathbb{1}[\cdot]$ and defined by a Sigmoid function (Stutz et al., 2021), i.e., $\tilde{\mathbb{1}}[S \leq q] = 1/(1 + \exp(-(q - S)/\tau_{\text{Sigmoid}}))$ with a tunable temperature parameter τ_{Sigmoid} .

CUT (Einbinder et al., 2022) measures the maximum deviation from the uniformity of conformity scores, defined as follows:

$$\widehat{\mathcal{L}}_{c, \text{CUT}}^{\text{SA}}(f) = \mathbb{E}_{\widehat{q}_f \sim \mathcal{Q}_f} \left[\sup_{\alpha \in [0, 1]} |(1 - \alpha) - \widehat{q}_f(\alpha)| \right], \quad (15)$$

where $\widehat{q}_f(\alpha)$ is the empirical batch-level quantile in \mathcal{B} of input $\alpha \in [0, 1]$.

B. Proof for Section 3

In this section, we prove Proposition 3.4 and Theorem 3.5 from Section 3.

B.1. Proof for Proposition 3.4

Proposition B.1. (*Proposition 3.4 restated, distribution of mini-batch quantiles*) Define an event $Z(j)$ as “the j -th smallest score $S_{(j)}$ from $\{S_i\}_{i=1}^n$ is selected as $\widehat{q}_f(\alpha)$ on a mini-batch \mathcal{B} ”. Then the probability of $Z(j)$ is:

$$\mathbb{P}(Z(j)) = \frac{\binom{j}{\lceil(1-\alpha)(s+1)\rceil} \binom{n-j-1}{s-\lceil(1-\alpha)(s+1)\rceil-1}}{\binom{n}{s}}.$$

Furthermore, we have the following asymptotic result

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z(j)) = \frac{e}{n} \mathbb{P}_{\text{Beta}}\left(\frac{j}{n}; \lceil(1-\alpha)(s+1)\rceil + 1; s - \lceil(1-\alpha)(s+1)\rceil\right),$$

where $\mathbb{P}_{\text{Beta}}(x; a, b)$ is the PDF of Beta distribution with two shape parameters a, b .

Proof. (Proof of Proposition 3.4) This proof follows the proof for Proposition 1 in (Kawaguchi & Lu, 2020).

Recall the Stirling's approximation that is used to prove this proposition:

$$n! \sim \sqrt{2\pi n} (n/e)^n.$$

To simplify the representation, we denote $a = \lceil(1 - \alpha)(s + 1)\rceil$ in this proof. Then, the $\mathbb{P}(Z(j))$ could be rewritten as:

$$\mathbb{P}(Z(j)) = \frac{\binom{j}{a} \binom{n-j-1}{s-a-1}}{\binom{n}{s}} = \frac{\frac{j!}{a!(j-a)!} \cdot \frac{(n-j-1)!}{(s-a-1)!(n-j-s+a)!}}{\frac{n!}{s!(n-s)!}} = \frac{s!}{a!(s-a-1)!} \cdot \frac{j!}{(j-a)!} \cdot \frac{(n-j-1)!}{(n-j-s+a)!} \cdot \frac{(n-s)!}{n!}. \quad (16)$$

Applying the Stirling's approximation when $n \rightarrow \infty$, we have that:

$$\begin{aligned} & \frac{j!}{(j-a)!} \cdot \frac{(n-j-1)!}{(n-j-s+a)!} \cdot \frac{(n-s)!}{n!} \\ &= \frac{\sqrt{2\pi j} \left(\frac{j}{e}\right)^j}{\sqrt{2\pi(j-a)} \left(\frac{j-a}{e}\right)^{j-a}} \cdot \frac{\sqrt{2\pi(n-j-1)} \left(\frac{n-j-1}{e}\right)^{n-j-1}}{\sqrt{2\pi(n-j-s+a)} \left(\frac{n-j-s+a}{e}\right)^{n-j-s+a}} \cdot \frac{\sqrt{2\pi(n-s)} \left(\frac{n-s}{e}\right)^{n-s}}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n} \\ &= \frac{j^j}{(j-a)^{j-a}} e^{-a} \cdot \frac{(n-j-1)^{n-j-1}}{(n-j-s+a)^{n-j-s+a}} e^{-s+a+1} \cdot \frac{(n-s)^{n-s}}{(n)^n} e^s \\ &= e \frac{j^j}{(j-a)^{j-a}} \cdot \frac{(n-j)^{n-j-1}}{(n-j)^{n-j-s+a}} \cdot \frac{(n)^{n-s}}{(n)^n} \\ &= e \frac{(n)^j \left(\frac{j}{n}\right)^j}{(n)^{j-a} \left(\frac{j-a}{n}\right)^{j-a}} \cdot (n-j)^{s-a-1} \cdot (n)^{-s} \\ &= e \frac{\left(\frac{j}{n}\right)^j}{\left(\frac{j-a}{n}\right)^{j-a}} (n)^a \cdot \left(\frac{n-j}{n}\right)^{s-a-1} (n)^{s-a-1} \cdot (n)^{-s} \\ &= e \cdot \left(\frac{j}{n}\right)^a \cdot \left(1 - \frac{j}{n}\right)^{s-a-1} \end{aligned} \quad (17)$$

Combining the Equation (16) and (17), we have that:

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(Z(j)) &= \frac{e}{n} \cdot \frac{s!}{a!(s-a-1)!} \cdot \left(\frac{j}{n}\right)^a \cdot \left(1 - \frac{j}{n}\right)^{s-a-1} = \frac{e}{n} \frac{\Gamma(a+1+s-a)}{\Gamma(a+1)\Gamma(s-a)} \cdot \left(\frac{j}{n}\right)^a \cdot \left(1 - \frac{j}{n}\right)^{s-a-1} \\ &= \frac{e}{n} \mathbb{P}_{\text{Beta}}\left(\frac{j}{n}; \lceil(1-\alpha)(s+1)\rceil + 1; s - \lceil(1-\alpha)(s+1)\rceil\right), \end{aligned}$$

where $\Gamma(n) = (n-1)!$ is gamma function and $\mathbb{P}_{\text{Beta}}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}$ is the probability density function of Beta distribution with shape parameters a and b at x . \square

B.2. Proof for Theorem 3.5

Theorem B.2. (*Theorem 3.5 restated, learning bounds of SA method*) Suppose that Assumption 3.1 and 3.2 hold. Assume $(1 - \alpha)(s + 1)$ is not an integer. If $\lceil(1 - \alpha)(s + 1)\rceil - (1 - \alpha)(s + 1) \geq \Omega(1/s)$ and $s \leq \sqrt{n}$, then the following inequality holds with probability at least $1 - \delta$:

$$\Omega(1/s) \leq |\widehat{\mathcal{L}}_c^{\text{SA}}(f) - \mathcal{L}_c(f)| \leq \tilde{O}(1/\sqrt{s}).$$

Proof. (Proof of Theorem 3.5)

We need the following helpful technical lemmas:

Lemma B.3. (*Lower bound for $\widehat{\ell}(f, \mathbb{E}_{\widehat{q}_f}[\widehat{q}_f]) - \mathbb{E}_{\widehat{q}_f}[\widehat{\ell}(f, \widehat{q}_f)]$*) Suppose that Assumptions 3.1 and 3.2 hold. Then, the following inequality holds:

$$|\widehat{\ell}(f, \mathbb{E}_{\widehat{q}_f}[\widehat{q}_f]) - \mathbb{E}_{\widehat{q}_f}[\widehat{\ell}(f, \widehat{q}_f)]| \geq \Omega(1/s).$$

Lemma B.4. (*Lower bound for $\widehat{\ell}(f, \mathbb{E}[\widehat{q}_f]) - \widehat{\ell}(f, Q_f)$*) Suppose that Assumption 3.1 holds. Assume $(1 - \alpha)(s + 1)$ is not an integer. If $\lceil(1 - \alpha)(s + 1)\rceil - (1 - \alpha)(s + 1) \geq \Omega(1/s)$ and $s \leq \sqrt{n}$, then the following inequality holds:

$$\widehat{\ell}(f, \mathbb{E}[\widehat{q}_f]) - \widehat{\ell}(f, Q_f) \geq \Omega(1/s).$$

Lemma B.5. (*Distance between $\mathcal{L}_c(f, \widehat{q}_f)$ and $\mathbb{E}_{\widehat{q}_f}[\mathcal{L}_c(f; \widehat{q}_f)]$*) Suppose that Assumption 3.1 holds. With probability at least $1 - \delta$, the following inequality holds:

$$|\mathcal{L}_c(f, \widehat{q}_f) - \mathbb{E}_{\widehat{q}_f}[\mathcal{L}_c(f; \widehat{q}_f)]| \leq \tilde{O}(1/\sqrt{s}).$$

The proofs of Lemma B.3, B.4, and B.5 are deferred to the end of this proof.

Recall that

$$\begin{aligned} \widehat{\mathcal{L}}_c^{\text{SA}}(f) &= \mathbb{E}_{\widehat{q}_f \sim \mathcal{Q}_f} [\widehat{\ell}(f, \widehat{q}_f)] = \mathbb{E}_{\widehat{q}_f \sim \mathcal{Q}_f} \left[\underbrace{\frac{1}{n} \sum_{i=1}^n \sum_{y \in \mathcal{Y}} \mathbb{1}[S_f(X_i, y) \leq \widehat{q}_f]}_{= \widehat{\ell}(f, \widehat{q}_f)} \right], \\ \mathcal{L}_c(f) &= \ell(f, Q_f) = \mathbb{E}_X \left[\sum_{y \in \mathcal{Y}} \mathbb{1}[S_f(X, y) \leq Q_f] \right]. \end{aligned}$$

We start with the lower bound in Theorem 3.5.

$$\begin{aligned} (\widehat{\mathcal{L}}_c^{\text{SA}}(f) - \mathcal{L}_c(f))^2 &= (\mathbb{E}_{\widehat{q} \sim \mathcal{Q}_f} [\widehat{\ell}(f, \widehat{q})] - \ell(f, Q_f))^2 = (\mathbb{E}_{\widehat{q} \sim \mathcal{Q}_f} [\widehat{\ell}(f, \widehat{q})] - \widehat{\ell}(f, \mathbb{E}[\widehat{q}_f]) + \widehat{\ell}(f, \mathbb{E}[\widehat{q}_f]) - \ell(f, Q_f))^2 \\ &= (\underbrace{\mathbb{E}_{\widehat{q} \sim \mathcal{Q}_f} [\widehat{\ell}(f, \widehat{q})] - \widehat{\ell}(f, \mathbb{E}[\widehat{q}_f])}_{\geq \Omega(1/s^2), \text{ Lemma B.3}})^2 + (\underbrace{\widehat{\ell}(f, \mathbb{E}[\widehat{q}_f]) - \ell(f, Q_f)}_{\geq 0})^2 \\ &\quad + (\underbrace{(\mathbb{E}_{\widehat{q} \sim \mathcal{Q}_f} [\widehat{\ell}(f, \widehat{q})] - \widehat{\ell}(f, \mathbb{E}[\widehat{q}_f]))(\widehat{\ell}(f, \mathbb{E}[\widehat{q}_f]) - \widehat{\ell}(f, Q_f))}_{\geq \Omega(1/s), \text{ Lemma B.3}}) + \underbrace{\widehat{\ell}(f, Q_f) - \ell(f, Q_f)}_{\geq -\tilde{O}(1/\sqrt{n}), \text{ Lemma C.4}} \\ &\geq \Omega(1/s^2), \end{aligned}$$

where the last inequality is due to $s \ll \sqrt{n}$.

The above inequality thus indicates the result:

$$|\widehat{\mathcal{L}}_c^{\text{SA}}(f) - \mathcal{L}_c(f)| \geq \Omega(1/s).$$

Next, we begin to prove the upper bound in Theorem 3.5:

$$\begin{aligned} &|\mathcal{L}_c(f, Q_f) - \mathbb{E}_{\widehat{q}_f}[\mathcal{L}_c(f; \widehat{q}_f)]| \\ &\leq |\mathcal{L}_c(f, Q_f) - \mathcal{L}_c(f; \widehat{q}_f)| + |\mathcal{L}_c(f; \widehat{q}_f) - \mathbb{E}_{\widehat{q}_f}[\mathcal{L}_c(f; \widehat{q}_f)]| \\ &\leq O(1/\sqrt{s}), \text{ according to Lemma B.5} \\ &\leq L|Q_f - \widehat{q}_f| + O(1/\sqrt{s}) \\ &= L|S_{f, (\lceil(1-\alpha)(s+1)\rceil)} - S_{f, (\lceil(1-\alpha-O(1/\sqrt{s}))(s+1)\rceil)}| + O(1/\sqrt{s}) \\ &\leq L \cdot L_2 \left| \frac{\lceil(1-\alpha)(s+1)\rceil}{s} - \frac{\lceil(1-\alpha-O(1/\sqrt{s}))(s+1)\rceil}{s} \right| + O(1/\sqrt{s}) \\ &\leq L \cdot L_2 \cdot O(1/\sqrt{s}) + O(1/\sqrt{s}) = O(1/\sqrt{s}) + O(1/\sqrt{s}) = O(1/\sqrt{s}), \end{aligned}$$

where the first inequality is due to the triangle inequality, the second inequality is due to the L -Lipschitz continuous of $\mathcal{L}_c(f, Q_f)$ in Lemma C.3 and Lemma B.5, and the third inequality is due to (24) and Assumption 3.1.

Combining the lower and upper learning bound, the proof of Theorem 3.5 is finished. \square

B.3. Proof for Lemma B.3

Proof. (of Lemma B.3)

We begin to prove the lower bound of $\widehat{\ell}(f, \mathbb{E}_{\widehat{q}_f}[\widehat{q}_f]) - \mathbb{E}_{\widehat{q}_f}[\widehat{\ell}(f; \widehat{q}_f)]$.

By the μ -strong-concavity of $\widehat{\ell}(f, q)$ in q (Assumption 3.2), we have

$$\begin{aligned}\widehat{\ell}(f, \mathbb{E}_{\widehat{q}_f}[\widehat{q}_f]) - \mathbb{E}_{\widehat{q}_f}[\widehat{\ell}(f; \widehat{q}_f)] &= \mathbb{E}_{\widehat{q}_f}[\widehat{\ell}(f, \mathbb{E}_{\widehat{q}_f}[\widehat{q}_f]) - \widehat{\ell}(f; \widehat{q}_f)] \geq \mathbb{E}_{\widehat{q}_f}[\partial_q \widehat{\ell}(f, \mathbb{E}_{\widehat{q}_f}[\widehat{q}_f])(\widehat{q}_f - \mathbb{E}[\widehat{q}_f]) + \frac{\mu}{2}(\widehat{q}_f - \mathbb{E}[\widehat{q}_f])^2] \\ &= \frac{\mu}{2} \mathbb{E}_{\widehat{q}_f}[(\widehat{q}_f - \mathbb{E}[\widehat{q}_f])^2] \geq L_1^2 \cdot \frac{\mu}{2} \frac{\mathbb{E}_j[(\lceil(1-\alpha)(n+1)\rceil - j)^2]}{n^2} \geq L_1^2 \cdot \frac{\mu}{2} \frac{\mathbb{E}[(\mathbb{E}[j] - j)^2]}{n^2} \\ &= L_1^2 \cdot \frac{\mu}{2} \cdot \frac{(\lceil(1-\alpha)(s+1)\rceil + 1)(s - \lceil(1-\alpha)(s+1)\rceil)}{(s+1)^2(s+2)} \geq L_1^2 \cdot \frac{\mu}{2} \cdot \Omega(1/s) = \Omega(1/s),\end{aligned}$$

where the second inequality is due to Proposition 3.4 and Bi-Lipschitz continuity of score $S_{(j)}$ in j (Assumption 3.1), the third inequality is due to the property of variance, i.e., $\mathbb{V}(X) = \mathbb{E}[(\mathbb{E}[X] - X)^2] \leq \mathbb{E}[(a - X)^2]$ for any a , and the third equality is due to the variance of Beta distribution (j satisfies Beta distribution, as Proposition 3.4) is $\frac{ab}{(a+b)^2(a+b+a)}$ with the shape parameter a, b . \square

B.4. Proof for Lemma B.4

Proof. (of Lemma B.4)

To prove Lemma B.4, we need the following technical lemma (proof is deferred to Section B.6).

Lemma B.6. Assume $(1-\alpha)(s+1)$ is not an integer. If $\lceil(1-\alpha)(s+1)\rceil - (1-\alpha)(s+1) \geq \Omega(1/s)$ and $s \leq \sqrt{n}$, then $\frac{\lceil(1-\alpha)(s+1)\rceil}{s+1} - \frac{\lceil(1-\alpha)(n+1)\rceil}{n} \geq \Omega(1/s)$.

By using the μ -strong-concavity of $\ell(f, q)$ in q , we have

$$\begin{aligned}\ell(f, \mathbb{E}[\widehat{q}_f]) - \widehat{\ell}(f, Q_f) &\geq -\partial_q \ell(f, \mathbb{E}[\widehat{q}_f])(Q_f - \mathbb{E}[\widehat{q}_f]) + \frac{\mu}{2}(Q_f - \mathbb{E}[\widehat{q}_f])^2 \\ &= \underbrace{\partial_q \ell(f, \mathbb{E}[\widehat{q}_f])}_{>0 \text{ due to no tie}} (\mathbb{E}[\widehat{q}_f] - Q_f) + \frac{\mu}{2} \underbrace{(Q_f - \mathbb{E}[\widehat{q}_f])^2}_{\geq 0} \geq B \cdot (\mathbb{E}[\widehat{q}_f] - Q_f),\end{aligned}$$

where the second inequality is due to no tie in the distribution of non-conformity scores and there exists a value $B > 0$ as in (25) such that $\partial_q \ell(f, q) \geq B = 0$. It suffices to show the lower bound of $\mathbb{E}[\widehat{q}_f] - Q_f$.

Due to j satisfies the Beta distribution with the two shape parameters $a = \lceil(1-\alpha)(s+1)\rceil + 1$ and $b = s - \lceil(1-\alpha)(s+1)\rceil$ (Proposition 3.4) and the mean of Beta random variable is $a/(a+1)$, we know that the corresponding probability for $\mathbb{E}[\widehat{q}_f]$ is $\frac{\lceil(1-\alpha)(s+1)\rceil}{s+1}$ and it is larger than $1-\alpha$ if $(1-\alpha)(s+1)$ is not an integer.

Recall that \widehat{Q}_f is the empirical quantile on a set of data. We regard \widehat{Q}_f as the $(\frac{\lceil(1-\alpha)(n+1)\rceil}{n})$ -quantile on \mathcal{D}_{tr} . Since $Q_f(\alpha)$ is increasing as α decreases, we use the Bi-Lipschitz continuity of $S_{(j)}$ (Assumption 3.1):

$$\begin{aligned}\mathbb{E}[\widehat{q}_f] - Q_f &= \mathbb{E}[\widehat{q}_f] - \widehat{Q}_f + \widehat{Q}_f - Q_f \geq L_1 \left(\frac{\lceil(1-\alpha)(s+1)\rceil}{s+1} - \frac{\lceil(1-\alpha)(n+1)\rceil}{n} \right) + \left(\widehat{Q}_f - \widehat{Q}_f(\alpha - \tilde{O}(1/\sqrt{n})) \right) \\ &\geq L_1 \Omega(1/s) - L_1 \tilde{O}(1/\sqrt{n}) \geq \Omega(1/s),\end{aligned}$$

where the two inequalities are due to Lemma B.6, Bi-Lipschitz continuity of $S_{(j)}$ (Assumption 3.1) and the concentration inequality for the empirical and population quantiles in (24). \square

B.5. Proof for Lemma B.5

Proof. (of Lemma B.5)

According to Chebyshev's inequality, we have that:

$$\begin{aligned} \mathbb{P}\{|\mathcal{L}_c(f, \hat{q}_f) - \mathbb{E}_{\hat{q}_f}[\mathcal{L}_c(f, \hat{q}_f)]| \leq \epsilon\} &\geq 1 - \frac{\mathbb{E}_{\hat{q}_f}[(\mathcal{L}_c(f, \hat{q}_f) - \mathbb{E}_{\hat{q}_f}[\mathcal{L}_c(f, \hat{q}_f)])^2]}{\epsilon^2} \\ &\geq 1 - \frac{L^2 \mathbb{E}_{\hat{q}_f}[(\hat{q}_f - \mathbb{E}_{\hat{q}_f}[\hat{q}_f])^2]}{\epsilon^2} \geq 1 - \frac{L^2 L_2^2}{\epsilon^2} \cdot \frac{((1-\alpha)(s+1)] + 1)(s - \lceil(1-\alpha)(s+1)\rceil)}{(s+1)^2(s+2)} \geq 1 - \delta, \end{aligned}$$

where the first and last inequality hold due to the Chebyshev's inequality, the second inequality is due to the L -Lipschitz continuous of $\mathcal{L}_c(f, Q_f)$ in Lemma C.3.

Rearranging the above inequality, the following inequality holds with probability at least $1 - \delta$:

$$|\mathcal{L}_c(f, \hat{q}_f) - \mathbb{E}_{\hat{q}_f}[\mathcal{L}_c(f, \hat{q}_f)]| \leq L \cdot L_2 \cdot \sqrt{\frac{((1-\alpha)(s+1)] + 1)(s - \lceil(1-\alpha)(s+1)\rceil)}{\delta \cdot (s+1)^2(s+2)}} \leq L \cdot L_2 \cdot O(1/\sqrt{s}) = O(1/\sqrt{s}),$$

where the first inequality is due to Assumption 3.1 and Lemma C.3. \square

B.6. Proof for Lemma B.6

Proof. (of Lemma B.6)

Recall that $(1-\alpha)(s+1)$ is not an integer. Then we begin to compare the upper and lower bound of $\frac{\lceil(1-\alpha)(n+1)\rceil}{n}$ and $\frac{\lceil(1-\alpha)(s+1)\rceil}{s+1}$, respectively:

$$(1-\alpha) + \frac{1-\alpha}{n} \leq \frac{\lceil(1-\alpha)(n+1)\rceil}{n} \leq (1-\alpha) + \frac{2-\alpha}{n}, \quad (18)$$

$$(1-\alpha) < \frac{\lceil(1-\alpha)(s+1)\rceil}{s+1} \leq (1-\alpha) + \frac{1}{s+1}, \quad (19)$$

where all inequalities are due to $(1-\alpha)(n+1) \leq \lceil(1-\alpha)(n+1)\rceil \leq (1-\alpha)(n+1) + 1$.

Due to $s < n$ and $\alpha \in [0, 1]$, we have that:

$$\begin{aligned} \frac{\lceil(1-\alpha)(s+1)\rceil}{s+1} - \frac{\lceil(1-\alpha)(n+1)\rceil}{n} &\leq \max\left\{(1-\alpha) + \frac{2-\alpha}{n} - (1-\alpha), (1-\alpha) + \frac{1}{s+1} - (1-\alpha) - \frac{1-\alpha}{n}\right\} \\ &= \max\left\{\frac{2-\alpha}{n}, \frac{1}{s+1} - \frac{1-\alpha}{n}\right\} \\ &\leq O(1/s). \end{aligned}$$

On the other hand, letting $B = \lceil(1-\alpha)(s+1)\rceil - (1-\alpha)(s+1)$, the lower bound for

$$\frac{\lceil(1-\alpha)(s+1)\rceil}{s+1} - \frac{\lceil(1-\alpha)(n+1)\rceil}{n} \geq \left(\frac{(1-\alpha)(s+1)}{s+1} + B\right) - \left((1-\alpha) + \frac{2-\alpha}{n}\right) = B - \frac{2-\alpha}{n} \geq \Omega(1/s),$$

where the last inequality is due to the assumption $\lceil(1-\alpha)(s+1)\rceil - (1-\alpha)(s+1) \geq \Omega(1/s)$ and $s \leq \sqrt{n}$. \square

C. Proof for Theorem 4.1

Theorem C.1. (*Theorem 4.1 restated, learning bound of DPSM*) Suppose that Assumption 3.1 holds. For any model $f \in \mathcal{F}$, the following inequality holds with probability at least $1 - \delta$:

$$|\bar{\mathcal{L}}_c^{DM}(f, q^*(f)) - \mathcal{L}_c(f)| \leq \tilde{O}(1/\sqrt{n}).$$

Proof. (Proof of Theorem 4.1)

To prove Theorem 4.1, we need the following three technical lemma and defer their proof after proving Theorem 4.1.

Lemma C.2. (Distance between Q_f and $q^* \in \mathcal{U}(f)$) Let $q_f^* \in \mathcal{U}(f) = \arg \min_{q \in \mathbb{R}} \widehat{\mathcal{L}}^{QR}(f, q)$ is a $(1 - \alpha)$ -quantile computed by minimizing QR loss. Suppose that Assumption 3.1 holds. Then the following inequality holds with probability at least $1 - \delta$:

$$|Q_f - q_f^*| \leq \tilde{O}(1/\sqrt{n}).$$

Lemma C.3. (Lipschitz continuity of $\widehat{\mathcal{L}}_c^{\text{DM}}(f, q)$) $\widehat{\mathcal{L}}_c^{\text{DM}}(f, q)$ is L -Lipschitz continuous in input q with $L = \frac{K}{4\tau_{\text{Sigmoid}}}$.

Lemma C.4. (Generalization error) Let $f : \mathcal{X} \rightarrow \mathbb{R}$ and $\{X_1, \dots, X_n\}$ be i.i.d. samples drawn from an underlying distribution. If there exists $M > 0$ such that $|f(X)| \leq M$ for all $X \in \mathcal{X}$, then with probability at least $1 - \delta$, we have

$$\left| \mathbb{E}_X[f(X)] - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \leq M \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Then we begin to prove Theorem 4.1. Recall $\bar{\mathcal{L}}_c^{\text{DM}}(f) = \widehat{\mathcal{L}}_c^{\text{DM}}(f, q_f^*)$, where $q_f^* \in \mathcal{U}(f)$, and define the empirical version of the conformal alignment loss $\widehat{\mathcal{L}}_c(f)$

$$\widehat{\mathcal{L}}_c(f) = \sum_{i \in \mathcal{D}} \left[\sum_{y \in \mathcal{Y}} \mathbb{1}[S_f(X_i, y) \leq Q_f] \right] = \widehat{\mathcal{L}}_c^{\text{DM}}(f, Q_f), \quad \text{see (4) and (9).}$$

We start with the following triangle inequality:

$$\begin{aligned} |\bar{\mathcal{L}}_c^{\text{DM}}(f) - \mathcal{L}_c(f)| &= |\widehat{\mathcal{L}}_c^{\text{DM}}(f, q_f^*) - \widehat{\mathcal{L}}_c(f) + \widehat{\mathcal{L}}_c(f) - \mathcal{L}_c(f)| \leq \underbrace{|\widehat{\mathcal{L}}_c^{\text{DM}}(f, q_f^*) - \widehat{\mathcal{L}}_c^{\text{DM}}(f, Q_f)|}_{L\text{-Lipschitz, Lemma C.3}} + \underbrace{|\widehat{\mathcal{L}}_c(f) - \mathcal{L}_c(f)|}_{\leq \tilde{O}(1/\sqrt{n}), \text{ Lemma C.4}} \\ &\leq L \underbrace{|q_f^*(f) - Q_f|}_{\leq \tilde{O}(1/\sqrt{n}), \text{ Lemma C.2}} + \tilde{O}(1/\sqrt{n}) \leq \tilde{O}(1/\sqrt{n}). \end{aligned}$$

□

C.1. Proof for Lemma C.2

Proof. (of Lemma C.2)

In this proof, define $q_f^*(\alpha) \in \mathcal{U}(f) = \arg \min_{q \in \mathbb{R}} \widehat{\mathcal{L}}^{QR}(f, q)$. We explicitly write the probability $1 - \alpha$ for our analysis.

First, we prove that $q_f^*(\alpha - O(1/\sqrt{n})) \leq Q_f(\alpha) \leq q_f^*(\alpha + O(q/\sqrt{n}))$

Define $Z_i = \mathbb{1}[S_{f,i} \leq Q(\alpha)]$ where $1 \leq i \leq n$. Thus, Z_i is a Bernoulli random variable. According to the definition of $Q_f(\alpha)$, we have that $\mathbb{P}\{Z_i = 1\} = 1 - \alpha$ and $\mathbb{P}\{Z_i = 0\} = \alpha$. Let $\widehat{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$ and $\mathbb{E}[\widehat{Z}] = 1 - \alpha$.

According to Chernoff bound, we know

$$\mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[\widehat{Z}] \right| \geq \varepsilon \mathbb{E}[\widehat{Z}] \right\} \leq 2 \exp\left(-\mathbb{E}[\widehat{Z}] \varepsilon^2 / 3\right) = 2 \exp\left(-n(1 - \alpha) \varepsilon^2 / 3\right).$$

By setting $\delta = 2 \exp(-n(1 - \alpha) \varepsilon^2 / 3)$, i.e., $\varepsilon = \sqrt{(3 \log(2/\delta)) / ((1 - \alpha)n)}$, we have with probability at least $1 - \delta$:

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}[V_i \leq Q_f(\alpha)] - (1 - \alpha) \right| \leq \varepsilon(1 - \alpha) = \sqrt{(3(1 - \alpha) \log(2/\delta)) / n} = \tilde{O}(1/\sqrt{n}). \quad (20)$$

Recall the definition of $q_f^*(\alpha) \in \arg \min_{q \in \mathbb{R}} \widehat{\mathcal{L}}^{QR}(f, q)$. Then we know the following upper bound and lower bound for $1 - \alpha$:

$$(1 - \alpha) \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}[S_{f,i} \leq q_f^*(\alpha)], \quad (1 - \alpha) \geq \frac{1}{n} \sum_{i=1}^n \mathbb{1}[S_{f,i} \leq q_f^*(\alpha + 1/n)].$$

Re-arranging (20) and using the above upper/lower bounds, with probability at least $1 - \delta$, we have

$$\begin{aligned}
 (1 - \alpha)(1 - \varepsilon) &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}[S_{f,i} \leq Q_f(\alpha)] \leq (1 - \alpha)(1 + \varepsilon) \\
 \Leftrightarrow 1 - \underbrace{(1 - (1 - \alpha)(1 - \varepsilon))}_{=\alpha'} &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}[S_{f,i} \leq Q_f(\alpha)] \leq 1 - \underbrace{(1 - (1 - \alpha)(1 + \varepsilon))}_{=\alpha''} \\
 \Rightarrow \frac{1}{n} \sum_{i=1}^n \mathbb{1}[S_{f,i} \leq q_f^*(\alpha' + 1/n)] &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}[S_{f,i} \leq Q_f(\alpha)] \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}[S_{f,i} \leq q_f^*(\alpha'')] \\
 \Leftrightarrow q_f^*(\alpha' + 1/n) &\leq Q_f(\alpha) \leq q_f^*(\alpha''). \tag{21}
 \end{aligned}$$

Finally, by using the definition of ε above, we analyze α' and α'' as follows

$$\alpha' = 1 - (1 - \alpha)(1 - \varepsilon) = \alpha + \varepsilon(1 - \alpha) = \alpha + \sqrt{3(1 - \alpha) \log(2/\delta)/n} = \alpha + \tilde{O}(1/\sqrt{n}), \tag{22}$$

$$\alpha'' = 1 - (1 - \alpha)(1 + \varepsilon) = \alpha - \varepsilon(1 - \alpha) = \alpha - \sqrt{3(1 - \alpha) \log(2/\delta)/n} = \alpha - \tilde{O}(1/\sqrt{n}). \tag{23}$$

Thus, plugging (22) and (23) into (21), we have

$$q_f^*(\alpha + \tilde{O}(1/\sqrt{n})) \leq Q_f(\alpha) \leq q_f^*(\alpha - \tilde{O}(1/\sqrt{n})). \tag{24}$$

Next, we prove that $|Q_f(\alpha) - q_f^*(\alpha)| \leq \tilde{O}(1/\sqrt{n})$ from (24):

$$|Q_f(\alpha) - q_f^*(\alpha)| \leq L_2 \left| \frac{\lceil (1 - \alpha)(n + 1) \rceil}{n} - \frac{\lceil (1 - \alpha - \tilde{O}(1/\sqrt{n}))(n + 1) \rceil}{n} \right| \leq \tilde{O}(1/\sqrt{n}),$$

where the first inequality is due to Assumption 3.1. \square

C.2. Proof for Lemma C.3

Proof. (of Lemma C.3)

Recall that $\tilde{\mathbb{1}}[S \leq q] = \frac{1}{1 + \exp(-\frac{q-S}{\tau_{\text{Sigmoid}}})}$. Define $u = \frac{q-S}{\tau_{\text{Sigmoid}}}$.

We start with the gradient of $\hat{\mathcal{L}}_c^{\text{DM}}(f, q)$ in q as follows

$$\begin{aligned}
 \frac{\partial \hat{\mathcal{L}}_c^{\text{DM}}(f, q)}{\partial q} &= \frac{\partial \sum_{y \in \mathcal{Y}} \tilde{\mathbb{1}}[S \leq q]}{\partial q} = \sum_{y \in \mathcal{Y}} \left(\frac{\partial (\frac{1}{1 + \exp(-u)})}{\partial u} \cdot \frac{\partial u}{\partial q} \right) = \sum_{y \in \mathcal{Y}} \left(\frac{\exp(-u)}{(1 + \exp(-u))^2} \cdot \frac{1}{\tau_{\text{Sigmoid}}} \right) \\
 &= \sum_{y \in \mathcal{Y}} \frac{\exp(-\frac{q-S}{\tau_{\text{Sigmoid}}})}{\tau_{\text{Sigmoid}} (1 + \exp(-\frac{q-S}{\tau_{\text{Sigmoid}}}))^2} > 0, \tag{25}
 \end{aligned}$$

where the first inequality is due to $\exp(x) > 0$. Thus, $\hat{\mathcal{L}}_c^{\text{DM}}(f, q)$ is a increasing function in input q .

For each component of the summation in above (25), define $h(u) = \frac{\exp(-u)}{(1 + \exp(-u))^2}$, then the gradient of $h(u)$ is:

$$\begin{aligned}
 h'(u) &= \frac{-\exp(-u)(1 + \exp(-u))^2 - \exp(-u) \cdot 2(1 + \exp(-u)) \cdot (-\exp(-u))}{(1 + \exp(-u))^4} \\
 &= \frac{(1 + \exp(-u)) \cdot \exp(-u) \cdot ((1 + \exp(-u)) + 2(\exp(-u)))}{(1 + \exp(-u))^4} \\
 &= \frac{(1 + \exp(-u)) \cdot \exp(-u) \cdot (\exp(-u) - 1)}{(1 + \exp(-u))^4}. \tag{26}
 \end{aligned}$$

We can take a closer look at the value of $h'(u)$ in (26): when $\exp(-u) \in (0, 1)$, or equivalently $u > 0$, we have $h'(u) > 0$; when $\exp(-u) \in (1, +\infty)$, or equivalently $u < 0$, we have $h'(u) < 0$. It means that $h(u)$ increases when $u > 0$ and decreases when $u < 0$. Further, when $u = 0$, we achieve the maximal value of $h(u)$, i.e., $\max_u h(u) = h(0) = 1/4$, or equivalently $\frac{\exp(-u)}{(1+\exp(-u))^2} \leq 1/4$.

Therefore, considering $\frac{\partial \hat{\mathcal{L}}_c^{\text{DM}}(f, q)}{\partial q} > 0$ as in (25) and plugging the above inequality into (25), $\left| \frac{\partial \hat{\mathcal{L}}_c^{\text{DM}}(f, q)}{\partial q} \right|$ is bounded as follows

$$\left| \frac{\partial \hat{\mathcal{L}}_c^{\text{DM}}(f, q)}{\partial q} \right| \leq \left| \sum_{y \in \mathcal{Y}} \frac{\exp(-\frac{q-S}{\tau_{\text{Sigmoid}}})}{\tau_{\text{Sigmoid}} \left(1 + \exp(-\frac{q-S}{\tau_{\text{Sigmoid}}})\right)^2} \right| \leq \sum_{y \in \mathcal{Y}} \frac{1}{4\tau_{\text{Sigmoid}}} = \frac{K}{4\tau_{\text{Sigmoid}}},$$

if $\tau_{\text{Sigmoid}} \neq 0$ and $|\mathcal{Y}| = K$. Therefore, $\mathcal{L}_c(f, q)$ is L -Lipschitz continuous for input q with $L = \frac{K}{4\tau_{\text{Sigmoid}}}$. \square

C.3. Proof for Lemma C.4

Proof. (of Lemma C.4)

Let $A_i = f(X_i)$ with $|A_i| \leq M$. According to Hoeffding's inequality, we have:

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n A_i - \mathbb{E}[A] \right| \geq \epsilon \right\} \leq 2 \exp \left(- \frac{2n\epsilon^2}{M^2} \right).$$

By setting $\delta = 2 \exp(-2n\epsilon^2/M^2)$, i.e., $\epsilon = M \sqrt{\log(2/\delta)/(2n)}$, we have the following inequality:

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n A_i - \mathbb{E}[A] \right| \leq M \sqrt{\frac{\log(2/\delta)}{2n}} \right\} \geq 1 - \delta.$$

Therefore, with probability at least $1 - \delta$, we have

$$\left| \mathbb{E}_X[f(X)] - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \leq M \sqrt{\frac{\log(2/\delta)}{2n}} \leq \tilde{O}(1/\sqrt{n}).$$

\square

D. Proofs for Results in Section 4.2

Lemma D.1. (*Lemma 4.5 restated, Hölderian error bound condition for QR loss*) Suppose there is no tie in $\{S_i\}_{i=1}^n$. Then, fixing f , the QR loss $\hat{\mathcal{L}}^{\text{QR}}(f, q)$ satisfies Hölderian error bound w.r.t. q for $\nu = 1$.

Proof. (of Lemma D.1)

Below we assume f is fixed and let $g(q) \triangleq \hat{\mathcal{L}}^{\text{QR}}(f, q)$, so we omit the dependence of g on f for simplicity of notations. Similarly, we denote the optimal solution set by $\mathcal{U} = \arg \min_q g(q)$ in this proof.

In the following, we consider two cases to prove Lemma D.1.

Case 1. Suppose we have two quantile variables q_1, q_2 such that $q_1 \notin \mathcal{U}$, $q_1 < q_2 \in \mathcal{U}$ and $q_2 = \arg \min_{q' \in \mathcal{U}} \|q' - q_1\|$ is

the quantile in the optimal set closest to q_1 . Then we have

$$\begin{aligned}
 n(g(q_1) - g(q_2)) &= \sum_{i=1}^n (\rho_\alpha(q_1, S_i) - \rho_\alpha(q_2, S_i)) \\
 &= \sum_{i=1}^n \left(\mathbb{1}[S_i \leq q_1] \left(\alpha(q_1 - S_i) - \alpha(q_2 - S_i) \right) + \mathbb{1}[q_1 < S_i < q_2] \left((1-\alpha)(S_i - q_1) - \alpha(q_2 - S_i) \right) \right. \\
 &\quad \left. + \mathbb{1}[q_2 \leq S_i] \left((1-\alpha)(S_i - q_1) - (1-\alpha)(S_i - q_2) \right) \right) \\
 &= \sum_{i=1}^n \left(\mathbb{1}[S_i \leq q_1] \left(\alpha(q_1 - q_2) \right) + \mathbb{1}[q_1 < S_i < q_2] \left(S_i - q_1 + \alpha(q_1 - q_2) \right) \right. \\
 &\quad \left. + \mathbb{1}[q_2 \leq S_i] \left((1-\alpha)(q_2 - q_1) \right) \right) \tag{27}
 \end{aligned}$$

$$\begin{aligned}
 &= \sum_{i=1}^n \left(\mathbb{1}[S_i \leq q_2] \alpha(q_1 - q_2) + \mathbb{1}[q_2 \leq S_i] (1-\alpha)(q_2 - q_1) + \mathbb{1}[q_1 < S_i < q_2] (S_i - q_1) \right) \\
 &= \sum_{i=1}^n \mathbb{1}[S_i \leq q_2] \alpha(q_1 - q_2) + \sum_{i=1}^n \mathbb{1}[q_2 \leq S_i] (1-\alpha)(q_2 - q_1) + \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2] (S_i - q_1) \\
 &= (1-\alpha)\alpha(q_1 - q_2) - \alpha(1-\alpha)(q_1 - q_2) + \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2] (S_i - q_1) \\
 &= \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2] (S_i - q_1). \tag{28}
 \end{aligned}$$

On the other hand, let $\tilde{n} = \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2]$. Denote $(1-c')q_1 + c'q_2$ as the weighted average between q_1 and q_2 with a weight parameter $c' \in (0, 1)$. Due to that the weighted average between q_1 and q_2 can be always smaller than the average score in the set $\{S_i : q_1 < S_i < q_2\}$ for a sufficiently small $c' \in (0, 1)$, we have

$$\begin{aligned}
 q_1 + c'(q_2 - q_1) &= (1-c')q_1 + c'q_2 \leq \frac{1}{\tilde{n}} \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2] S_i \\
 \Leftrightarrow \tilde{n}(q_1 + c'(q_2 - q_1)) &\leq \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2] (S_i - q_1 + q_1) \\
 \Leftrightarrow c'\tilde{n}(q_2 - q_1) &\leq \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2] (S_i - q_1) \stackrel{(28)}{=} n(g(q_1) - g(q_2)) \\
 \Leftrightarrow \frac{c'\tilde{n}}{n}(q_2 - q_1) &\leq g(q_1) - g(q_2).
 \end{aligned}$$

Recall that $q_2 = \arg \min_{q' \in \mathcal{U}} \|q' - q_1\|$ and $(q_2 - q_1) = \text{dist}(q_1, \mathcal{U})$. Plug it into the above equality, we have

$$\frac{c'\tilde{n}}{n} \text{dist}(q_1, \mathcal{U}) \leq g(q_1) - g(q_2).$$

Denoting $c = n/(\tilde{n}c') > 0$, then we have

$$\text{dist}(q_1, \mathcal{U}) \leq c(g(q_1) - \min_{q'} g(q)), \tag{29}$$

which satisfies the HEB condition shown as Definition 4.3 with the exponent $\nu = 1$.

Case 2. Suppose that $q_1 \in \mathcal{U}$, $q_1 < q_2 \notin \mathcal{U}$ and $q_1 = \arg \min_{q' \in \mathcal{U}} \|q' - q_2\|$. We can start from (27) as follows

$$\begin{aligned} n(g(q_1) - g(q_2)) &= \sum_{i=1}^n \left(\mathbb{1}[S_i \leq q_1] (\alpha(q_1 - q_2)) + \mathbb{1}[q_1 < S_i < q_2] (S_i - q_1 + \alpha(q_1 - q_2)) \right. \\ &\quad \left. + \mathbb{1}[q_2 \leq S_i] ((1 - \alpha)(q_2 - q_1)) \right) \\ &= \sum_{i=1}^n \left(\mathbb{1}[S_i \leq q_1] \alpha(q_1 - q_2) + \mathbb{1}[q_1 \leq S_i] (1 - \alpha)(q_2 - q_1) + \mathbb{1}[q_1 < S_i < q_2] (S_i - q_2) \right) \\ &= \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2] (S_i - q_2). \end{aligned}$$

With the same notation of $\tilde{n} = \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2]$, we know that there must exist a constant $c' \in (0, 1)$ such that

$$q_2 + c'(q_1 - q_2) = c'q_1 + (1 - c')q_2 \geq \frac{1}{\tilde{n}} \sum_{i=1}^n \mathbb{1}[q_1 < S_i < q_2] S_i$$

Therefore, we have

$$g(q_1) - g(q_2) \leq \frac{\tilde{n}}{n} c'(q_1 - q_2) = -\frac{\tilde{n}}{n} c' \text{dist}(q_2, \mathcal{U}).$$

Denoting $c = n/(\tilde{n}c) > 0$, then we have

$$\text{dist}(q_2, \mathcal{U}) \leq c(g(q_2) - \min_{q' \in \mathcal{U}} g(q')). \quad (30)$$

By combining (29) and (30), we show that the QR loss satisfies the Hölderian error bound condition with a constant $c > 0$ and an exponent $\nu = 1$. \square

E. Additional Experimental Setup Details

Dataset and Split. We consider the benchmark datasets CIFAR-100 (Krizhevsky et al., 2009), Caltech-101 (Fei-Fei et al., 2004), and iNaturalist (Van Horn et al., 2018). We split the original testing datasets into used calibration and testing datasets. Table 2 summarizes key statistics of the used datasets which we elaborate on in the following.

Hyperparameters for training. We set datasets, base models, batch size, training epochs, training parameters (learning rate, learning schedule, momentum, gamma, and weight decay), and λ as hyperparameter choices. We search for hyperparameters on batch size $\in \{64, 128\}$, epochs $\in \{30, 40, 60\}$, learning rate (η) $\in \{0.001, 0.005, 0.01, 0.05, 0.1\}$, learning rate schedule $\in \{[3], [25], [25, 40]\}$, Momentum = 0.9, weight decay $\in \{0.1, 0.97\}$, and $\lambda = \{0.01, 0.05, 0.1, 0.5, 1.0, 5.0\}$ to select the best combination of hyperparameters of each methods. We also search the learning rate for lower function (γ) $\in \{0.0001, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5\}$ for DPSM. The hyperparameters employed to get the results presented in the main paper are summarized in Table 3.

Table 2. Description of the data sets are given in the table. *The number of classes in the iNaturalist data set depends on the taxonomy level (e.g., species, genus, family). We employ "Fungi" species which has 341 different categories.

Data	Number of Classes	Number of Training Data	Number of Validation Data	Number of Calibration Data	Number of Test Data
CIFAR-100	100	45000	5000	3000	7000
Caltech-101	101	4310	1256	1111	2000
iNaturalist	341*	15345	1705	1410	2000

Table 3. The below table shows the details we used to train our models. We reported the hyperparameters which gives the best predictive efficiency. We employed SGD optimizer for all training unless specified.

Data	Architecture	Batch size	Epochs	η	lr schedule	Momentum	weight decay	γ	λ
CIFAR-100	DenseNet	64	40	0.1	25	0.9	0.1	0.01	0.05
	ResNet	128	40	0.1	25	0.9	0.1	0.01	0.01
Caltech-101	DenseNet	128	60	0.05	25, 40	0.9	0.1	0.1	1.0
	ResNet	128	60	0.05	25, 40	0.9	0.1	0.05	0.1
iNaturalist	DenseNet	128	60	0.001	3	0.9	0.97	0.001	1.0
	ResNet	128	60	0.001	3	0.9	0.97	0.001	0.5

HPS → Training/Calibration/Testing. In our study, we primarily apply cross-entropy loss for the classification model. For the ConfTr and DPSM methods, we minimize both the classification loss and the HPS non-conformity score-based predictive inefficiency loss during training. Post-training, we estimate the HPS non-conformal scores using equation 11 for all three methods during calibration and testing. We report the marginal coverage and the prediction set size with the corresponding mean and standard deviation, with the results presented in Table 4.

HPS → Training, APS → Calibration/Testing. In our study, we primarily apply cross-entropy loss for the classification model. For the ConfTr and DPSM methods, we minimize both the classification loss and the differentiable HPS non-conformity score-based inefficiency loss during training. After training, we estimate APS non-conformal scores using equation 12. We calculate the marginal coverage and the prediction set size, with the results presented in Table 5.

HPS → Training, RAPS → Calibration/Testing. In our study, we primarily apply cross-entropy loss for the classification model. For the ConfTr and DPSM methods, we minimize both the classification loss and the differentiable HPS non-conformity score-based inefficiency loss during training. After training, we estimate RAPS non-conformal scores using equation 13. Then, we calculate the marginal coverage and the prediction set size, with the results presented in Table 6.

F. Additional Experiments

F.1. Additional Experiments for Marginal Coverage

Table 4. HPS → Training/Calibration/Testing: The APSS on three different datasets with two different deep models trained and calibrated with HPS when $\alpha = 0.1$. \downarrow indicates the percentage improvement in predictive efficiency compared to the best existing method, whereas \uparrow denotes a percentage decrease in predictive efficiency. All results are the average over 10 different runs, with the mean and standard deviation reported. DPSM significantly outperforms almost the best baselines with 20.32% prediction set size reduction across all datasets.

Model	Marginal Coverage				Prediction Set Size			
	CE	CUT	ConfTr	DPSM	CE	CUT	ConfTr	DPSM
CalTech-101								
DenseNet	0.90 ± 0.006	0.90 ± 0.005	0.90 ± 0.008	0.90 ± 0.003	3.50 ± 0.10	1.62 ± 0.030	4.10 ± 0.19	0.90 ± 0.003 ($\downarrow 44.44\%$)
ResNet	0.90 ± 0.004	0.90 ± 0.007	0.90 ± 0.006	0.90 ± 0.005	1.57 ± 0.018	1.64 ± 0.049	1.52 ± 0.040	0.91 ± 0.005 ($\downarrow 44.51\%$)
CIFAR-100								
DenseNet	0.90 ± 0.007	0.90 ± 0.009	0.90 ± 0.007	0.90 ± 0.006	2.59 ± 0.053	2.27 ± 0.09	2.28 ± 0.07	2.17 ± 0.086 ($\downarrow 4.82\%$)
ResNet	0.90 ± 0.006	0.90 ± 0.005	0.90 ± 0.007	0.90 ± 0.007	3.39 ± 0.10	3.01 ± 0.11	3.77 ± 0.14	2.94 ± 0.08 ($\downarrow 2.32\%$)
iNaturalist								
DenseNet	0.90 ± 0.009	0.90 ± 0.011	0.90 ± 0.011	0.90 ± 0.008	94.58 ± 3.45	77.13 ± 3.72	79.93 ± 3.70	61.22 ± 2.49 ($\downarrow 20.63\%$)
ResNet	0.90 ± 0.019	0.90 ± 0.007	0.90 ± 0.012	0.90 ± 0.008	99.48 ± 8.95	73.09 ± 2.00	76.73 ± 3.87	70.04 ± 1.99 ($\downarrow 4.17\%$)

DPSM generates smaller prediction sets. Table 4 presents the set sizes and coverage rates for different methods using HPS score across training, calibration, and testing phases. DPSM outperforms all existing baselines with 20.32% reduction in terms of prediction set size across all datasets. Table 5 presents the set sizes and coverage rates for different methods using HPS for training and APS for calibrating. DPSM outperforms nearly all existing baselines with 20.61% reduction in terms of prediction set size across all datasets, except $\uparrow 8.71\%$ increase on CIFAR-100 with DenseNet and $\uparrow 0.54\%$ increase on iNaturalist with ResNet in terms of prediction set size. Table 6 presents the set sizes and coverage rates for different methods

Table 5. HPS → Training, APS → Calibration/Testing: The APSS on three different datasets with two different deep models trained with HPS and calibrated with APS when $\alpha = 0.1$. \downarrow indicates the percentage improvement in predictive efficiency compared to the best existing method, whereas \uparrow denotes a percentage decrease in predictive efficiency. All results are the average over 10 different runs, with the mean and standard deviation reported. DPSM significantly outperforms almost the best baselines with 20.61% prediction set size reduction across all datasets.

Model	Marginal Coverage				Prediction Set Size			
	CE	CUT	ConfTr	DPSM	CE	CUT	ConfTr	DPSM
CalTech-101								
DenseNet	0.90 ± 0.006	0.90 ± 0.007	0.90 ± 0.008	0.90 ± 0.005	8.44 ± 0.15	3.87 ± 0.11	8.64 ± 0.21	1.58 ± 0.022 (\downarrow 59.17%)
ResNet	0.90 ± 0.006	0.90 ± 0.004	0.90 ± 0.007	0.90 ± 0.005	4.50 ± 0.059	4.59 ± 0.072	3.61 ± 0.08	1.74 ± 0.031 (\downarrow 51.80%)
CIFAR-100								
DenseNet	0.90 ± 0.007	0.90 ± 0.009	0.90 ± 0.008	0.90 ± 0.006	3.38 ± 0.12	2.41 ± 0.11	3.08 ± 0.11	2.64 ± 0.86 (\uparrow 8.71%)
ResNet	0.90 ± 0.006	0.90 ± 0.11	0.90 ± 0.007	0.90 ± 0.007	3.98 ± 0.13	3.81 ± 0.08	4.90 ± 0.18	3.53 ± 0.11 (\downarrow 7.35%)
iNaturalist								
DenseNet	0.90 ± 0.009	0.90 ± 0.010	0.90 ± 0.011	0.90 ± 0.010	101.97 ± 3.16	88.93 ± 3.06	90.79 ± 3.17	75.98 ± 2.99 (\downarrow 14.56%)
ResNet	0.90 ± 0.013	0.90 ± 0.006	0.90 ± 0.012	0.90 ± 0.009	95.81 ± 3.80	79.00 ± 2.21	88.70 ± 3.88	79.43 ± 2.39 (\uparrow -0.54%)

Table 6. HPS → Training, RAPS → Calibration/Testing: The APSS on three different datasets with two different deep models trained with HPS and calibrated with RAPS when $\alpha = 0.1$, where $\lambda_{\text{RAPS}} = 0.01$ and $k_{\text{reg}} = 5$. \downarrow indicates the percentage improvement in predictive efficiency compared to the best existing method, whereas \uparrow denotes a percentage decrease in predictive efficiency. All results are the average over 10 different runs, with the mean and standard deviation reported. DPSM significantly outperforms almost the best baselines with 19.04% prediction set size reduction across all datasets.

Model	Marginal Coverage				Prediction Set Size			
	CE	CUT	ConfTr	DPSM	CE	CUT	ConfTr	DPSM
CalTech-101								
DenseNet	0.90 ± 0.007	0.90 ± 0.008	0.90 ± 0.009	0.90 ± 0.006	6.58 ± 0.12	3.53 ± 0.11	6.88 ± 0.17	1.50 ± 0.020 (\downarrow 57.51%)
ResNet	0.90 ± 0.005	0.90 ± 0.004	0.90 ± 0.007	0.90 ± 0.004	3.89 ± 0.047	3.98 ± 0.59	3.19 ± 0.69	1.67 ± 0.025 (\downarrow 47.65%)
CIFAR-100								
DenseNet	0.90 ± 0.007	0.90 ± 0.007	0.90 ± 0.006	0.90 ± 0.006	2.73 ± 0.043	2.14 ± 0.55	2.69 ± 0.053	2.34 ± 0.035 (\uparrow 8.55%)
ResNet	0.90 ± 0.006	0.90 ± 0.11	0.90 ± 0.007	0.90 ± 0.007	3.25 ± 0.14	2.93 ± 0.60	4.02 ± 0.11	2.93 ± 0.05
iNaturalist								
DenseNet	0.90 ± 0.013	0.90 ± 0.008	0.90 ± 0.015	0.90 ± 0.008	97.27 ± 4.23	82.88 ± 2.47	86.77 ± 6.17	68.17 ± 2.18 (\downarrow 17.74%)
ResNet	0.90 ± 0.012	0.90 ± 0.011	0.90 ± 0.015	0.90 ± 0.009	97.43 ± 4.17	76.75 ± 4.47	81.18 ± 4.93	76.81 ± 2.68 (\uparrow 0.08%)

using HPS for training and RAPS for calibrating. DPSM outperforms existing baselines on several datasets, achieving a 19.04% reduction in prediction set size across all datasets, except for CIFAR-100. On iNaturalist with ResNet, it shows a marginal 0.08% increase in prediction set size. Combined with above three tables, DPSM improve the predictive efficiency with 19.99% reduction in term of prediction set size across all settings and all datasets. We also visualize the coverage rate and APSS with confidence intervals of all methods using DenseNet and HPS score in Figure 5. It clearly confirms that DPSM achieves significantly smaller prediction set size while maintaining the valid coverage.

DPSM converges to stable error for bilevel optimization. To further explore how DPSM effectively generates smaller prediction sets, we analyze the convergence of DPSM by plotting the loss of the upper level function (i.e., a combination of classification loss and conformal alignment loss) and the lower level function (i.e., QR loss) over training 100 epochs with ResNet model. Figure 6 (a) and (b) show the upper-level loss and lower-level loss over epochs, respectively. We also report the results of 40-epoch training regime in Figure 7 and 8 for reference. As shown, the upper-level loss of DPSM exhibits an initial increase during the first 2 epochs, reaching the peak, then steadily decreases before stabilizing around epoch 35. In contrast, the lower-level loss decreases sharply within the first 2 epochs, followed by a gradual reduction until convergence near the end of training. These results empirically demonstrate that DPSM effectively converges in terms of both upper and lower level training errors, validating its bilevel optimization approach. To investigate how the learned quantiles influence the optimization error, we compute both conformal and QR losses using the learned quantiles and the optimal (dataset-level) quantiles. The corresponding optimization errors—defined as the loss differences between learned

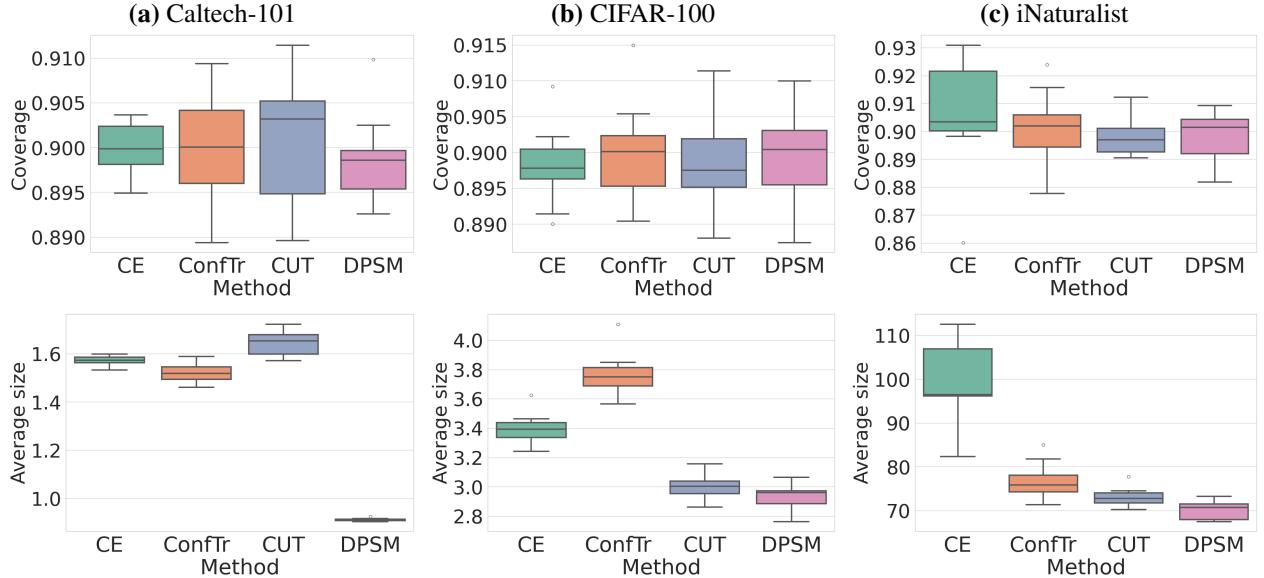


Figure 5. Box plots of coverage (Top row) and APSS (Bottom row) of all methods using ResNet and HPS score. DPSM achieves significantly smaller prediction set size while maintaining the valid coverage.

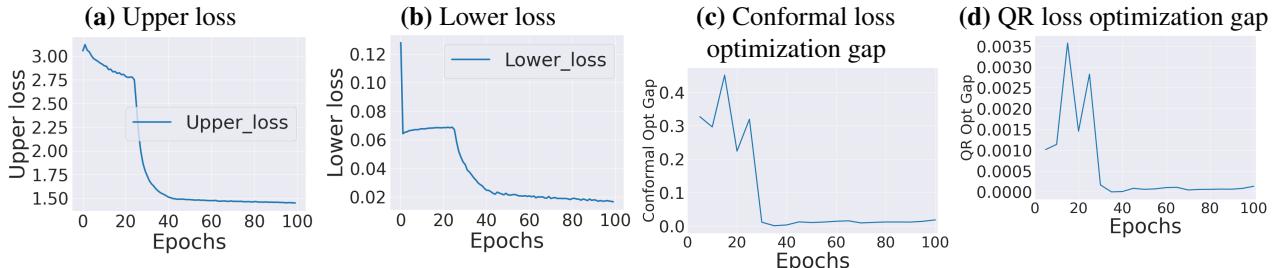


Figure 6. Justification experiments for the convergence of DPSM on CIFAR-100 using ResNet and HPS score. (a) Upper level loss (i.e., a combination of classification loss and conformal alignment loss); (b) Lower level loss (i.e., QR loss); (c) Optimization gap of conformal loss, defined as the difference between conformal losses using learned batch-level quantiles and dataset-level quantiles on the training set; (d) Optimization gap of the lower-level QR loss, defined similarly as the loss difference between learned batch-level quantiles and dataset-level quantiles.

and optimal quantiles—are visualized in Figure 6 (c) and (d). Both errors converge to nearly 0, indicating that the learned quantiles effectively approximate the optimal quantiles over training.

DPSM estimates empirical quantiles with small error. To compare the precision of empirical quantiles estimation in ConfTr and DPSM, we plot the estimation error between \hat{Q}_f^n (quantiles evaluated on the whole training dataset) and \hat{q}_f (quantiles evaluated in ConfTr or learned in DPSM on mini-batches) with ResNet. Figure 9 (a) plots this estimation error over training epochs. For the first 25 epochs, the estimation errors for DPSM are significantly larger compared to ConfTr. However, as the training progresses, the estimation errors for DPSM decrease rapidly, converging close to 0 after epoch 32. This result verifies the theoretical result for smaller estimation error in learning bound analysis from Theorem 4.1. Furthermore, the rapid reduction in estimation error also reflects the convergence of the lower loss (i.e., QR loss), highlighting the effectiveness of DPSM in accurately estimating quantiles.

Learning bound of DPSM is much tighter than ConfTr. To approximately compare the learning bounds of DPSM and ConfTr, we compare the conformal alignment losses of ConfTr and DPSM during training in terms of the average soft set size, as shown in Figure 9 (b). The soft set size of DPSM is consistently smaller than that of ConfTr during training. Combining the empirical results of the smaller estimation error of quantiles from Figure 9 (a), we can conclude that the learning bound of DPSM is much tighter than the learning bound of ConfTr, providing the empirical verification of Theorem 4.1 and 3.5. Although learning bound cannot be empirically computed, we approximate it using a common strategy in ML literature (Yuan et al., 2019; Yang et al., 2021b), which estimates generalization error by the absolute gap between training and test errors. For CP, we use APSS evaluated on train and test sets to approximate the learning errors. Specifically, for

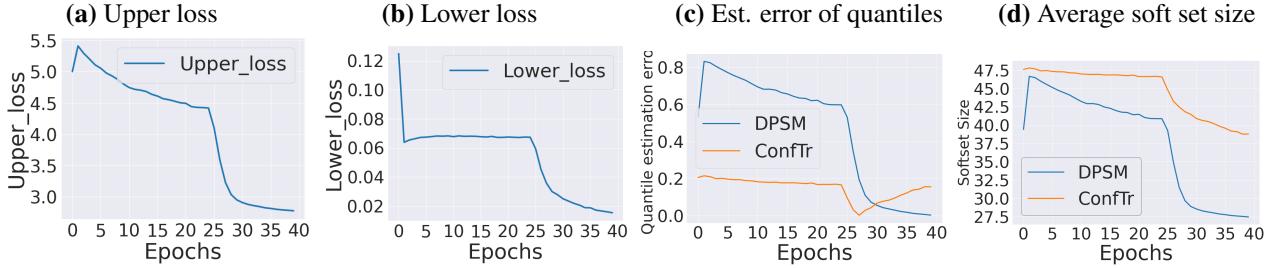


Figure 7. Justification experiments for effectiveness of DPSM trained with 40 epochs on CIFAR-100 using DenseNet and HPS score. (a) Upper level loss (i.e., a combination of classification loss and conformal alignment loss) in DPSM; (b) Lower level loss (i.e., QR loss) in DPSM; (c) Estimation error between the \hat{Q}_f^n (quantiles evaluated on the whole training data) and \hat{q}_f (quantiles evaluated in ConfTr or learned in DPSM on mini-batches); and (d) Average soft set size of DPSM and ConfTr (using Sigmoid function).

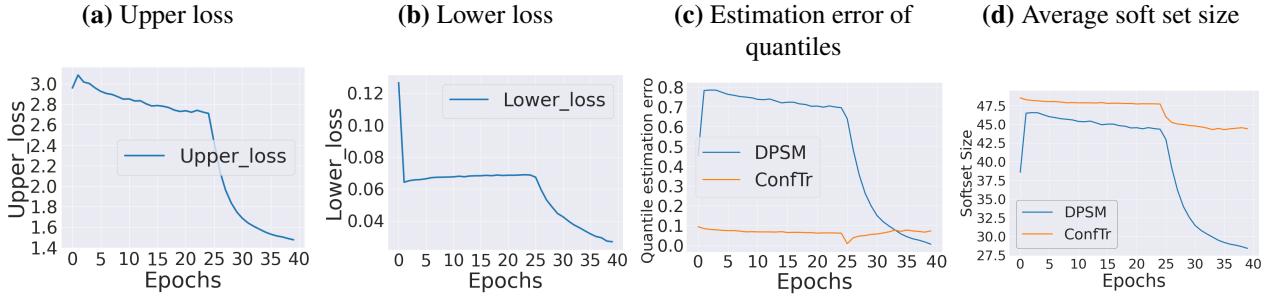


Figure 8. Justification experiments for effectiveness of DPSM trained with 40 epochs on CIFAR-100 using ResNet and HPS score. (a) Upper level loss (i.e., a combination of classification loss and conformal alignment loss) in DPSM; (b) Lower level loss (i.e., QR loss) in DPSM; (c): Estimation error between the \hat{Q}_f^n (quantiles evaluated on the whole training data) and \hat{q}_f (quantiles evaluated in ConfTr or learned in DPSM on mini-batches); and (d): Average soft set size of DPSM and ConfTr.

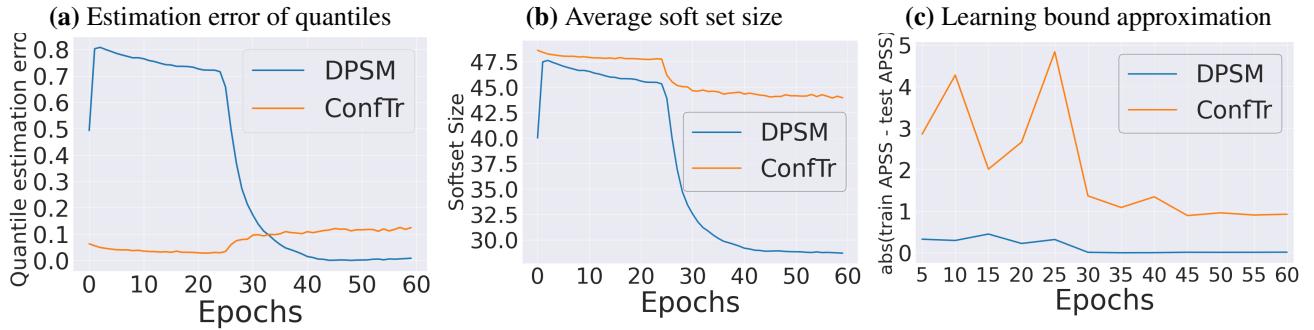


Figure 9. Justification experiments for the learning bound of DPSM on CIFAR-100 using ResNet and HPS score. (a) Estimation error between the \hat{Q}_f^n (dataset-level quantiles on training data) and \hat{q}_f (batch-level quantiles evaluated in ConfTr or learned in DPSM); (b) Average soft set size of DPSM and ConfTr (using Sigmoid function); (c) Approximated learning error comparison between DPSM and ConfTr, measured by their gaps between the training and testing APSS.

DPSM, at each iteration, we: (i) compute APSS on the training set using the learned quantiles as thresholds. It includes optimization error since the learned quantiles are not optimal (true dataset-level quantiles); (ii) compute the APSS on the testing set using the dataset-level quantiles as thresholds. The gap between these two APSS values is employed as an approximation of the learning bound. We apply the same strategy to the SA-based ConfTr, where the training APSS is computed using the quantiles evaluated on mini-batches from the training data, and the test APSS is computed using the dataset-level quantiles from the test data. This comparison is shown in Figure 9 (c), which demonstrates that the approximated learning error is improved by DPSM.

Assumption 3.1 is empirically valid. Figure 10 (a) illustrates the conformity scores plotted against their corresponding normalized order. The x-axis represents the normalized order, while the y-axis represents conformity scores. From this figure, it is clear that the curve does not remain near the x-axis or y-axis, indicating that the gradient of conformity scores

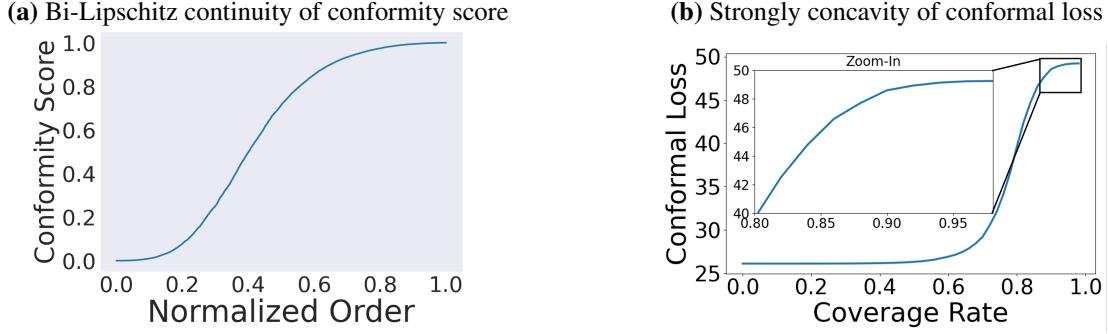


Figure 10. Verification studies on CIFAR-100 with ResNet model using HPS scoring function on calibration dataset. **(a)**: HPS scores over corresponding normalized order produced by ConfTr. The x-axis is the normalized order, the y-axis is the corresponding conformity score; **(b)**: The soft set size measure of ConfTr. The input coverage rate is from [0.02, 0.98] with 0.02 range with its zoom-in version where $1 - \alpha$ is close to target coverage 0.9; When coverage rate is close to 0.9, the curve for the soft set size exhibits a concave shape.

Table 7. The WSC (\uparrow better), SSCV (\downarrow better) and CovGap (\downarrow better) of all methods on CIFAR-100 with HPS score: The best results are in **bold**. These results show that DPSM achieves the best performance for class-conditional coverage (the smallest CovGap). For size-stratified coverage (SSCV), DPSM has a worse measure compared with CE and CUT, but is better than ConfTr. For WSC, CUT and DPSM have comparable performance.

Measures	DenseNet				ResNet			
	CE	CUT	ConfTr	DPSM	CE	CUT	ConfTr	DPSM
WSC	0.88 ± 0.016	0.90 ± 0.022	0.88 ± 0.018	0.89 ± 0.011	0.88 ± 0.012	0.88 ± 0.020	0.88 ± 0.020	0.89 ± 0.019
SSCV	0.12 ± 0.024	0.09 ± 0.019	0.21 ± 0.061	0.17 ± 0.034	0.09 ± 0.018	0.11 ± 0.017	0.14 ± 0.022	0.12 ± 0.019
CovGap	4.54 ± 0.49	5.20 ± 0.29	4.56 ± 0.28	4.43 ± 0.41	4.71 ± 0.38	4.71 ± 0.32	4.70 ± 0.34	4.69 ± 0.26

with respect to normalized index is both upper and lower bounded. This observation empirically supports the validity of Assumption 3.1.

Assumption 3.2 is empirically valid. Figure 10 (b) visualizes the soft set size of ConfTr, with input as coverage rate $\in [0.02, 0.98]$ with range 0.02. When coverage rate approaches 0.9, the curves of all methods exhibit a concave shape (zoom-in version also shown), providing empirical verification for Assumption 3.2.

F.2. Additional Experiments for Conditional Coverage

DPSM achieves comparable conditional coverage performance and strong class-conditional coverage performance compared to baselines. To evaluate the impact of prediction set size reduction on the conditional coverage performance of DPSM, we report three metrics on CIFAR-100 using the HPS score in Table 7: (i) WSC (Worst-Slab Coverage, \uparrow better), introduced in (Romano et al., 2020); (ii) SSCV (Size-Stratified Coverage, \downarrow better), from (Angelopoulos et al., 2021); and (iii) CovGap (Average Class Coverage Gap, \downarrow better), proposed in (Ding et al., 2024) to measure class-conditional coverage. The results show that DPSM achieves the best class-conditional coverage, reflected by the lowest CovGap. For SSCV, DPSM performs slightly worse than CE and CUT, but better than ConfTr. In terms of WSC, DPSM and CUT achieve comparable performance. We further visualize the distribution of class-conditional coverage and class-wise average prediction set size in Figure 11 and 12, providing fine-grained insights into class-conditional performance. DPSM demonstrates slightly more concentrated class-wise coverage and generally smaller class-wise prediction set sizes.

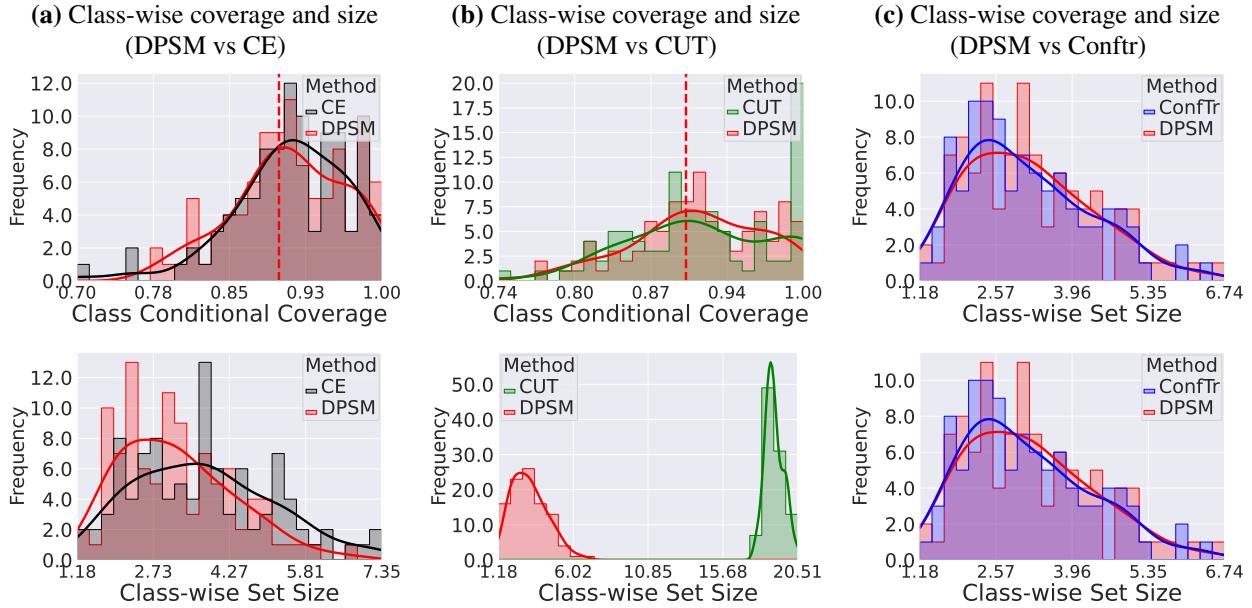


Figure 11. Class conditional coverage and class-wise prediction set size of all methods on CIFAR-100 with DenseNet and HPS. To better compare the class conditional coverage, we compare the class-conditional coverage between DPSM and 3 baselines in (a), (b) and (c) separately. DPSM shows a bit more concentration in terms of class-wise coverage to the nominal coverage (90%) and smaller prediction set size.

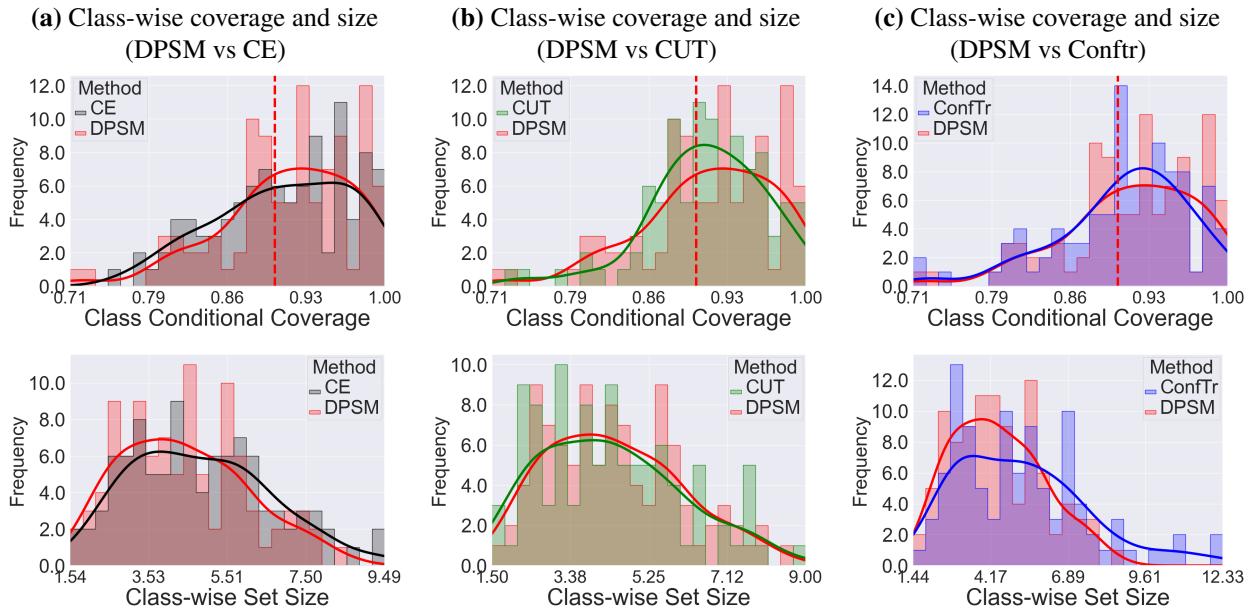


Figure 12. Class conditional coverage and class-wise prediction set size of all methods on CIFAR-100 with ResNet and HPS. To better compare the class conditional coverage, we compare the class-conditional coverage between DPSM and 3 baselines in (a), (b) and (c) separately. DPSM shows a bit more concentration in terms of class-wise coverage to the nominal coverage (90%) and smaller prediction set size.