Conformal Prediction with Learned Features

Shayan Kiyani [1] **George Pappas** [1] **Hamed Hassani** [1]

Abstract

In this paper, we focus on the problem of conformal prediction with conditional guarantees. Prior work has shown that it is impossible to construct nontrivial prediction sets with full conditional coverage guarantees. A wealth of research has considered relaxations of full conditional guarantees, relying on some *predefined* uncertainty structures. Departing from this line of thinking, we propose Partition Learning Conformal Prediction (PLCP), a framework to improve conditional validity of prediction sets through *learning* uncertainty-guided features from the calibration data. We implement PLCP efficiently with alternating gradient descent, utilizing off-the-shelf machine learning models. We further analyze PLCP theoretically and provide conditional guarantees for infinite and finite sample sizes. Finally, our experimental results over four real-world and synthetic datasets show the superior performance of PLCP compared to state-of-the-art methods in terms of coverage and length in both classification and regression scenarios.

1. Introduction

Consider a distribution $\mathcal{D}$ over a domain $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes the space of covariates and $\mathcal{Y}$ denotes the space of labels. Let $f : \mathcal{X} \to \mathcal{Y}$ be a (pre-trained) *model* that provides for every input x a *point estimate* of the corresponding label y. Using the model f and a set of new calibration samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, generated i.i.d. from $\mathcal{D}$, the goal of conformal prediction is to construct for every input x a *prediction set* $C(x)$ that is guaranteed to cover the true label y with high probability. Formally, we say that the prediction sets $C(x) \subseteq \mathcal{Y}$ have *marginal* coverage guarantee if for a test sample (X_{n+1}, Y_{n+1}) we have

$$\Pr\left(Y_{n+1} \in C\left(X_{n+1}\right)\right) = 1 - \alpha, \tag{1}$$

where α is the miscoverage rate, and the probability is taken over the randomness in calibration and test points.

Oftentimes in practice, methods that only guarantee marginal coverage fail to provide valid coverage with respect to specific subgroups or under changing conditions (Romano et al., 2020a; Guan, 2021; Lei & Wasserman, 2014). This issue is particularly evident in applications such as healthcare, where obtaining valid prediction sets for different patient demographics is crucial. For instance, a marginal method might perform well on average over new patients but fail to construct accurate prediction sets for certain age groups or medical conditions.

Ultimately, we may seek to construct prediction sets that achieve *full conditional coverage* which requires for every $x \in \mathcal{X}$

$$\Pr\left(Y_{n+1} \in C\left(X_{n+1}\right) \mid X_{n+1} = x\right) = 1 - \alpha. \tag{2}$$

Despite the importance of achieving conditional guarantees, there are some fundamental limitations. Prior work (Vovk, 2012; Foygel Barber et al., 2019; Lei & Wasserman, 2014) has shown that it is impossible to construct nontrivial prediction sets with distribution-free, full conditional coverage when we have access to a finite-size calibration set. Consequently, relaxations of (2) have been considered. For instance, (Gibbs et al., 2023; Tibshirani et al., 2019) develop frameworks to guarantee coverage under a predefined class of covariate shifts. Another line of work (Romano et al., 2020a; Jung et al., 2023; Barber et al., 2019) considers predefined groups of the covariates and guarantees coverage conditioned on those groups. A more detailed discussion on the existing methods and their implications is provided in the related works section 1.1.

In this paper, we take a new approach and, instead of considering predefined structures, propose to *learn* structures from the calibration data that are *informative about uncertainty quantification*. Our algorithmic framework aims at learning such structures in conjunction with constructing the prediction sets in an iterative fashion. To better illustrate our approach and contributions, we will proceed with the following toy example.

[1]The Electrical and Systems Engineering Department, University of Pennsylvania, University of Pennsylvania, USA. Correspondence to: Shayan Kiyani <shayank@seas.upenn.edu>, George Pappas <pappasg@seas.upenn.edu>, Hamed Hassani <hassani@seas.upenn.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

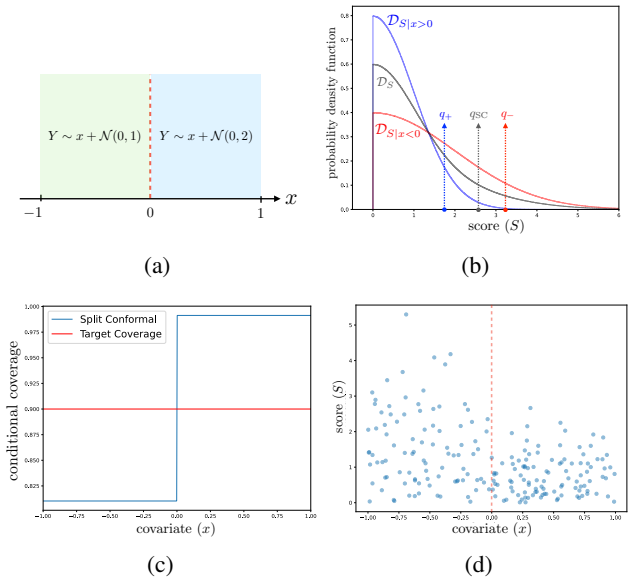

Figure 1: (a) Distribution of the labels conditioned on the covariate (x). (b) Conditional and marginal distributions of the score. (c) Coverage of the prediction sets of Split Conformal conditioned on x. (d) Samples of the form (x_i, s_i). From these samples, we aim to learn the partition/feature h of the covariate space shown by the dashed red line.

Example. Let $\mathcal{X} = [-1, +1]$ and assume that the covariate X is uniformly distributed over $\mathcal{X}$. The label Y is distributed as follows (see also Figure 1-(a)):

- If $x < 0$, then $Y \sim x + \mathcal{N}(0, 1)$,
- If $x \geq 0$, then $Y \sim x + \mathcal{N}(0, 2)$.

For simplicity, we assume in this example that we have infinitely many data points available from this distribution. It is easy to see that the optimal[1] regression model for this distribution is $f(x) = \mathbb{E}[Y \mid X = x] = x$. As a result, considering the conformity score $S(x, y) = |y - f(x)|$, the distribution of S follows the folded Gaussian distribution: i.e. for $x < 0$ we have $\mathcal{D}_{S|x} = |N(0, 1)|$, and for $x \geq 0$ we have $\mathcal{D}_{S|x} = |N(0, 2)|$. Figure 1-(b) depicts the conditional and marginal distributions of S with their 0.9-quantiles.

In this setting, the standard Split Conformal method only looks at the marginal distribution of S and constructs the prediction sets according to its quantile value q_{SC}. This choice of prediction sets, although being marginally valid as in (1), will lead to over-coverage when $x < 0$, and under-coverage when $x \geq 0$ (see Figure 1-(c)). For this specific example, this limitation can be overcome by constructing the prediction sets based on the sign of the covariate. Let q_+ (rep. q_-) be the $(1 - \alpha)$-quantile of the distribution of S conditioned on $x \geq 0$ (rep. $x < 0$). Then the following prediction sets will provide a valid coverage on both positive and negative covariates (in fact, here we have full conditional coverage as in (2)):

- If $x < 0$, then $C(x) = \{y \in \mathbb{R} \text{ s.t. } S(x, y) \leq q_-\}$,
- If $x \geq 0$, then $C(x) = \{y \in \mathbb{R} \text{ s.t. } S(x, y) \leq q_+\}$.

A few points about the above example are in order: (i) This example showcases how we can obtain richer, conditionally-valid prediction sets by using an appropriate partitioning of the covariate space, and constructing the prediction sets according to the partitions. Put differently, the prediction sets were obtained using a new feature, $h(x) = \mathrm{sign}(x)$, of the covariates. (ii) Further, in the above example, $f(x) = x$ is the optimal pointwise prediction of the label. Thus, the feature h is not useful in obtaining a more accurate pointwise estimate of the label; but rather, it is useful in quantifying the *uncertainty* in estimating the label. (iii) Finally, in reality informative features such as h are not known apriori and they have to be *learnt* from data. This is a challenging task as we need to find features, or boundaries in the covariate space, that can separate points $x, x' \in \mathcal{X}$ based on how different the conditional distributions $\mathcal{D}_{S|x}$ and $\mathcal{D}_{S|x'}$ are. Yet, these conditional distributions are not known and only a finite number of samples of the form (x_i, s_i) are given (see Figure 1-(d)).

In light of the example above, our goal is to identify, in a data-driven manner, a partitioning of the covariate space such that points in the same partition share some similarities in terms of their prediction sets. For example, for two points in the same partition, we would like their respective quantile of the conditional distribution of the score to be close to each other. The main challenge here is that the distributions are unknown and only a finite-size calibration set is given.

Algorithmic Contributions. We formalize the problem of learning such uncertainty-guided partitions in Section 2. We will first derive an optimization objective that measures the quality of a partitioning based on how far the conditional quantiles of the points in that partition differ from each other. Our main algorithm, PLCP, optimizes this objective by iteratively updating the partitioning and prediction sets over a given calibration data set. The partitioning is chosen over a function class, e.g. linear functions or neural networks. In this sense, PLCP can systematically utilize off-the-shelf machine learning models very efficiently to construct the prediction sets.

Theoretical Contributions. We introduce the notion of "Mean Squared Conditional Error (MSCE)" defined as

$$\mathrm{MSCE}(\mathcal{D}, \alpha, C) = \mathbb{E}\left[\left(\mathrm{cov}(X) - (1 - \alpha)\right)^2\right] \qquad (3)$$

[1]Here, optimality is measured in terms of the mean squared error.

$$\text{where:} \quad \text{cov}(x) = \Pr[Y \in C(X) \mid X = x],$$

which measures the deviation of prediction sets $C(x)$ from (2). For the ease of notation, we will drop the arguments inside $\text{MSCE}(\mathcal{D}, \alpha, C)$ and simply use MSCE. In section 3, we establish two insightful theoretical guarantees for the prediction sets constructed by PLCP. In the infinite data regime, we demonstrate that the MSCE of PLCP's prediction sets scales as $O(\frac{1}{\sqrt{m}})$, where m denotes the number of regions in the learned partitioning of the covariate space. Hence, as the number of partitions increases, the deviation from the nominal coverage, $1-\alpha$, diminishes across the covariate space. Further, we present a finite sample guarantee for the MSCE of the prediction sets obtained by PLCP. We show that the MSCE scales as $O\left(\sqrt{\frac{m \log n + \text{complexity}(\mathcal{H})}{n}} + \frac{1}{\sqrt{m}}\right)$, with high probability, where $\mathcal{H}$ represents the class of functions used by PLCP, and $\text{complexity}(\mathcal{H})$ quantifies its complexity (i.e. the covering number). Such PAC-style guarantees are common in the conformal prediction literature (Vovk, 2012; Park et al., 2020; Jung et al., 2023). Finally, at the end of the section we provide implied coverage guarantees (both marginal and conditional) for PLCP.

Experimental Results. Across four diverse datasets and tasks, we compared our method with established approaches such as Split Conformal (Lei et al., 2016; Papadopoulos et al., 2002), CQR (Romano et al., 2019), LocalCP (Hore & Barber, 2023), Batch-GCP (Jung et al., 2023), and Conditional Calibration (Gibbs et al., 2023). PLCP consistently outperformed Split Conformal in terms of conditional coverage and interval length. Unlike BatchGCP, which relies on predefined groups (e.g., race, gender), our method requires no such prior knowledge. In the experiments, PLCP matched BatchGCP's performance on known groups and effectively identified and covered additional meaningful groups. Compared to Conditional Calibration, PLCP achieved comparable coverage but with notably shorter prediction intervals. This success is attributed to PLCP's effective integration of advanced machine learning models to extract relevant features for uncertainty quantification.

1.1. Related Work

We begin by highlighting some of the key developments in obtaining conditional guarantees.

Group-Conditional Methods. These methods consider a *predefined* collection of groups $\mathcal{G}$ within the covariate space and guarantee coverage conditioned on each group $g \in G$ (Jung et al., 2023; Barber et al., 2019; Vovk et al., 2003; Javanmard et al., 2022). Our approach departs significantly from such methods as it *learns* a partitioning of the covariate space directly from the calibration data, not pre-establishing them, in order to identify diverse behaviors in $\Pr[Y|X]$. This distinction is crucial in cases with limited knowledge of the uncertainty of the pre-trained predictive model on varying subpopulations.

Covariate Shifts. An alternative approach to relax (2) is to provide conditional guarantees with respect to a pre-defined family of covariate shifts (Gibbs et al., 2023; Tibshirani et al., 2019; Cauchois et al., 2023). The set of covariate shifts are fixed prior to the calibration phase, and hence they may not capture covariate shifts tied to the uncertainty of the pre-trained predictive model. Gibbs et al. (2023) aims to mitigate this by expanding the covariate shift space, but this can lead to conservatism, creating unnecessarily large prediction sets, as pointed out by Hore & Barber (2023). In contrast, our method learns the uncertainty patterns of the predictive model during calibration, leading to smaller and more precise prediction sets as demonstrated in Section 4. We provide a detailed comparison with Conditional Calibration (Gibbs et al., 2023) in Remark B.4, in the Appendix.

Designing Better Conformity Scores. An alternative approach in the conformal inference literature to improve conditional validity is to design new conformity scores (Lei et al., 2016; Romano et al., 2019; Chernozhukov et al., 2021; Deutschmann et al., 2023; Feldman et al., 2021; Romano et al., 2020b). These scores aim to better encapsulate the complexities inherent in $\Pr[Y|X]$ in different applications. In contrast, in the pipeline of conformal prediction, our method applies after selecting the conformity score. This would enable us to capture more sophisticated uncertainty patterns of the pre-trained predictive model. Nonetheless, score-selection methods can potentially be integrated into our framework.

Further Works. Guan (2021); Hore & Barber (2023) aim at relaxing (2) using a predefined pairwise similarity function on the covariate space. The choice of these functions is generally heuristic and rather less dependent on the data. Moreover, Izbicki et al. (2020); LeRoy & Zhao (2021); Izbicki et al. (2022); Amoukou & Brunel (2023) aim to estimate the distribution of scores using calibration data. Such methods can become inapplicable in modern high-dimensional datasets as accurate estimation of the whole conditional distribution of scores is not possible from calibration data. In contrast to all these works, we aim at systematically learning low-dimensional features from the calibration data that characterizes the amount of uncertainty in the labels. In section 4, we showcase PLCP's ability to success at high dimensional datasets including image data. On a parallel thread, other works also looked at class-conditional (conditioned on Y) coverage guarantees (Ding et al., 2024; Si et al., 2023). While both problems, i.e. class conditional coverage and covariate conditional coverage, are pertinent to the field of conformal prediction, they represent distinct challenges and have developed along separate trajectories.

2. Algorithm

Preliminaries and Notations. Recall that given a calibration set $\{(X_i, Y_i)\}_{i=1}^n$ consisting of i.i.d. samples from a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, and a (pre-trained) model f, conformal prediction aims to construct for every input $x \in \mathcal{X}$ a prediction set $C(x)$ that is guaranteed to cover the true label y with high probability. These sets are often constructed using a given conformity score $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. For example, one commonly used conformity score for regression tasks is $S(x, y) = |y - f(x)|$. Define the random variable $S = S(X, Y)$, where $(X, Y) \sim \mathcal{D}$. To keep the notation simple, we use the notation $\mathcal{D}$ to also refer to the distribution of (X, S). Furthermore, we use $\mathcal{D}_S$ to denote the marginal distribution of S, and $\mathcal{D}_{S|x}$ to denote the conditional distribution of S given an input $x \in \mathcal{X}$. Denoting $S_i = S(X_i, Y_i)$, note that $\{(X_i, S_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}$.

One standard approach to conformal prediction is the so-called Split Conformal method (SC), which lets

$$C_{\text{SC}}(x) = \{y \in \mathcal{Y} | S(x, y) \leq q_{\text{SC}}\},$$

where q_{SC} is the $\lceil (1-\alpha)n \rceil$-th largest element in the set $\{S_1, S_2, \cdots, S_n, \infty\}$. The quantity q_{SC} can be considered as an estimate of the $(1-\alpha)$-quantile of the marginal distribution of S, i.e. $\mathcal{D}_S$. In this way, the prediction sets are mainly constructed based on the distribution $\mathcal{D}_S$, and the information from the covariates X_i is not used directly. As a consequence, the Split Conformal method (and its variants) can only provide *marginal* coverage guarantees.

In what follows, we will use the pinball loss defined as

$$\ell_\alpha(q, s) = \begin{cases} \alpha(q - s) & \text{if } q \geq s, \\ (1-\alpha)(s - q) & \text{if } q < s. \end{cases} \tag{4}$$

It is well known that, for random variable $S \sim \mathcal{D}_S$, minimizing the expected pinball loss over q yields the $(1-\alpha)$-quantile of the distribution, which we denote by $q_{1-\alpha}(S)$. I.e. $q_{1-\alpha}(S) = \underset{q \in \mathbb{R}}{\text{argmin}}\ \mathbb{E}_{S \sim \mathcal{P}}\, \ell_\alpha(q, S)$.

We will use bold symbols like $\boldsymbol{q}$ for vectors. For any $\boldsymbol{q} = (q_1, q_2, \cdots, q_m) \in \mathbb{R}^m$ and probability vector $\boldsymbol{p} = (p_1, \cdots, p_m)$, we will use the notation $\boldsymbol{q}_{i \sim \boldsymbol{p}}$, to point to the random variable that takes the value q_i with probability p_i.

Algorithmic Principles. As mentioned in Sec. 1, at a high level, our main goal is to learn a partitioning of the covariate space $\mathcal{X}$, which is representative of the uncertainty structure of the label, and construct the prediction sets accordingly. For instance, assume that we would like to partition $\mathcal{X}$ into two groups. Intuitively speaking, we would like one of the groups to include points x such that $\mathcal{D}_{Y|x}$ is "more" noisy, and the other group to include points x such that $\mathcal{D}_{Y|x}$ is "less" noisy. For more noisy x's, the distribution $\mathcal{D}_{S|x}$ takes higher values, and hence its $(1-\alpha)$-quantile is high. And for the less noisy x's, the $(1-\alpha)$-quantile of $\mathcal{D}_{S|x}$ is low.

To better illustrate the algorithmic principles, let us first assume that we have infinite data. We will shortly focus on the finite-size setting for which our algorithm is designed. We would like to partition the covariate space $\mathcal{X}$ into m groups $G_1, \cdots, G_m$. Fix a group G_i. The prediction set for each $x \in G_i$ will take the same form, i.e. $S(x, y) \leq q_i$. Hence, our goal will be to find both the groups G_i and the values q_i (i.e. the prediction sets).

Let us first see how q_i's could be found depending on the choice of G_i's. In order to guarantee coverage conditioned on the group G_i, q_i can simply be chosen as the $(1-\alpha)$-quantile of the distribution of S conditioned on G_i, i.e.

$$q_i = q_{1-\alpha}(S | X \in G_i). \tag{5}$$

Thus, fixing the groups $G_1, \cdots, G_m$, one can compute the values q_i as the corresponding quantiles and construct the prediction sets:

$$C(x) = \begin{cases} \{y \in \mathcal{Y} | S(x, y) \leq q_1\} & \text{if } x \in G_1, \\ \vdots & \vdots \\ \{y \in \mathcal{Y} | S(x, y) \leq q_m\} & \text{if } x \in G_m. \end{cases} \tag{6}$$

Now, assume that the values $q_1, \cdots, q_m$ are given and we would like to find the groups G_i accordingly. Ideally, we want to find the groups G_i in a way that if $x \in G_i$, then the $(1-\alpha)$-quantile of the conditional distribution $\mathcal{D}_{S|x}$, i.e. $q_{1-\alpha}(S | X = x)$, is very close to q_i. But this may not be possible given the specific values of q_i's. Instead, our key insight is to assign x to the group G_i whose associated value q_i is closest to $q_{1-\alpha}(S | X = x)$. To quantify closeness, note that $q_{1-\alpha}(S | X = x)$ is written using the pinball loss (4):

$$q_{1-\alpha}(S | X = x) = \underset{q \in \mathbb{R}}{\text{argmin}}\ \mathbb{E}_{S|X=x}\, \ell_\alpha(q, S) \quad \forall x \in \mathcal{X}.$$

Accordingly, the quantity $\mathbb{E}_{S|X=x}\, \ell_\alpha(q, S)$ measures how close a value q is to the quantile $q_{1-\alpha}(S | X = x)$. This measure can also be interpreted in terms of conditional coverage. Note that $\frac{d}{dq}\, \mathbb{E}_{S|X=x}\, [\ell_\alpha(q, S)] = \Pr\{S(X, Y) \leq q | X = x\} - (1-\alpha)$. As a result, $\mathbb{E}_{S|X=x}\, [\ell_\alpha(q, S)]$ can be considered as a measure of miscoverage for the prediction set $S(x, y) \leq q$. Using this measure, the closest point q_{i^*} to $q_{1-\alpha}(S | X = x)$ can be found as

$$i^* = \underset{i \in [1, \cdots, m]}{\text{argmin}}\ \mathbb{E}_{S|X=x}\, \ell_\alpha(q_i, S). \tag{7}$$

In summary, we have identified two algorithmic principles to derive the groups G_i and the values q_i (i.e. the prediction sets). Given the groups G_i, choose q_i using (5); and given the q_i's, for any $x \in \mathcal{X}$ assign its group according to (7).

The Algorithm. We will now proceed with implementing principles (5) and (7) in the finite-size setting. One way

to do this is to derive an equivalent optimization objective which admits an unbiased estimate using finite samples.

Consider the following optimization problem over the variable $\boldsymbol{q} = (q_1, q_2, \cdots, q_m) \in \mathbb{R}^m$:

$$\boldsymbol{q}^\infty = \underset{\boldsymbol{q}=(q_1,\cdots,q_m)\in\mathbb{R}^m}{\operatorname{argmin}} \; \mathbb{E}_X \left[\min_{i\in[1,\cdots,m]} \mathbb{E}_{S|X=x} \, \ell_\alpha(q_i, S) \right]. \tag{8}$$

The above optimization problem is convex as the pinball loss is convex. Further, by defining G_i^∞ according to (7):

$$G_i^\infty = \left\{ x \in \mathcal{X} \text{ s.t. } i = \underset{t\in[1,\cdots,m]}{\operatorname{argmin}} \; \mathbb{E}_{S|X=x} \, \ell_\alpha(q_t^\infty, S) \right\}, \tag{9}$$

it is easy to see that the pair $\boldsymbol{q}^\infty$ and $\{G_i^\infty\}_{i=1}^m$ satisfy both (5) and (7). Hence, we will focus on (8). Let us rewrite (8) in a slightly different but equivalent form. Let Δ_m be the m-dimensional simplex; i.e. the set of all the probability vectors $\boldsymbol{p} = (p_1, \cdots, p_m) \in \mathbb{R}^m$. We have

$$\min_{i\in[1,\cdots,m]} \mathbb{E}_{S|X=x} \, \ell_\alpha(q_i, S) = \min_{\boldsymbol{p}\in\Delta_m} \sum_{i=1}^m p_i \, \mathbb{E}_{S|X=x} \, \ell_\alpha(q_i, S).$$

Using the above relation, we can rewrite (8) as:

$$\begin{aligned}
\boldsymbol{q}^\infty &= \underset{\boldsymbol{q}\in\mathbb{R}^m}{\operatorname{argmin}} \; \mathbb{E}_X \min_{\boldsymbol{p}\in\Delta_m} \sum_{i=1}^m p_i \, \mathbb{E}_{S|X=x} \, \ell_\alpha(q_i, S) \\
&= \underset{\substack{\boldsymbol{q}\in\mathbb{R}^m \\ h:\mathcal{X}\to\Delta_m}}{\operatorname{argmin}} \; \mathbb{E}_X \sum_{i=1}^m h^i(x) \, \mathbb{E}_{S|X=x} \, \ell_\alpha(q_i, S) \\
&= \underset{\substack{\boldsymbol{q}\in\mathbb{R}^m \\ h:\mathcal{X}\to\Delta_m}}{\operatorname{argmin}} \; \mathbb{E}_{(X,S)} \sum_{i=1}^m h^i(x) \ell_\alpha(q_i, S).
\end{aligned}$$

Here, in the second and third step the minimization is over $\boldsymbol{q}$ and all the functions $h = (h^1, \cdots, h^m) : \mathcal{X} \to \Delta_m$. The second equality follows from the fact that the function

$$\boldsymbol{p}^\infty(x) = \underset{\boldsymbol{p}\in\Delta_m}{\operatorname{argmin}} \sum_{i=1}^m p_i \, \mathbb{E}_{S|X=x} \, \ell_\alpha(q_i, S),$$

is a mapping from $\mathcal{X}$ to Δ_m. Our optimization problem can thus be written as

$$h^\infty, \boldsymbol{q}^\infty = \underset{\substack{\boldsymbol{q}\in\mathbb{R}^m \\ h:\mathcal{X}\to\Delta_m}}{\operatorname{argmin}} \; \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}} \left[\sum_{i=1}^m h^i(X) \ell_\alpha(q_i, S) \right]. \tag{10}$$

With finite-size data, we perform the following two common relaxations on the above objective: (i) Replace the objective with its empirical version (i.e. replace the expectation with the sum over the calibration data); (ii) Instead of optimizing over all the functions $h : \mathcal{X} \to \Delta_m$, which clearly overfits in the finite-size setting, we optimize over a function class $\mathcal{H}$. E.g., $\mathcal{H}$ could be the class of linear functions or neural networks with a soft-max layer at the output. Our final optimization problem, using the finite-size calibration set $\{(X_i, S_i)\}_{i=1}^n$ and function class $\mathcal{H}$, becomes:

$$h^*, \boldsymbol{q}^* = \underset{\substack{\boldsymbol{q}\in\mathbb{R}^m \\ h\in\mathcal{H}}}{\operatorname{argmin}} \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^m h^i(X_j) \ell_\alpha(q_i, S_j). \tag{11}$$

Partition Learned Conformal Prediction (PLCP) algorithm is formulated utilizing (11), in Algorithm 1.

Algorithm 1 Partition Learned Conformal Prediction (PLCP)

Require: Data: $\{(X_i, Y_i)\}_{i=1}^n$, Conformity score: $S(x, y)$, Number of groups: m, Family of functions: $\mathcal{H}$

1: Compute $S_i = S(X_i, Y_i), \quad \forall i \in [1, \cdots, n]$.

2: Solve the optimization problem,

$$h^*, \boldsymbol{q}^* = \underset{\substack{\boldsymbol{q}\in\mathbb{R}^m \\ h\in\mathcal{H}}}{\operatorname{argmin}} \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^m h^i(X_j) \ell_\alpha(q_i, S_j).$$

Ensure: Prediction set $C^*(x) = \{y \mid S(x, y) \le \boldsymbol{q}^*_{i\sim h^*(x)}\}$

Remark 2.1. Oftentimes in practice, the machine learning models, such as Neural Networks, are parametric (h_θ, where θ is the set of parameters). In that case, we can simply implement PLCP with alternating gradient descent, performing few steps of gradient descent on q and θ at each iteration.

3. Theoretical Results

We introduced in (3) a new notion to measure conditional coverage called the mean squared coverage error (MSCE). In words, MSCE penalizes the deviation of the coverage of prediction sets, conditioned on each point x, from the nominal value $1 - \alpha$. Further, as we will show, any bound on the MSCE can be simply translated into a bound on coverage. As a result, minimizing MSCE can be a valid objective as a relaxation of (2). In what follows, we provide guarantees for the MSCE of PLCP in the presence of finite and infinite data. Fallback coverage guarantees will also be provided at the end of this section. All the proofs has been moved to the Appendix E. First, let us introduce our main assumptions.

Assumption 3.1. The set $\{(X_i, Y_i)\}_{i=1}^n$ are generated i.i.d.

Assumption 3.2. The conformity score $s(.,.)$ should be bounded. Without loss of generality, we can assume $s(\cdot,\cdot) \in [0, 1]$.

Definition 3.3. A distribution $\mathcal{P}$, is called L-lipschitz if we have for every real valued numbers $q \le q'$,

$$\Pr_{X\sim\mathcal{P}}(X \le q') - \Pr_{X\sim\mathcal{P}}(X \le q) \le L\,(q' - q)\,.$$

Assumption 3.4. The conditional distribution $\mathcal{D}_{S|X}$ should be L-Lipschitz, almost surely with respect to $\mathcal{D}_X$.

Assumption 3.1 is often necessary to obtain concentration guarantees and has been used in the literature of of conformal prediction (Sesia & Romano, 2021; Jung et al., 2023; Lei & Wasserman, 2012; Guan, 2021). These works have also considered regularity assumptions similar to or stronger than 3.4. Also, Assumption 3.2 has been used in the same references. Note that the above assumptions do not compromise the distribution-free nature of conformal prediction. Given that obtaining distribution-free guarantees as in (2) is impossible (Vovk, 2012; Foygel Barber et al., 2019), the adoption of regularity conditions to extend beyond mere marginal coverage guarantees is a well-established path.

3.1. Infinite data

At the heart of our analysis, we have Proposition 3.5 that connects the pinball loss to the MSCE of prediction sets. To motivate this Proposition, let us look at the following prediction set created by the true quantile function, $q_{1-\alpha}(S|X=x)$,

$$C_{\text{opt}}(x) = \{y \in \mathcal{Y} \mid S(X,Y) \leq q_{1-\alpha}(S|X=x)\}.$$

This prediction set is optimal as it guarantees full conditional coverage (2). The true quantile function can also be expressed as the minimizer of the pinball loss; i.e. defining the function $q_{1-\alpha}(x) := q_{1-\alpha}(S|X=x)$, we have

$$q_{1-\alpha}(\cdot) \in \underset{f:\mathcal{X}\to\mathbb{R}}{\text{argmin}}\ \mathbb{E}_{(X,S)\sim\mathcal{D}}\, \ell_\alpha(f(X),S). \tag{12}$$

This suggests that minimizing pinball loss can potentially lead to prediction sets with a better conditional coverage behaviour – an intuition used in prior work, such as (Jung et al., 2023; Gibbs et al., 2023), to obtain conditional guarantees using the pinball loss. In the following proposition, we formalize this intuition, by bounding the MSCE of prediction sets constructed by an arbitrary function $g(x) : \mathcal{X} \to \mathbb{R}$.

Proposition 3.5. Under assumption 3.4, for every function $g(x) : \mathcal{X} \to \mathbb{R}$, we have

$$\text{MSCE}(C_g) \leq 2L\,\mathbb{E}\left[\ell_\alpha(g(X),S) - \ell_\alpha(q_{1-\alpha}(X),S)\right],$$

$$\text{where:}\quad C_g(x) = \{y \in \mathcal{Y} | S(x,y) \leq g(x)\}.$$

In light of Proposition 3.5, we proceed with analyzing the coverage of the prediction sets created by PLCP. In the first step, we look at the case that we have infinitely many data, and the function class $\mathcal{H} = \Delta_m^{\mathcal{X}}$. Here $\Delta_m^{\mathcal{X}}$ denotes all the possible functions from $\mathcal{X}$ to simplex Δ_m. In this case we show that the performance of function $h^\infty(x)$ defined in (10), in terms of expected pinball loss, is different from the optimal quantile function $q_{1-\alpha}(S|X=x)$ by an additive error $O(1/\sqrt{m})$.

Theorem 3.6. *For any distribution $\mathcal{D}$ and number of groups m, we have the following bound,*

$$\left|\min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m}\mathbb{E}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^i(X)\ell_\alpha(q_i,S)\right] - \mathbb{E}\left[\ell_\alpha(q_{1-\alpha}(X),S)\right]\right| \leq 2\sqrt{\frac{\text{var}(q_{1-\alpha}(X))}{m-1}}, \tag{13}$$

where $\text{var}(\cdot)$ *denotes the usual variance.*

An immediate implication of Theorem 3.6 is the MSCE property of the prediction sets of PLCP. Recall the definitions of $h^\infty, \boldsymbol{q}^\infty$ as in (10). By using the results of Proposition 3.5 and Theorem 3.6, we obtain the following corollary.

Corollary 3.7. *Under Assumption 3.4, the prediction sets $C_\infty = \{y \in \mathcal{Y} | S(x,y) \leq \boldsymbol{q}^\infty_{i\sim h^\infty(x)}\}$, satisfies the following conditional coverage guarantee*

$$\text{MSCE}(C_\infty) \leq 4L\sqrt{\frac{\text{var}(q_{1-\alpha}(X))}{m-1}},$$

where $\text{var}(\cdot)$ *denotes the variance.*

Corollary 3.7 indicates the effect of the number of groups, m, on the MSCE in the infinite-sample (population) regime. For further discussion on this result, please see Remarks B.1, B.2 in the appendix.

3.2. Finite data

Next, we turn into the finite-size setting and analyze the performance of prediction sets provided by PLCP. Naturally, we should expect that the complexity of the function class $\mathcal{H}$ used by PLCP to play a role. In this context, the following statement captures the complexity of a function class using the well-known notion of covering number.

Definition 3.8. (ε-net). Let (T,d) be a metric space. Consider a subset $K \subset T$ and let $\varepsilon > 0$. A subset $\mathcal{N} \subseteq K$ is called an ε-net of K if every point in K is within distance ε of some point of $\mathcal{N}$, i.e. $\forall x \in K \quad \exists y \in \mathcal{N} : d(x,y) \leq \varepsilon$.

Definition 3.9. (Covering numbers). The smallest possible cardinality of an ε-net of K is called the covering number of K and is denoted $\mathcal{N}(K,d,\varepsilon)$.

Proposition 3.10. The following function $d : \Delta_m^{\mathcal{X}} \times \Delta_m^{\mathcal{X}} \to \mathbb{R}^+ \cup \{0\}$ is a metric over $\Delta_m^{\mathcal{X}}$:

$$d(h_1,h_2) = \sup_{x\in\mathcal{X}} \sum_{i=1}^m \left|h_1^i(x) - h_2^i(x)\right|,$$

We will see that $\mathcal{N}(\mathcal{H},d,\frac{1}{n})$ will play an important role in our finite sample analysis. Assuming that a class of functions has "bounded complexity", we would expect a trade-off between m, the number of groups, and n, the

number of i.i.d. samples. On the one hand, corollary 3.12 suggests that increasing m should benefit the conditional coverage of PLCP; and on the other hand, increasing m in the finite sample regime would lead to a smaller number of samples per each group, that can hurt the coverage property of PLCP prediction sets. The next Theorem characterizes this trade-off precisely. Before presenting the theorem we need to define the approximation gap $\lambda_{\mathcal{H}}$ of a class $\mathcal{H}$ as

$$\lambda_{\mathcal{H}} = \min_{h\in\mathcal{H},\boldsymbol{q}\in\mathbb{R}^m} \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}} \left[\sum_{i=1}^{n} h^i(X)\ell_\alpha(q_i,S)\right] - \min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m} \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}} \left[\sum_{i=1}^{m} h^i(X)\ell_\alpha(q_i,S)\right]. \tag{14}$$

In words, the approximation gap $\lambda_{\mathcal{H}}$ is the error we suffer due to restricting to the function class $\mathcal{H}$. Essentially, when $\lambda_{\mathcal{H}} = 0$ (a.k.a. the minimizer is achieved by the class $\mathcal{H}$) then this becomes an analogous to the "realizable case" terminology that exists in the learning theory literature.

Theorem 3.11. *Under assumptions 3.1 and 3.2 we have, with probability* $1-\delta$*,*

$$\left|\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m} h^{*i}(X)\ell_\alpha(q_i^*,S)\right] - \mathbb{E}\left[l_\alpha(q_{1-\alpha}(X),S)\right]\right| \leq 10\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)+\ln\left(\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)\right)+\ln\left(\mathcal{N}(\mathcal{H},d,\frac{1}{n})\right)}{n}} + 2\sqrt{\frac{var(q_{1-\alpha}(X))}{m}} + \lambda_{\mathcal{H}},$$

where $h^*, \boldsymbol{q}^*$ *are defined in algorithm 1.*

Applying Proposition 3.5 to Theorem 3.11 leads to the following corollary, which is our main theoretical result on the finite-size behavior of the predictions sets obtained by PLCP.

Corollary 3.12. *Under assumptions 3.1, 3.4, and 3.2 we have, with probability* $1-\delta$*,*

$$\mathrm{MSCE}(C^*) \leq 20L\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)+m\ln(n)+\ln\left(\mathcal{N}(\mathcal{H},d,\frac{1}{n})\right)}{n}} + 4L\sqrt{\frac{var(q_{1-\alpha}(X))}{m}} + 2L\lambda_{\mathcal{H}},$$

where C^* *is the prediction sets constructed by PLCP.*

For further discussion, see Remark B.3 in the appendix.

3.3. Fallback Coverage Guarantees

The MSCE bounds provided in Corollaries 3.7, 3.12, can be simply translated to fallback coverage guarantees.

Given prediction sets $C(x), x\in\mathcal{X}$, assume that

$$\mathrm{MSCE} := E\left[(\mathrm{Cov}(X)-(1-\alpha))^2\right] \leq p,$$

for some $p \geq 0$. Using Jensen's inequality, we have

$$1-\alpha-\sqrt{p} \leq E[\mathrm{Cov}(X)] = \Pr(Y\in C(X)) \leq 1-\alpha+\sqrt{p}$$

This means any bound on MSCE gives a bound on the *marginal* coverage of the prediction sets.

Furthermore, for *any set* $A\subseteq\mathcal{X}$ such that $P[A]\geq\delta$, we have,

$$1-\alpha-\sqrt{\frac{p}{\delta}} \leq E[\mathrm{Cov}(X)\mid X\in A] = \Pr(Y\in C(X)\mid X\in A) \leq 1-\alpha+\sqrt{\frac{p}{\delta}}$$

This indicates that a bound on MSCE can be translated to *conditional* coverage guarantees.

We can now use the results of Corollaries 3.7, 3.12 to obtain specific coverage guarantees for PCLP. These fallback coverage guarantees are provided in Corollary C.1 in the appendix.

4. Experimental Results

We have conducted extensive numerical experiments to evaluate the performance of PLCP in terms of two key metrics: the coverage probability and the average length of the prediction intervals. These metrics are evaluated with respect to specific groups within the test data. For a given group of interest G, conditional coverage and length are examined as quantities: $\Pr[Y\in C(X)\mid X\in G]$ and $\mathbb{E}[\mathrm{length}(C(X))\mid X\in G]$, where $C(X)$ denotes the prediction set for a covariate point X. The group G may represent a subpopulation within the data or a set of covariates that have undergone shifts relative to the calibration set. These shifts include features that are both Out of Distribution (OOD) and In Distribution (ID) w.r.t. the training set utilized for the predictor. We compare PLCP against three baselines: (i) the Split Conformal method (Papadopoulos et al., 2002; Lei et al., 2016); (ii) the BatchGCP method (Jung et al., 2023); and (iii) the Conditional Calibration method (Gibbs et al., 2023). In all experiments, we set the miscoverage rate, α, to 0.1. For a discussion on how to tune m, the number of regions in PLCP, see Remark B.5 in the Appendix.

We have provided further experimental results in Appendix A by comparing PLCP with several other baselines such as the method of Conformalized Quantile Regression (CQR) developed in (Romano et al., 2019) as well as the LocalCP method (Hore & Barber, 2023). These baselines are selected due to their emphasis on calibrated prediction sets that cater to conditional coverage. In Appendix A, we also look at two metrics recommended by (Feldman et al., 2021)(Feldman et al., 2021), beyond coverage and set size plots.

4.1. Comparison with Split Conformal

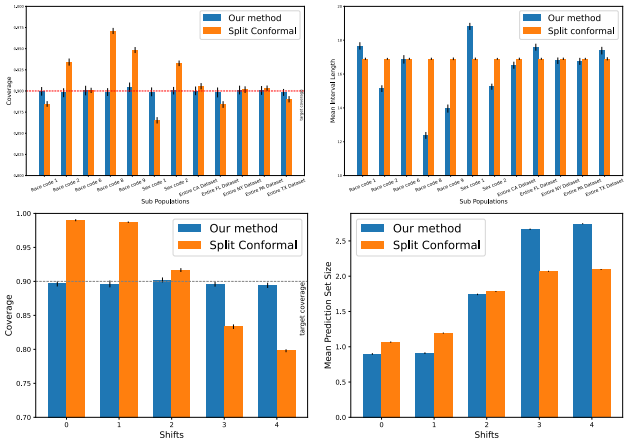

Figure 2: Left-hand-side plots show coverage and right-hand-side plots show mean prediction set size. Row 1: US Census Data; Row 2: MNIST with Gaussian Blur.

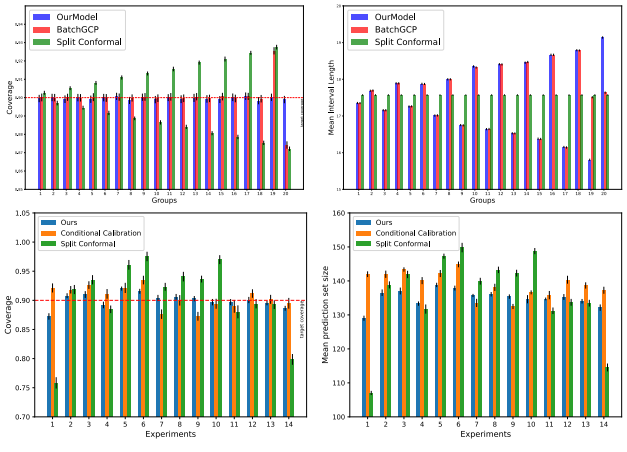

Figure 3: Left-hand-side plots show coverage and right-hand-side plots show mean prediction set size. Row 1: Synthetic Regression Task; Row 2: RxRx1 WILDS Dataset.

Here we compare PLCP with the Split Conformal method. While there are other methods with marginal guarantees, such as Full Conformal Prediction (Vovk et al., 2005) and Jackknife+ (Barber et al., 2021), our choice of Split Conformal is driven by its implementation efficiency and the fact that its conditional coverage performance is similar to those alternatives. The experiments are conducted in two scenarios depending on calibration data being In Distribution (ID) or Out of Distribution (OOD) relative to the training data.

2018 US Census Data (In-Distribution). We study the 2018 US Census Data from the Folktables library (Ding et al., 2021) for income prediction, a dataset rich in demographic details like gender, race, and age. Focusing on the five most populated US states (CA, FL, NY, PA, TX), we aim to model diverse demographic and geographical subpopulations. The objective is to assess PLCP's performance in scenarios where the train, calibration, and test sets are In Distribution (ID). Data are divided into three segments: 60% for training, 20% for calibration, and 20% for testing. We use a linear regression model as the predictor $f(x)$, with conformity measured by $S(x, y) = |f(x) - y|$. PLCP is implemented using a two-layer ReLU neural network (200 and 100 neurons). Performance evaluation across various racial, gender, and state-wise subpopulations is shown in Figure 2. PLCP achieves near-perfect coverage across all subpopulations, which showcases PLCP's ability to learn and leverage features that are informative w.r.t. the uncertainty inherent in the predictor $f(x)$.

MNIST Data (Out-of-Distribution). We divide the MNIST dataset into 35,000 training images and 25,000 for calibration/testing. The calibration/test data is further split into five subgroups each subjected to different Gaussian blur levels. Such added blurriness creates an OOD scenario compared to the training data. These 25,000 blurred images are then randomly divided into a 15,000-image calibration set and a 10,000-image test set. The conformity score for each input image x is given as $S(x, y) = 1 - \pi^y(x)$, where $\{\pi^i(x)\}_{i=1}^{10}$ denote the softmax probabilities of class membership output by the trained predictive model. PLCP is implemented with $m = 8$ to denote eight distinct groups, employing a Convolutional Neural Network (CNN) architecture with three convolution layers and two feed-forward layers. The results are given in Figure 2. PLCP demonstrates near-perfect coverage across all blur levels, effectively adjusting to the OOD conditions.

4.2. Comparison with Methods that Provide Conditional Guarantees with Respect to Predefined Structures

Group-Conditional Methods. In this section, we used a synthetic regression task, proposed originally in (Jung et al., 2023), to compare PLCP with BatchGCP. The data distribution is given as follows: The covariate $X = (X_1, \cdots, X_{100})$ is a vector in $\mathbb{R}^{100}$. The values in the first ten coordinates of X are independent uniform binary variables, and the remaining 90 coordinates are i.i.d. Gaussian variables with mean of zero and variance of σ_x. The label y is then generated as follows: $y = \langle \theta, X \rangle + \epsilon_X$ where $\epsilon_X \sim \mathcal{N}\left(0, \sigma_x^2 + \sum_{i=1}^{d} X_i \cdot \sigma_i^2\right)$, and $\sigma_i^2 = i$. We generate from this distribution 150K training samples (to train the regression model to predict the label), 50K calibration data points, and 50K test data points. We evaluate all algorithms over 100 independent trials and report average performance.

We define 20 overlapping groups based on the first ten binary components of X. Specifically, for each i in the range 1 to 10, Group $2i - 1$ corresponds to $X_i = 0$ and Group $2i$ to

$X_i = 1$. BatchGCP is implemented given the knowledge of the first 18 groups (associated with the first nine binary digits), however, no information about the group structures or the data distribution is provided to PLCP. We ran PLCP with $m = 25$ (25 groups), using a linear classifier (as $\mathcal{H}$).

As seen in Figure 3, the Split Conformal method results in over-coverage for groups with lower variance and under-coverage for those with higher variance. Notably, PLCP achieves coverage performance comparable to BatchGCP for the first 18 groups. However, for groups 19 and 20, whose information was not given to BatchGCP, PLCP exhibits superior coverage. As illustrated in the Mean Interval Length plot of Figure 3, PLCP adaptively modifies the interval lengths for these two groups, recognizing the importance of X_{10}, a feature learned from the data. This showcases the power of PLCP in learning features/boundaries in covariate space that are informative about the uncertainty inside the predictions of the pre-trained model.

Methods Based on a Family of Covariate Shifts. Our last experiment is on the RxRx1 dataset (Taylor et al., 2019) from the WILDS repository (Koh et al., 2021), a dataset previously used in (Gibbs et al., 2023). This repository includes benchmark datasets particularly designed for evaluating performance under distribution shifts. The RxRx1 dataset comprises images of cells, where the task is to predict one of 1339 genetic treatments applied to these cells, spanning 51 distinct experiments. The inherent execution and environmental variations introduce covariate shifts among images from different experiments. One main challenge in this task is the lack of explicit features indicating experiment origin, which complicates the direct use of methods that rely on predefined groups or covariate shifts for coverage. We compare PLCP with the Conditional Calibration algorithm (Gibbs et al., 2023). The Conditional Calibration approach bifurcates the calibration data: first, it uses ℓ_2-regularized multinomial linear regression to estimate the likelihood of each image's experimental origin, then applies these probabilities to define covariate shifts for the second part of the data. In contrast, PLCP inherently adapts to this task by directly learning features important for uncertainty quantification using the entire calibration set. For this experiment, we ran PLCP using a CNN with a single convolution layer, with ReLU activation followed by a linear layer, configured with $m = 20$ groups.

For genetic treatment prediction (the predictive model), we employ a ResNet50 architecture, $f(x)$, pre-trained on 37 experiments from the WILDS repository. The remaining images from the 14 experiments are divided into calibration and test sets uniformly at random. The conformity scores are computed as follows: for each image x, let $\{f^i(x)\}_{i=1}^{1339}$ represent the weights assigned by $f(x)$ to the 1339 treatments. Temperature Scaling, followed by a softmax operation, is applied to derive the probability weights $\pi^i(x) := \exp(Tf_i(x))/(\sum_j \exp(Tf_j(x)))$, with T as the temperature parameter. We let $S(x, y) := \sum_{i:\pi_i(x)>\pi_y} \pi_i(x)$. Figure 3 provides a comparative evaluation of PLCP, Conditional Calibration, and Split Conformal. PLCP exhibits a superior performance in terms of set size while matching the coverage performance of Conditional Calibration. This performance, combined with the principled approach of our method, showcases its advantage for applications where identifying the correct uncertainty structures from the covariates is not straightforward.

Acknowledgements

Shayan Kiyani, Hamed Hassani, and George J. Pappas are supported by the NSF Institute for CORE Emerging Methods in Data Science (EnCORE).

Impact Statement

In this paper, we focus on developing a new algorithmic framework for Conformal Prediction which has immediate applications in areas such as healthcare. We do not anticipate any negative societal impact.

References

Amoukou, S. I. and Brunel, N. J. Adaptive conformal prediction by reweighting nonconformity score. *arXiv preprint arXiv:2303.12695*, 2023.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 2019. URL https://api.semanticscholar.org/CorpusID:88524668.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. 2021.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, (just-accepted):1–22, 2023.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Deutschmann, N., Rigotti, M., and Martinez, M. R. Adaptive conformal regression with jackknife+ rescaled scores. *arXiv preprint arXiv:2305.19901*, 2023.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

Feldman, S., Bates, S., and Romano, Y. Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems*, 34: 2060–2071, 2021.

Foygel Barber, R., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *arXiv e-prints*, art. arXiv:1903.04684, March 2019. doi: 10.48550/arXiv.1903.04684.

Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees, 2023.

Guan, L. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 2021. URL https://api.semanticscholar.org/CorpusID:235446816.

Hore, R. and Barber, R. F. Conformal prediction with local weights: randomization enables local guarantees. *arXiv preprint arXiv:2310.07850*, 2023.

Izbicki, R., Shimizu, G., and Stern, R. Flexible distribution-free conditional predictive bands using density estimators. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3068–3077. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/izbicki20a.html.

Izbicki, R., Shimizu, G., and Stern, R. B. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *The Journal of Machine Learning Research*, 23(1):3772–3803, 2022.

Javanmard, A., Shao, S., and Bien, J. Prediction sets for high-dimensional mixture of experts models. *arXiv preprint arXiv:2210.16710*, 2022.

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Dk7QQp8jHEo.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. Wilds: A benchmark of in-the-wild distribution shifts. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 71–96, 2014.

Lei, J. and Wasserman, L. A. Distribution free prediction bands. *ArXiv*, abs/1203.5422, 2012. URL https://api.semanticscholar.org/CorpusID:11908092.

Lei, J., G'Sell, M. G., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. A. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094 – 1111, 2016. URL https://api.semanticscholar.org/CorpusID:13741419.

LeRoy, B. and Zhao, D. Md-split+: Practical local conformal inference in high dimensions. *arXiv preprint arXiv:2107.03280*, 2021.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European Conference on Machine Learning*, 2002. URL https://api.semanticscholar.org/CorpusID:42084298.

Park, S., Bastani, O., Matni, N., and Lee, I. Pac confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxVI04YvB.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2), apr 30 2020a. https://hdsr.mitpress.mit.edu/pub/qedrwcz3.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020b.

Sesia, M. and Romano, Y. Conformal prediction using conditional histograms. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:239769316.

Si, W., Park, S., Lee, I., Dobriban, E., and Bastani, O. Pac prediction sets under label shift. *arXiv preprint arXiv:2310.12964*, 2023.

Taylor, J., Earnshaw, B., Mabey, B., Victors, M., and Yosinski, J. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, volume 22, pp. 23, 2019.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. Conformal prediction under covariate shift. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:115140768.

Vovk, V. Conditional validity of inductive conformal predictors. In Hoi, S. C. H. and Buntine, W. (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vovk, V., Lindsay, D., Nouretdinov, I., and Gammerman, A. Mondrian confidence machine. *Technical Report*, 2003.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

A. Additional Experiments

In this section we add two more baselines to the experiment setups of the Section 4.1. We already compared PLCP with Split Conformal prediction, a conventional tool to achieve marginal guarantee, and similar to PLCP, it does not need to know any further structures of the data. Here we also consider and report the performance of two other algorithms, which are known to show better conditional coverage behavior than the Split Conformal solution; and they also do not need to have any prior knowledge of the structure of the data. The two methods are the method of Conformalized Quantile Regression (CQR) developed in (Romano et al., 2019) as well as the LocalCP method (Hore & Barber, 2023). In short, CQR combines conformal prediction with classical quantile regression and makes the interval length more adaptable across the input space. LocalCP uses a weighted version of Split Conformal prediction, where the weights come from a local similarity kernel (e.g. a gaussian). These methods are selected due to their emphasis on calibrated prediction sets that cater to conditional coverage. Furthermore, to make a more refined comparison, in addition to coverage and length plots, we also look at two metrics recommended by (Feldman et al., 2021); Namely, the correlation between the size of prediction intervals and the indicators of coverage, assessed using both the Hilbert-Schmidt Independence Criterion (HSIC) and the Pearson's correlation coefficient. In the tables provided (table 1, 2), the numbers represent percentages of improvement in comparison to the Split Conformal solution. This presentation mirrors the approach taken in the paper (Feldman et al., 2021) where correlations are similarly reported. In the same paper, the authors argue that for a prediction set to adhere to the full conditional coverage criterion (see (2)), its interval length and its coverage indicator must exhibit independence. Therefore, the improvement percentages detailed in the subsequent tables can be understood as the percentage reduction in the relevant correlation metric when measured against the Split Conformal solution. Tables 1, 2 and the plots in Figure 4 show the superior performance of PLCP compared to the baselines.

Table 1: Results for the 2018 US Census

	PLCP	CQR	LocalCP
HSIC	56.07	51.89	32.12
Pearson	64.06	59.21	37.96

Table 2: Results for the MNIST dataset

	PLCP	CQR	LocalCP
HSIC	73.36	58.01	40.88
Pearson	86.08	75.19	47.51

B. Remarks

Remark B.1. Theorems 3.6 and Corollary 3.7 can be interpreted through the lens of quantization, a fundamental concept in information theory and signal processing. By constraining the number of groups to m, the quantity $\boldsymbol{q}^{\infty}_{i\sim h^{\infty}(x)}$ takes only m distinct values. This scenario can be conceptualized as an m-level quantization ($\log(m)$ bits) of the signal $q_{1-\alpha}(X)$. With this perspective, Theorem 3.6 provides an upper bound on the quantization error for the optimal m-level approximation of the optimal quantile function $q_{1-\alpha}(\cdot)$. Specifically, the theorem can be restated as follows: $\log\left(\frac{4\text{var}(q_{1-\alpha}(X))}{\epsilon^2}\right)$ bits suffice to achieve a quantization error below ϵ for the signal $q_{1-\alpha}(X)$.

Remark B.2. The authors propose to explore a more comprehensive theory that bridges the gap between quantization theory and conformal prediction with conditional guarantees as a promising avenue for future research. Such a theory could potentially lead to the development of more nuanced, distribution-dependent impossibility results, thereby extending the findings of (Vovk, 2012; Foygel Barber et al., 2019).

Remark B.3. While similarities can be drawn between the finite sample theory presented in our work and the classical PAC-learning framework, there exist fundamental differences in both algorithm design and analysis, making our analysis more complex. In the context of supervised classification, the learner is provided with covariates and their corresponding labels, and the challenge lies in developing a classifier that generalizes effectively to unseen data. In a similar manner, the task in conformal prediction can be described as predicting the hypothetical label $q_{1-\alpha}(S|X=x)$ for every covariate

point x. However, examining the calibration data after applying the conformity measure, i.e. $\{(X_i, S_i)\}_{i=1}^n$, reveals that the learner has access to covariates and noisy approximations, $\{S_i\}_{i=1}^n$, of the hypothetical labels $\{q_{1-\alpha}(S|X=x)\}_{i=1}^n$. Therefore, the learner's objective is to devise a classifier that not only generalizes well to new data but also effectively navigates these noisy labels. This challenge is further compounded when considering that the random variable S do not represent an unbiased estimation of $q_{1-\alpha}(S|X=x)$. Often in practical scenarios, S tends to concentrate around the mean of its distribution, whereas $q_{1-\alpha}(S|X=x)$ corresponds to a point in the upper tail of S's probability density function. This discrepancy indicates that our methodology and theoretical analysis are tackling a scenario of greater sophistication compared to the standard supervised classification problem, encompassing not just prediction accuracy but also the subtleties of dealing with biased and noisy label information.

Remark B.4. The optimization problem considered in Gibbs et al. (2023), similar to ours, is based on minimizing a pinball loss-oriented objective over a class of functions. However, these two approaches are fundamentally different. The role of the function class in the method of Gibbs et al. (2023), as mentioned by the authors, is to learn\approximate the quantile value function ($q_{1-\alpha}(x)$ in our notation). In other words, they address a *quantile regression* problem over calibration data. On the contrary, the function class in our method is used to learn a partitioning\boundary in the covariate space, effectively tackling a *classification* problem over the calibration data. Although this boundary is related to the conditional quantile function ($q_{1-\alpha}(x)$), the relationship is considerably more complex. To illustrate, consider two scenarios in which the conditional quantile functions differ by a constant real number c (a situation that could arise if two datasets experience a constant label shift relative to each other). Then, the quantile regression function produced by the method of Gibbs et al. (2023) in these scenarios, should ideally differ by the same constant c. Conversely, the output of the function obtained by our method should ideally remain unchanged across both scenarios. This distinction stems from the fact that the model obtained by our method aims to identify regions in the covariate space whose members exhibit similar uncertainty levels (with respect to the predictions of the underlying pretrained predictive model), rather than mirroring the exact values of the conditional quantile function ($q_{1-\alpha}(x)$).

Remark B.5. In practical applications, the hyperparameter m is optimized through cross-validation. Our approach to tune m, as informed by the theoretical insights in Section 3, leverages the observed bell-shaped relationship between m and accuracy (measured by MSCE). Starting with $m = 1$, we note that the accuracy initially improves with increasing m, but after a certain point it begins to decline. To identify the optimal m, we employ the doubling trick: setting aside 20 percent of the calibration data for validation, we increment m from a small value, evaluate PLCP on the validation set, and continue doubling m until the validation metric worsens. We then fine-tune by bisecting between the last two m values. Once the optimal m is determined, we re-run PLCP on the entire calibration set using this optimized m value.

C. Additional Corollaries

Corollary C.1. ***Infinite sample****: Under Assumption 3.4, the following fallback coverage guarantees hold,*

(a) Marginal validity:

$$\left|\Pr(Y \in C_\infty(X)) - (1-\alpha)\right| \leq O\left(m^{-\frac{1}{4}}\right)$$

(b) Conditional validity: For any set $A \subseteq \mathcal{X}$ such that $\Pr(A) \geq \gamma$

$$\left|\Pr(Y \in C_\infty(X) \mid X \in A) - (1-\alpha)\right| \leq O\left(m^{-\frac{1}{4}}\gamma^{-\frac{1}{2}}\right)$$

Finite sample*: Under assumptions 3.1, 3.4, and 3.2 we have,*

(a) Marginal validity: With probability $1-\delta$,

$$\left|\Pr(Y \in C(X)) - (1-\alpha)\right| \leq O\left(\sqrt{m^{-\frac{1}{2}} + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + m\ln(n) + \ln\left(\mathcal{N}\left(\mathcal{H}, d, \frac{1}{n}\right)\right)}{n}} + \lambda_{\mathcal{H}}}\right)$$

(b) Conditional validity: For any set $A \subseteq \mathcal{X}$ such that $P[A] \geq \gamma$, with probability $1-\delta$,

$$\left|\Pr(Y \in C(X) \mid X \in A) - (1-\alpha)\right| \leq O\left(\gamma^{-\frac{1}{2}}\sqrt{m^{-\frac{1}{2}} + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + m\ln(n) + \ln\left(\mathcal{N}\left(\mathcal{H}, d, \frac{1}{n}\right)\right)}{n}} + \lambda_{\mathcal{H}}}\right)$$

D. Additional Lemmas

Theorem D.1. *(Chernoff Bound). Let $\{X_i\}_{i=1}^n$ be independent random variables bounded such that for each $i \in [n], X_i \in [0,1]$. Let $S_n = \sum_{i=1}^n X_i$ denote their sum. Then for all $\epsilon > 0$,*

$$\Pr_{\{X_i\}_{i=1}^n}\left(|S_n - \mathbb{E}[S_n]| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{n}\right)$$

Lemma D.2. *(Lipschitz Continuity of the Pinball Loss Function). The pinball loss function exhibits Lipschitz continuity with a constant of 1. This implies that for any four real numbers $y_1, y_2, y_3, y_4 \in \mathbb{R}$, the following inequality holds true:*

$$|\ell_\alpha(y_1, y_2) - \ell_\alpha(y_3, y_4)| \leq |(y_1 - y_2) - (y_3 - y_4)|.$$

Proof. Our objective is to establish that $\ell_\alpha(y_1, y_2) - \ell_\alpha(y_3, y_4) \leq |(y_1 - y_2) - (y_3 - y_4)|$. The converse inequality can be derived analogously due to symmetry. We consider the following four scenarios:

Scenario 1: When $y_1 \geq y_2$ and $y_3 \geq y_4$, we have:

$$\ell_\alpha(y_1, y_2) - \ell_\alpha(y_3, y_4) = \alpha(y_1 - y_2) - \alpha(y_3 - y_4) \leq |(y_1 - y_2) - (y_3 - y_4)|.$$

Scenario 2: For $y_1 < y_2$ and $y_3 < y_4$, it follows that:

$$\ell_\alpha(y_1, y_2) - \ell_\alpha(y_3, y_4) = (1-\alpha)(y_2 - y_1) - (1-\alpha)(y_4 - y_3) \leq |(y_1 - y_2) - (y_3 - y_4)|.$$

Scenario 3: In the case where $y_1 \geq y_2$ but $y_3 < y_4$, the equation becomes:

$$\begin{aligned}\ell_\alpha(y_1, y_2) - \ell_\alpha(y_3, y_4) &= \alpha(y_1 - y_2) - (1-\alpha)(y_4 - y_3) \\ &= \alpha(y_1 - y_2 - (y_3 - y_4)) + (y_3 - y_4) \leq |(y_1 - y_2) - (y_3 - y_4)|.\end{aligned}$$

Scenario 4: Lastly, when $y_1 < y_2$ and $y_3 \geq y_4$, we have:

$$\begin{aligned}\ell_\alpha(y_1, y_2) - \ell_\alpha(y_3, y_4) &= (1-\alpha)(y_2 - y_1) - \alpha(y_3 - y_4) \\ &= (1-\alpha)(y_2 - y_1 - (y_4 - y_3)) + (y_4 - y_3) \leq |(y_1 - y_2) - (y_3 - y_4)|.\end{aligned}$$

□

Lemma D.3. *Let B_∞^m denote the unit ball in $\mathbb{R}^m$ under the sup norm, defined by $B_\infty^m = \{x \in \mathbb{R}^m : \|x\|_\infty \leq 1\}$. For any positive real number δ, smaller than 2, the covering number $N(B_\infty^m, \delta)$ of B_∞^m can be bounded above by $\lceil \frac{2}{\delta} \rceil^m$.*

Proof. Consider a grid of cubes in $\mathbb{R}^m$, each of side length δ, where $0 < \delta < 2$. These cubes are the Cartesian product of intervals of length δ, centered at grid points in $\mathbb{R}^m$. A cube centered at a point $y = (y_1, y_2, \ldots, y_m)$ is defined as the set $\{x \in \mathbb{R}^m : |x_i - y_i| \leq \delta/2, \forall i = 1, \ldots, m\}$. The grid points form a regular lattice in $\mathbb{R}^m$, with each coordinate being an integer multiple of δ.

Any point $x \in B_\infty^m$ will lie within a distance of $\frac{\delta}{2}$ (in the sup norm) from some grid point. The unit ball under the sup norm can be inscribed within a cube of side length 2, centered at the origin. This larger cube can be covered by at most $\lceil \frac{2}{\delta} \rceil$ intervals of length δ along any axis. Therefore, in m dimensions, the total number of cubes required to cover B_∞^m is bounded above by $\lceil \frac{2}{\delta} \rceil^m$. □

E. Proofs of Section 3

Proof of Proposition 3.5. Starting from the right-hand side, we have

$$\mathbb{E}[\ell_\alpha(g(X), S) - \ell_\alpha(q_{1-\alpha}(X), S)] = \mathbb{E}_X \, \mathbb{E}_{S|X}[\ell_\alpha(g(X), S) - \ell_\alpha(q_{1-\alpha}(X), S)] \tag{15}$$
$$= \mathbb{E}_X[\gamma(X, g(X)) - \gamma(X, q_{1-\alpha}(X))], \tag{16}$$

where $\gamma(x, q) := \mathbb{E}_{S|X=x}[\ell_\alpha(q, S)]$.

Note that the derivative of γ with respect to q is

$$\gamma^{'}(x, q) = \frac{d}{dq}\gamma(x, q) = \Pr[S < q \mid X = x] - (1 - \alpha), \quad \forall q \in \mathbb{R}. \tag{17}$$

This will allow us to connect the pinball loss to conditional coverage of the prediction sets. To do so, we first show that we can bound $\gamma(.,.)$ in the following way

$$\gamma(x, q) - \gamma(x, q_{1-\alpha}(S \mid X = x)) \leq \frac{\gamma^{'}(x, q)^2}{2L} \tag{18}$$

To show (18), we assume that $q \geq q_{1-\alpha}(S \mid X = x)$ and note that for the other case, i.e. $q < q_{1-\alpha}(S \mid X = x)$, the proof follows similarly.

We can write

$$\begin{aligned}
\gamma(x, q) - \gamma(x, q_{1-\alpha}(S \mid X = x)) &= \gamma(x, q) - \gamma\left(x, q - \frac{\gamma^{'}(x, q)}{L}\right) + \gamma\left(x, q - \frac{\gamma^{'}(x, q)}{L}\right) - \gamma(x, q_{1-\alpha}(S \mid X = x)) \\
&\overset{(a)}{\geq} \gamma(x, q) - \gamma\left(x, q - \frac{\gamma^{'}(x, q)}{L}\right) \\
&\overset{(b)}{\geq} \int_{q - \frac{\gamma^{'}(x,q)}{L}}^{q} \gamma^{'}(x, \tilde{q}) d\tilde{q} \\
&\overset{(c)}{\geq} \int_{q - \frac{\gamma^{'}(x,q)}{L}}^{q} \left[\gamma^{'}(x, q) - L(q - \tilde{q})\right] d\tilde{q} \\
&= \frac{\gamma^{'}(x, q)^2}{L} - \frac{\gamma^{'}(x, q)^2}{2L} \\
&= \frac{\gamma^{'}(x, q)^2}{2L},
\end{aligned}$$

where (a) follows from (12), (b) is due to the fundamental theorem of calculus, and (c) follows from assumption 3.4 which results in the L-Lipschitz continuity of $\gamma^{'}(x, q)$ in term of q. This will conclude the proof of (18).

Continuing from (16), we have

$$\begin{aligned}
\mathbb{E}[\ell_\alpha(g(X), S) - \ell_\alpha(q_{1-\alpha}(X), S)] &\overset{(18)}{\geq} \mathbb{E}_X\left[\frac{\gamma^{'}(X, g(X))^2}{2L}\right] \\
&\overset{(17)}{=} \mathbb{E}_X\left[\frac{(\Pr[S < g(X) \mid X = x] - (1 - \alpha))^2}{2L}\right] \\
&= \frac{\text{MSCE}(C_g)}{2L},
\end{aligned}$$

which concludes the proof of the proposition. □

Proof of Theorem 3.6. By using the Jensen inequality we have,

$$\sum_{i=1}^{m} h^i(X)\ell_\alpha(q_i, S) \geq \ell_\alpha\left(\sum_{i=1}^{m} h^i(X) q_i, S\right) \quad \forall\, h \in \mathcal{X}^{\Delta_m}, \forall\, \boldsymbol{q} \in \mathbb{R}^m,$$

which is a consequence of the convexity of the pinball loss. Taking an expectation and a minimum from both sides we have,

$$\begin{aligned}\min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m}\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^i(X)\ell_\alpha(q_i,S)\right] &\geq \min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m}\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\ell_\alpha(\sum_{i=1}^m h^i(X)q_i,S)\right]\\ &\geq \mathbb{E}\left[\ell_\alpha(q_{1-\alpha}(X),S)\right]\end{aligned}$$

where the second inequality comes from (12). This means the quantity inside the absolute value on the left-hand-side of the Theorem statement (13) is positive.

At a high level, the proof will proceed as follows: We first show that we can bound the left-hand-side quantity in (13) by looking at an arbitrary partitioning of the space $\mathcal{X}$. Then, we further refine the bound by looking at a very specific partitioning of the space $\mathcal{X}$. Now, assume that $E=\{E_i\}_{i=1}^{m+1}$ is a partitioning on the set $\mathcal{X}$; i.e.

$$E=\{E_i\}_{i=1}^{m+1} \text{ such that } \bigcup_{i=1}^{m+1} E_i=\mathcal{X}, \text{and } E_i\cap E_j=\emptyset \quad \forall i\neq j,$$

and $\tilde{\boldsymbol{q}}$ is an arbitrary vector in $\mathbb{R}^m$. We can write,

$$\begin{aligned}\min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m}\mathbb{E}_{(X,S)}\left[\sum_{i=1}^m h^i(X)\ell_\alpha(q_i,S)\right] &= \min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m}\mathbb{E}_X\left[\mathbb{E}_{S|X}\left[\sum_{i=1}^m h^i(X)\ell_\alpha(q_i,S)\right]\right]\\ &= \min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m}\mathbb{E}_X\left[\sum_{i=1}^m \mathbb{E}_{S|X}\left[h^i(X)\ell_\alpha(q_i,S)\right]\right]\\ &= \min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m}\mathbb{E}_X\left[\sum_{i=1}^m h^i(X)\mathbb{E}_{S|X}\left[\ell_\alpha(q_i,S)\right]\right]\\ &\overset{(a)}{=} \min_{\boldsymbol{q}\in\mathbb{R}^m}\mathbb{E}_X\left[\min_{i\in[m]}\mathbb{E}_{S|X}\left[\ell_\alpha(q_i,S)\right]\right]\\ &\overset{(b)}{\leq} \sum_{i=1}^m\int_{x\in E_i} p(x)\mathbb{E}_{S|X=x}\left[\ell_\alpha(\tilde{q}_i,S)\right]\,dx \end{aligned}\tag{19}$$

Where (a) comes from the fact that the min over h is taken over all the functions in $\Delta_m^{\mathcal{X}}$, and (b) is due to the fact that E partitions the set $\mathcal{X}$.

A similar approach can be applied to reformulate $\mathbb{E}\left[\ell_\alpha(q_{1-\alpha}(X),S)\right]$ in terms of the partition E.

$$\begin{aligned}\mathbb{E}_{(X,S)}\left[\ell_\alpha\left(q_{1-\alpha}(S|X),S\right)\right] &= \mathbb{E}_X\,\mathbb{E}_{S|X}\left[\ell_\alpha\left(q_{1-\alpha}(S|X),S\right)\right]\\ &= \sum_{i=1}^m\int_{x\in E_i} p(x)\,\mathbb{E}_{S|X=x}\left[\ell_\alpha\left(q_{1-\alpha}(S|X=x),S\right)\right]\,dx\end{aligned}\tag{20}$$

Now putting together (19) and (20), we can write,

$$\begin{aligned}&\min_{h\in\Delta_m^{\mathcal{X}},\boldsymbol{q}\in\mathbb{R}^m}\mathbb{E}_{(X,S)}\left[\sum_{i=1}^m h^i(X)\ell_\alpha(q_i,S)\right]-\mathbb{E}_{(X,S)}\left[\ell_\alpha\left(q_{1-\alpha}(S|X),S\right)\right]\\ &\leq \sum_{i=1}^m\int_{x\in E_i} p(x)\left[\mathbb{E}_{S|X=x}\left[\ell_\alpha\left(\tilde{q}_i,S\right)-\ell_\alpha\left(q_{1-\alpha}(S|X=x),S\right)\right]\right]dx\\ &\leq \sum_{i=1}^m\int_{x\in E_i} p(x)\left|\,\tilde{q}_i-q_{1-\alpha}(S|X=x)\,\right|dx\end{aligned}\tag{21}$$

Here, the last inequality is due to Lemma D.2 which addresses the Lipschitzness of the pinball loss.

The relation (21) holds for any (arbitrary) partition E and vector $\tilde{\boldsymbol{q}}$. Consequently, by using a carefully crafted partition of the space $\mathcal{X}$, we will be able to obtain a tighter bound for our problem. Let us fix the following partitioning of the set $\mathcal{X}$ and

vector, keeping the same notation E and $\tilde{\boldsymbol{q}}$. Let η be an arbitrary positive real number,

$$E_i = \left\{ x \in \mathcal{X} \text{ such that: } \left| q_{1-\alpha}(S|X=x) - \mathbb{E}[q_{1-\alpha}(S|X)] + \eta - \frac{(2i-1)\eta}{m-1} \right| \leq \frac{\eta}{m-1} \right\} \quad \forall\, i \in [1, \cdots, m-1],$$

$$\tilde{q}_i = \mathbb{E}[q_{1-\alpha}(S|X)] - \eta + \frac{(2i-1)\eta}{m-1} \quad \forall\, i \in [1, \cdots, m-1],$$

and

$$E_m = \left\{ x \in \mathcal{X} \text{ such that: } \left| q_{1-\alpha}(S|X=x) - \mathbb{E}[q_{1-\alpha}(S|X)] \right| \geq \eta \right\},$$

$$\tilde{q}_m = \mathbb{E}[q_{1-\alpha}(S|X)].$$

By definition, $\bigcup_{i=1}^m E_i$ constitutes a partition of $\mathcal{X}$, and $\tilde{\boldsymbol{q}} = (\tilde{q}_1, \cdots, \tilde{q}_m)$ is a vector in $\mathcal{R}^m$. Substituting back into the inequality (21), we obtain:

$$\begin{aligned}
\min_{h \in \Delta_m^{\mathcal{X}}, \boldsymbol{q} \in \mathbb{R}^m} \mathbb{E}_{(X,S)} \left[\sum_{i=1}^m h^i(X) \ell_\alpha(q_i, S) \right] - \mathbb{E}_{(X,S)} \left[\ell_\alpha \left(q_{1-\alpha}(S|X), S \right) \right] \\
&\leq \sum_{i=1}^m \int_{x \in E_i} p(x) \left| \tilde{q}_i - q_{1-\alpha}(S|X=x) \right| dx \\
&\leq \frac{\eta}{m-1} + \int_{x \in E_m} p(x) \left| q_{1-\alpha}(S|X=x) - \mathbb{E}[q_{1-\alpha}(S|X)] \right| dx
\end{aligned} \tag{22}$$

To finish the proof, it remains to bound the second term in (22). We now show that for every $\eta > 0$ we have,

$$\int_{x \in E_m} p(x) \left| q_{1-\alpha}(S|X=x) - \mathbb{E}[q_{1-\alpha}(S|X)] \right| dx \leq \frac{\mathrm{Var}(q_{1-\alpha}(S|X))}{\eta} \tag{23}$$

To see the above inequality, we can write

$$\begin{aligned}
\mathrm{Var}(q_{1-\alpha}(S|X)) &= \int_{E_m} p(x) \left(q_{1-\alpha}(S|X=x) - \mathbb{E}[q_{1-\alpha}(S|X)) \right)^2 dx + \int_{{E_m}^c} p(x) \left(q_{1-\alpha}(S|X=x) - \mathbb{E}[q_{1-\alpha}(S|X)) \right)^2 dx \\
&\geq \int_{E_m} p(x) \left(q_{1-\alpha}(S|X=x) - \mathbb{E}[q_{1-\alpha}(S|X)) \right)^2 dx \\
&\geq \eta \int_{E_m} p(x) \left| q_{1-\alpha}(S|X=x) - \mathbb{E}[q_{1-\alpha}(S|X)) \right| dx
\end{aligned}$$

Rearranging the last inequality proves (23).

Finally, by plugging (23) into (22), we obtain

$$\begin{aligned}
\min_{h \in \Delta_m^{\mathcal{X}}, \boldsymbol{q} \in \mathbb{R}^m} \mathbb{E}_{(X,S)} \left[\sum_{i=1}^m h^i(X) \ell_\alpha(q_i, S) \right] - \mathbb{E}_{(X,S)} \left[\ell_\alpha \left(q_{1-\alpha}(S|X), S \right) \right] \\
&\leq \frac{\eta}{m-1} + \frac{\mathrm{Var}(q_{1-\alpha}(S|X))}{\eta},
\end{aligned}$$

and choosing $\eta = \sqrt{(m-1)\,\mathrm{Var}(q_{1-\alpha}(S|X))}$ gives us

$$\min_{h \in \Delta_m^{\mathcal{X}}, \boldsymbol{q} \in \mathbb{R}^m} \mathbb{E}_{(X,S)} \left[\sum_{i=1}^m h^i(X) \ell_\alpha(q_i, S) \right] - \mathbb{E}_{(X,S)} \left[\ell_\alpha \left(q_{1-\alpha}(S|X), S \right) \right] \leq 2 \sqrt{\frac{\mathrm{Var}(q_{1-\alpha}(S|X))}{m-1}}$$

. □

Proof of Theorem 3.11. We start by defining the following random variable

$$Z_{h,\boldsymbol{q}} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} h^i(X_j)\ell_\alpha(q_i, S_j) - \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m} h^i(X)\ell_\alpha(q_i, S)\right]. \tag{24}$$

Utilizing Lemma D.2 on Lipschitzness of pinball loss and assumption 3.2 we have with probability at least $1-\delta$ that,

$$|Z_{h,\boldsymbol{q}}| \le \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{n}}. \tag{25}$$

The rest of the proof is dedicated to show that a similar bound to (25) holds with high probability simultaneously for all the possible values of q and h. Recall the definitions of $\boldsymbol{q}^*$ (11) and $\boldsymbol{q}^\infty$ (10). Here one key observation is, since the random variable S only takes values in $[0,1]$, both $\boldsymbol{q}^\infty$ and $\boldsymbol{q}^*$ should have all their entries between 0 and 1. This can be shown for $\boldsymbol{q}^*$ by looking at the first order condition of (11) (similarly for $\boldsymbol{q}^\infty$ by looking at the (10)), for every $i \in [1, \cdots, m]$,

$$\begin{aligned}
&\frac{d}{dq_i}\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} h^i(X_j) l_\alpha(q_i, S_j) = 0 \\
&\iff \frac{1}{n}\sum_{j=1}^{n} h^i(X_j)\left[\mathbf{1}\left[S_j \le q_i\right] - (1-\alpha)\right] = 0
\end{aligned}$$

This means for every h the optimal q_i should be essentially a weighted quantile of $S_1, S_2, \cdots, S_n$. This results in $q_i^* \in [0,1]$, for every $i \in [1, \cdots, m]$. Consequently, the rest of the proof tries to show a similar bound holds for all the $h \in \mathcal{H}, \boldsymbol{q} \in \mathcal{V}$, where $\mathcal{V} = \{\boldsymbol{q} \in \mathcal{R}^m |\quad 0 \le q_i \le 1 \quad \forall i \in \{1, 2, \cdots, m\}\}$. To do so we look at ε-net sets over $\mathcal{H}$ and $\mathcal{V}$. A union bound will give us a bound over the ε-net sets, and then leveraging the ε-net properties we can argue a bound over $\mathcal{H}$ and $\mathcal{V}$.Let's continue with with proving the following lemma that will pave the way for the rest of arguments,

Lemma E.1. *For every $h_1, h_2 \in \mathcal{H}$ and $\boldsymbol{q^1}, \boldsymbol{q^2} \in \mathcal{V}$ we have,*

$$|Z_{h_1,\boldsymbol{q^1}} - Z_{h_2,\boldsymbol{q^2}}| \le 2||\boldsymbol{q^1} - \boldsymbol{q^2}||_\infty + 2d(h_1, h_2).$$

Proof.

$$\begin{aligned}
|Z_{h_1,\boldsymbol{q^1}} - Z_{h_2,\boldsymbol{q^2}}| &= |Z_{h_1,\boldsymbol{q^1}} - Z_{h_1,\boldsymbol{q^2}} + Z_{h_1,\boldsymbol{q^2}} - Z_{h_2,\boldsymbol{q^2}}| \\
&\le |Z_{h_1,\boldsymbol{q^1}} - Z_{h_1,\boldsymbol{q^2}}| + |Z_{h_1,\boldsymbol{q^2}} - Z_{h_2,\boldsymbol{q^2}}|
\end{aligned} \tag{26}$$

Now we bound each term separately,
First term:

$$\begin{aligned}
|Z_{h_1,\boldsymbol{q^1}} - Z_{h_1,\boldsymbol{q^2}}| &= \left|\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} h_1^i(X_j)\left(\ell_\alpha(q_i^1, S_j) - \ell_\alpha(q_i^2, S_j)\right) - \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m} h_1^i(X)\left(\ell_\alpha(q_i^1, S) - \ell_\alpha(q_i^2, S)\right)\right]\right| \\
&\overset{(a)}{\le} \left|\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} h_1^i(X_j)\left(\ell_\alpha(q_i^1, S_j) - \ell_\alpha(q_i^2, S_j)\right)\right| + \left|\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m} h_1^i(X)\left(\ell_\alpha(q_i^1, S) - \ell_\alpha(q_i^2, S)\right)\right]\right| \\
&\overset{(b)}{\le} \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} h_1^i(X_j)\left|\ell_\alpha(q_i^1, S_j) - \ell_\alpha(q_i^2, S_j)\right| + \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m} h_1^i(X)\left|\ell_\alpha(q_i^1, S) - \ell_\alpha(q_i^2, S)\right|\right] \\
&\overset{(c)}{\le} \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} h_1^i(X_j)\left|q_i^1 - q_i^2\right| + \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m} h_1^i(X)\left|q_i^1 - q_i^2\right|\right] \\
&\overset{(d)}{\le} 2||\boldsymbol{q^1} - \boldsymbol{q^2}||_\infty
\end{aligned}$$

Where (a) and (b) are triangle inequalities and (c) is followed by the fact that pinball loss is 1-lipschitz in terms of it's first argument(look at Lemma D.2). (d) is also by the definition of $||\boldsymbol{q^1} - \boldsymbol{q^2}||_\infty$.
Second term:

$$
\begin{aligned}
|Z_{h_1\boldsymbol{q^2}} - Z_{h_2\boldsymbol{q^2}}| &= \left| \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m}\left(h_1^i(X_j) - h_2^i(X_j)\right) l_\alpha(q_i^2, S_j) - \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m}\left(h_1^i(X) - h_2^i(X)\right) l_\alpha(q_i^2, S)\right]\right| \\
&\overset{(a)}{\leq} \left| \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m}\left(h_1^i(X_j) - h_2^i(X_j)\right) l_\alpha(q_i^2, S_j)\right| + \left|\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m}\left(h_1^i(X) - h_2^i(X)\right) l_\alpha(q_i^2, S)\right]\right| \\
&\overset{(b)}{\leq} \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m}\left|\left(h_1^i(X_j) - h_2^i(X_j)\right) l_\alpha(q_i^2, S_j)\right| + \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m}\left|\left(h_1^i(X) - h_2^i(X)\right) l_\alpha(q_i^2, S)\right|\right] \\
&\overset{(c)}{\leq} \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m}\left|h_1^i(X_j) - h_2^i(X_j)\right| \left|q_i^2\right| + \mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^{m}\left|h_1^i(X) - h_2^i(X)\right| \left|q_i^2\right|\right] \\
&\overset{(d)}{\leq} 2d(h_1, h_2),
\end{aligned}
$$

where (a) and (b) are triangle inequalities and (c) is followed by the fact that pinball loss is 1-lipschitz in terms of it's first argument(look at Lemma D.2). (d) is also by the definition of $d(h_1, h_2)$. Plugging back to the (26) concludes the Lemma (E.1). □

Continuing the proof of Theorem 3.11, let's say ε_1-net($\mathcal{H}$) and ε_2-net($\mathcal{V}$) are two minimal ϵ-net sets (ε-net sets with minimum cardinality). Then applying a union bound we have, with probability $1-\delta$,

$$
|Z_{h,\boldsymbol{q}}| \leq \sqrt{\frac{\ln\left(\frac{2\mathcal{N}(\mathcal{H},d,\epsilon_1)\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)}{\delta}\right)}{n}} \qquad \forall h \in \varepsilon_1 - \text{net}(\mathcal{H}), \quad \forall \boldsymbol{q} \in \varepsilon_2 - \text{net}(\mathcal{V}). \tag{27}
$$

Now we can do the following calculation for any arbitrary $h \in \mathcal{H}$ and $\boldsymbol{q} \in \mathcal{V}$. If $\tilde{h} \in \varepsilon_1$-net($\mathcal{H}$) and $\tilde{\boldsymbol{q}} \in \varepsilon_2$-net($\mathcal{V}$) such that $d(h, \tilde{h}) \leq \epsilon_1$ and $||\boldsymbol{q} - \tilde{\boldsymbol{q}}||_\infty \leq \epsilon_2$ then we have,

$$
\begin{aligned}
|Z_{h,\boldsymbol{q}}| \leq |Z_{h,\boldsymbol{q}} - Z_{\tilde{h},\tilde{\boldsymbol{q}}}| + |Z_{\tilde{h},\tilde{\boldsymbol{q}}}| &\leq 2||\boldsymbol{q} - \tilde{\boldsymbol{q}}||_\infty + 2d(h, \tilde{h}) + |Z_{\tilde{h},\tilde{\boldsymbol{q}}}| \\
&\leq 2\epsilon_1 + 2\epsilon_2 + \sqrt{\frac{\ln\left(\frac{2\mathcal{N}(\mathcal{H},d,\epsilon_1)\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)}{\delta}\right)}{n}}.
\end{aligned}
$$

Where the second inequality follows from Lemma E.1 and the third one is based on the fact that $\tilde{h} \in net_{\epsilon_1}(\mathcal{H})$ and $\tilde{\boldsymbol{q}} \in net_{\epsilon_2}(\mathcal{V})$. Now putting $\epsilon_1 = \epsilon_2 = \frac{1}{n}$ we have,

$$
\begin{aligned}
|Z_{h,\boldsymbol{q}}| &\leq \frac{2}{n} + \frac{2}{n} + \sqrt{\frac{\ln\left(\frac{2\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2))\mathcal{N}(\mathcal{H},d,\frac{1}{n})}{\delta}\right)}{n}} \\
&\leq 5\sqrt{\frac{\ln\left(\frac{2\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)\mathcal{N}(\mathcal{H},d,\frac{1}{n})}{\delta}\right)}{n}} \\
&= 5\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + \ln\left(\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)\right) + \ln\left(\mathcal{N}(\mathcal{H},d,\frac{1}{n})\right)}{n}}
\end{aligned}
$$

where the second inequality happens for sufficiently large n.

As a result of our calculations, we know have the following two inequalities hold at the same time with probability $1-\delta$,

$$|Z_{h^*,\boldsymbol{q}^*}| \leq 5\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)+\ln\left(\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)\right)+\ln\left(\mathcal{N}(\mathcal{H},d,\frac{1}{n})\right)}{n}}$$
$$|Z_{h^\infty,\boldsymbol{q}^\infty}| \leq 5\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)+\ln\left(\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)\right)+\ln\left(\mathcal{N}(\mathcal{H},d,\frac{1}{n})\right)}{n}}$$

This leads to,

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^{*i}(X)\ell_\alpha(q_i^*,S)\right]-\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^{\infty i}(X)\ell_\alpha(q_i^\infty,S)\right]\right| \\ \leq 10\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)+\ln\left(\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)\right)+\ln\left(\mathcal{N}(\mathcal{H},d,\frac{1}{n})\right)}{n}}\end{aligned} \tag{28}$$

Now putting everything together,

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^{*i}(X)\ell_\alpha(q_i^*,S)\right]-\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\ell_\alpha(q_{1-\alpha}(X),S)\right]\right| &\leq \left|\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^{*i}(X)\ell_\alpha(q_i^*,S)\right]\right. \\ &\quad \left.-\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^{\infty i}(X)\ell_\alpha(q_i^\infty,S)\right]\right| \\ &\quad +\left|\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^{\infty i}(X)\ell_\alpha(q_i^\infty,S)\right]\right. \\ &\quad \left.-\min_{h\in\mathcal{X}^{\Delta_m},\boldsymbol{q}\in\mathbb{R}^m}\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^{i}(X)\ell_\alpha(q_i,S)\right]\right| \\ &\quad +\left|\min_{h\in\mathcal{X}^{\Delta_m},\boldsymbol{q}\in\mathbb{R}^m}\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\sum_{i=1}^m h^{i}(X)\ell_\alpha(q_i,S)\right]\right. \\ &\quad \left.-\mathop{\mathbb{E}}_{(X,S)\sim\mathcal{D}}\left[\ell_\alpha(q_{1-\alpha}(X),S)\right]\right| \\ &\leq 10\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)+\ln\left(\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)\right)+\ln\left(\mathcal{N}(\mathcal{H},d,\frac{1}{n})\right)}{n}} \\ &\quad +2\sqrt{\frac{\text{var}(q_{1-\alpha}(X))}{m}}+\lambda_{\mathcal{H}},\end{aligned}$$

where the last inequality follows from (28), the definition of realizability gap (14), and Theorem 3.6. Lemma D.3 concludes the proof by proving an upperbound on the $\mathcal{N}(\mathcal{V},||.||_\infty,\epsilon_2)$.

□

F. Additional Figures

In this section you can find additional plots associated with the experiments.

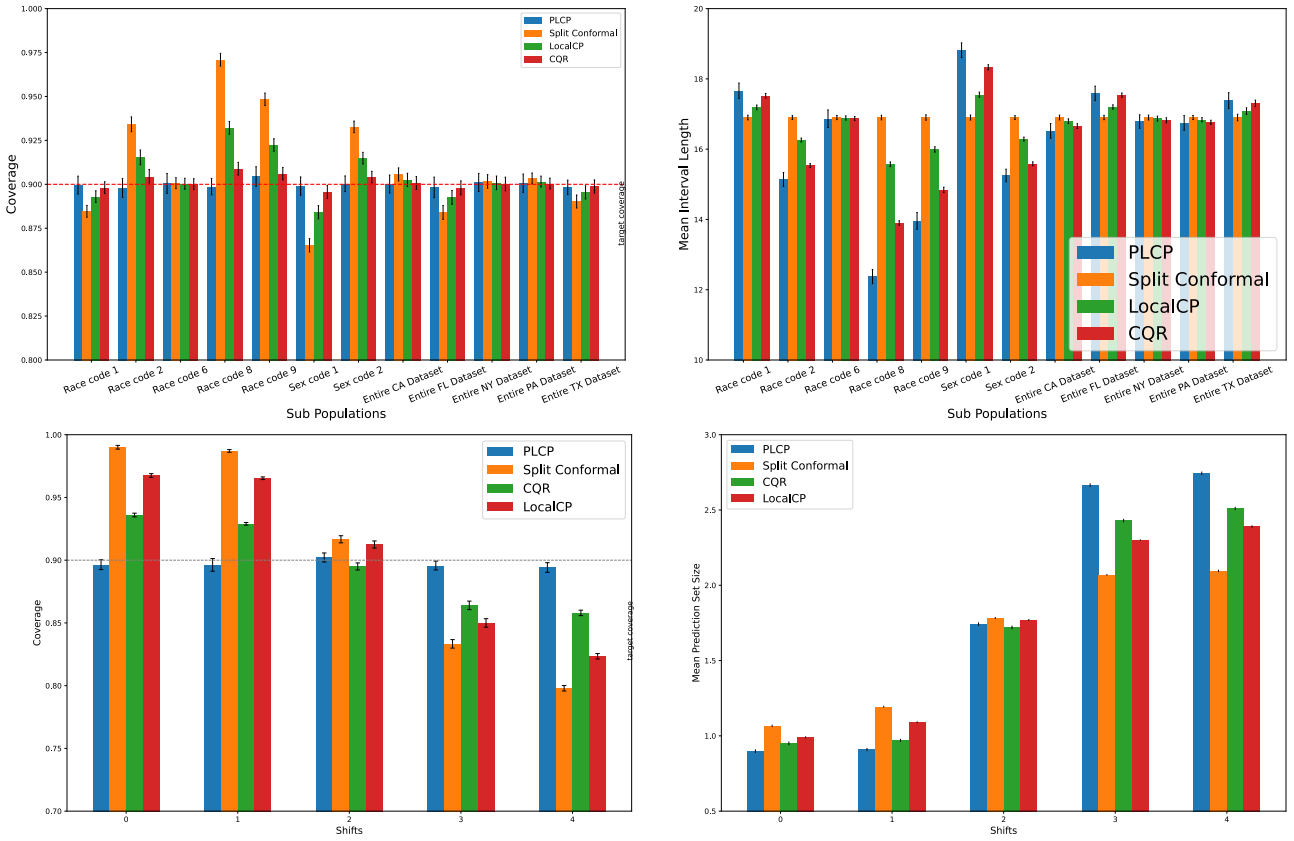

Figure 4: Left-hand-side plots show coverage and right-hand-side plots show mean prediction set size. Row 1: US Census Data; Row 2: MNIST with Gaussian Blur.

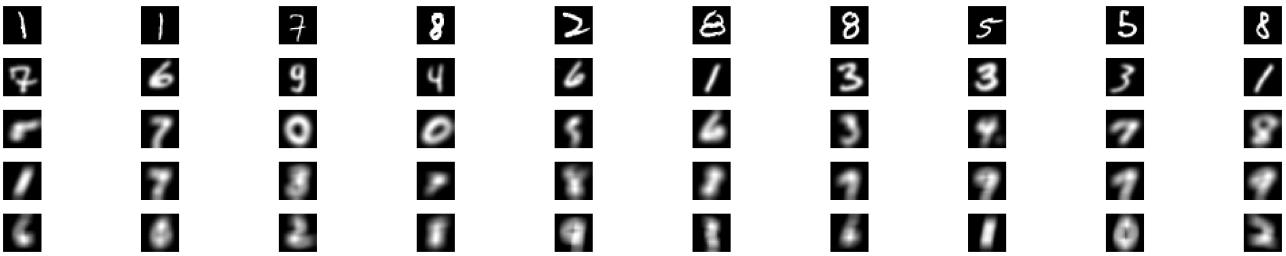

Figure 5: Sample images from 5 groups with increasing levels of gaussian blur applied from top to bottom.