Volume Optimality in Conformal Prediction with Structured Prediction Sets

Chao Gao [*1] **Liren Shan** [*2] **Vaidehi Srinivas** [*3] **Aravindan Vijayaraghavan** [*3]

Abstract

Conformal Prediction is a widely studied technique to construct prediction sets of future observations. Most conformal prediction methods focus on achieving the necessary coverage guarantees, but do not provide formal guarantees on the size (volume) of the prediction sets. We first prove an impossibility of volume optimality where any distribution-free method can only find a trivial solution. We then introduce a new notion of volume optimality by restricting the prediction sets to belong to a set family (of finite VC-dimension), specifically a union of k-intervals. Our main contribution is an efficient distribution-free algorithm based on dynamic programming (DP) to find a union of k-intervals that is guaranteed for any distribution to have near-optimal volume among all unions of k-intervals satisfying the desired coverage property. By adopting the framework of distributional conformal prediction (Chernozhukov et al., 2021), the new DP based conformity score can also be applied to achieve approximate conditional coverage and conditional restricted volume optimality, as long as a reasonable estimator of the conditional CDF is available. While the theoretical results already establish volume-optimality guarantees, they are complemented by experiments that demonstrate that our method can significantly outperform existing methods in many settings.

[*]Equal contribution [1]Department of Statistics, University of Chicago, Chicago, Illinois, USA [2]Toyota Technological Institute at Chicago, Chicago, Illinois, USA [3]Department of Computer Science, Northwestern University, Evanston, Illinois, USA. Correspondence to: Chao Gao <chaogao@uchicago.edu>, Liren Shan <lirenshan@ttic.edu>, Vaidehi Srinivas <vaidehi@u.northwestern.edu>, Aravindan Vijayaraghavan <aravindv@northwestern.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Conformal inference has emerged as a powerful black-box method for quantifying uncertainty in model predictions, providing confidence sets or prediction sets that contain the true value with a specified probability (Gammerman et al., 1998; Vovk et al., 2005). Consider a prediction problem where $\mathcal{X}$ is the covariate space (feature space), and $\mathcal{Y}$ is the label space. Given a dataset of n labeled samples $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$, a conformal prediction algorithm uses these n samples (often called calibration samples) to construct for a test $X_{n+1} \in \mathcal{X}$ with (unknown) true value $Y_{n+1} \in \mathcal{Y}$, a prediction set that we will denote by $\widehat{C}(X_{n+1}) \subseteq \mathcal{Y}$,[1] satisfying the *coverage* requirement for some desired parameter $\alpha \in (0, 1)$:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}(X_{n+1})\right) \geq 1 - \alpha. \tag{1}$$

Here the probability $\mathbb{P}$ refers to the joint distribution over all $n + 1$ pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ including the test sample (X_{n+1}, Y_{n+1}). Unlike traditional approaches, conformal inference is distribution-free, relying only on the assumption of exchangeability of the joint distribution $\mathbb{P}$ over the $(n + 1)$ samples.

While most conformal methods provide guarantees on coverage, they do not provide any control on the size or volume of the prediction sets.[2] In fact, the choice of $\widehat{C}(X_{n+1}) = \mathcal{Y}$ also satisfies the coverage requirement. Consequently, the size of these sets is often validated empirically, without formal guarantees. This raises the important question of *volume optimality*, which is the focus of this paper:

Question: *Given calibration samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn i.i.d. from a distribution P, can we find among all data-dependent sets $\widehat{C} \subset \mathcal{Y}$ satisfying the desired coverage requirement for $(X_{n+1}, Y_{n+1}) \sim P$, the one with*

[1]It may be more accurate to use $\widehat{C}(X_1, Y_1, \ldots, X_n, Y_n, X_{n+1})$ instead of $\widehat{C}(X_{n+1})$ to reflect that $\widehat{C}$ is a function of the calibration samples and X_{n+1}.

[2]Although conformal prediction ensures that the coverage probability does not exceed $1 - \alpha + 1/n$, it does not prevent overly conservative predictions.

the smallest volume, as quantified by the Lebesgue measure $\mathrm{vol}(\widehat{C}) = \lambda(\widehat{C})$?

The volume of the prediction set in conformal prediction is also sometimes referred to as 'efficiency' has been stated as an important consideration in many prior works (see e.g., Shafer & Vovk, 2008; Vovk et al., 2016; Angelopoulos & Bates, 2023). However, most works that we are aware of do not give theoretical guarantees of volume optimality, and mainly reason about volume control through empirical evaluations.

There are few works that provide guarantees of volume optimality.[3] Notable exceptions include Lei et al. (2013) in the unsupervised setting, Lei (2014); Vovk et al. (2016); Sadinle et al. (2016) in the classification setting, Lei et al. (2015) for functional data and recent works of Izbicki et al. (2020; 2022) and Kiyani et al. (2024) in the regression setting. As summarized by Angelopoulos et al. (2024), a sufficient condition that leads to volume optimality of conformal prediction is consistent estimation of the conditional density function of Y given X. This is essentially the strategy adopted by previous work (Lei et al., 2013; Izbicki et al., 2020; 2022). In comparison, our method, by incorporating a framework of Chernozhukov et al. (2021), builds on the estimation of the conditional CDF via a new conformity score computed by dynamic programming, and thus also works in settings where good conditional density estimation is impossible or density does not even exist.

1.1. Our Results

An Impossibility Result. We first prove an impossibility result in a one-dimensional setting where any distribution-free method that satisfies the coverage requirement can only find a trivial solution whose volume is sub-optimal. See Theorem 2.1 for a formal statement. This result provides an explanation for the lack of such volume-optimality guarantees in the conformal prediction literature, and also motivates our new notion of volume-optimality that we introduce in this work.

Structured Prediction Sets and Restricted Volume Optimality. Motivated by the impossibility result, our goal is to find a prediction set $\widehat{C} \in \mathcal{C} \subset 2^{\mathcal{Y}}$ whose volume is competitive with the optimum volume of any set in the family $\mathcal{C}$ as given by

$$\mathrm{OPT}_{\mathcal{C}}(P, 1-\alpha) = \inf_{C \in \mathcal{C}} \{\mathrm{vol}(C) : P(C) \geq 1-\alpha\}. \quad (2)$$

[3] Some works also guarantee that the coverage is not much more than $1-\alpha$, e.g., $\mathbb{P}\left(Y_{n+1} \in \widehat{C}(X_{n+1}) \mid X_{n+1}\right) \leq 1-\alpha+o(1)$ to argue that the prediction set is not too big. However, smallness according to the measure $\mathbb{P}$ does not necessarily reflect a small volume (or Lebesgue measure) for the set.

As long as $\mathcal{C}$ has bounded VC-dimension, for any distribution P we can obtain good empirical estimates of the probability measure of any set $C \in \mathcal{C}$ via a standard uniform concentration inequality, which allows us to overcome the impossibility result in Theorem 2.1. In the rest of the paper, we focus on the setting when $\mathcal{Y} = \mathbb{R}$ and $\mathcal{C} = \mathcal{C}_k$ which is the collection of unions of k intervals.

Conformalized Dynamic Programming. Equipped with our new notion of volume optimality, we propose a new conformity score based on dynamic programming. The proposed method is shown to not only achieve approximate conditional coverage as in Chernozhukov et al. (2021) and Romano et al. (2019), but also conditional volume optimality with respect to unions of k intervals, as long as a reasonable estimator of the conditional CDF is available. Our method of learning a predictive set via CDF can be regarded an extension of the framework of Izbicki et al. (2020); Chernozhukov et al. (2021).

1.2. Paper Organization

We will start with the unsupervised setting with label-only data in Section 2. The extension of the theory and algorithm to the supervised setting is given in Section 3. The numerical comparisons between our proposed methodology and existing methods in the literature are presented in Section 4. Due to the page limit, all technical proofs and additional numerical experiments will be presented in the appendix.

2. Unsupervised Setting

2.1. Approximate Volume Optimality

Suppose $Y_1, \cdots, Y_n, Y_{n+1}$ are independently drawn from a distribution P on $\mathbb{R}$. The goal is to predict Y_{n+1} based on the first n samples $Y_1, \cdots, Y_n$. To be specific, we would like to construct a data-dependent set $\widehat{C} = \widehat{C}(Y_1, \cdots, Y_n)$ such that

$$\mathbb{P}(Y_{n+1} \in \widehat{C}) \geq 1-\alpha. \quad (3)$$

Among all data-dependent sets that satisfy (3), our goal is to find the one with the smallest volume, quantified by the Lebesgue measure $\mathrm{vol}(\widehat{C}) = \lambda(\widehat{C})$. When the distribution P is known, one can directly minimize $\lambda(C)$, subject to $P(C) \geq 1-\alpha$ without even using the data. In particular, when $P \ll \lambda$, an optimal solution is given by the density level set

$$C_{\mathrm{opt}} = \left\{\frac{dP}{d\lambda} > t\right\} \cup D,$$

for some $t > 0$ and D is some subset of $\{dP/d\lambda = t\}$.

In general, P may not be absolutely continuous and the density need not exist. Nonetheless, we can still define the

optimal volume by

$$\mathrm{OPT}(P, 1-\alpha) = \inf\{\mathrm{vol}(C) : P(C) \geq 1-\alpha\}.$$

Note that without any assumption on P, the above optimization problem may not have a unique solution. Moreover, it is possible that the infimum cannot be achieved by any measurable set. Therefore, a natural relaxation is to consider approximate volume optimality. For some $\varepsilon \in (0, \alpha)$, a prediction set $\widehat{C}$ is called ε-optimal if

$$\mathrm{vol}(\widehat{C}) \leq \mathrm{OPT}(P, 1-\alpha+\varepsilon), \tag{4}$$

either in expectation or with high probability.

The notion of volume optimality defined by (4) is quite different from those considered in the literature. A popular quantity that has already been studied is the volume of set difference $\mathrm{vol}\left(\widehat{C}\Delta C_{\mathrm{opt}}\right)$ (Lei et al., 2013; Izbicki et al., 2020; Chernozhukov et al., 2021). However, this much stronger notion requires that the optimal solution C_{opt} must not only exist but also be unique. Usually additional assumptions need to be imposed in the neighborhood of the boundary of C_{opt} in order that the set difference vanishes in the large sample limit. In comparison, the definition (4) only requires the volume to be controlled, which can be achieved even if $\widehat{C}$ is not close to C_{opt}, or when C_{opt} does not even exist. Indeed, from a practical point of view, any set with coverage and volume control would serve the purpose of valid prediction. Insisting the closeness to a questionable target C_{opt} comes at the cost of unnecessary assumptions on the data generating process.

Another notion considered in the literature is close to our formulation (4). Instead of relaxing the coverage probability level from $1-\alpha$ to $1-\alpha+\varepsilon$, one can consider the following approximate volume optimality,

$$\mathrm{vol}(\widehat{C}) \leq \mathrm{OPT}(P, 1-\alpha) + \varepsilon. \tag{5}$$

Results of interval length optimality in the sense of (5) have been studied by Chernozhukov et al. (2021); Kiyani et al. (2024). However, the ε in (5) is usually proportional to the scale of the distribution P, or may depend on P in some other ways. In comparison, the ε in (4) has the unit of probability, and as we will show later, can be made independent of the distribution P, which leads to more natural and cleaner theoretical results with fewer assumptions.

2.2. Impossibility of Distribution-Free Volume Optimality

It is known that conformal prediction achieves the coverage property (3) in a distribution-free sense, meaning that (3) holds uniformly for all distributions P. One naturally hopes that the approximate volume optimality (4) can also be established in a distribution-free way. Perhaps not surprisingly, this goal is too ambitious. The theorem below rigorously proves the impossibility of the task. The detailed proof is deferred to Appendix D.1.

Theorem 2.1. *Consider observations $Y_1, Y_2, \ldots, Y_n, Y_{n+1}$ sampled i.i.d. from a distribution P on $\mathbb{R}$. Suppose $\widehat{C} = \widehat{C}(Y_1, \cdots, Y_n)$ satisfies $\mathbb{P}(Y_{n+1} \in \widehat{C}) \geq 1-\alpha$ for all distribution P. Then, for any $\varepsilon \in (0, \alpha)$, there exists some distribution P on $\mathbb{R}$, such that the expected volume of the prediction set is at least*

$$\mathbb{E}\,\mathrm{vol}(\widehat{C}) \geq \mathrm{OPT}(P, 1-\alpha+\varepsilon).$$

The above impossibility result can be regarded as a consequence of a nonparametric testing lower bound. Consider the following hypothesis testing problem,

$$\begin{aligned} H_0 &: \quad P = P_0 \\ H_1 &: \quad P \in \{P : \mathsf{TV}(P, P_0) > 1-\delta\}. \end{aligned}$$

It is well known that a testing procedure with both vanishing Type-1 and Type-2 errors does not exist without further constraining the alternative hypothesis, even when δ is arbitrarily close to 0 (LeCam & Schwartz, 1960; Barron, 1989). In the setting of distribution-free inference with simultaneous coverage and volume guarantees, two different probability measures naturally arise. The coverage guarantee is defined with respect to the joint distribution P^{n+1}, which governs the full dataset of n training samples and one test point. In contrast, the expected volume of the prediction set is measured under the product distribution $P^n \otimes \lambda$, where P^n represents the joint distribution of n training samples. When restricting the support of P to the unit interval $[0, 1]$, λ becomes the uniform probability, and thus both P^{n+1} and $P^n \otimes \lambda$ are probability distributions. It turns out achieving approximate volume optimality is related to hypothesis testing between P^{n+1} and $P^n \otimes \lambda$ with total variation separation.

2.3. Distribution-Free Restricted Volume Optimality

The impossibility result implies a volume lower bound $\mathrm{OPT}(P, 1-\alpha+\varepsilon)$, where the coverage level $1-\alpha+\varepsilon$ can be arbitrarily close to 1. This means that, at least in the worst case, the volume cannot be smaller than that of the support of P.

To avoid this triviality, in this section, we consider a weaker notion of volume optimality by only considering prediction sets that are unions of k intervals. We use $\mathcal{C}_k$ to denote the collection of all sets that are unions of k intervals. The restricted optimal volume with respect to the class $\mathcal{C}_k$ is defined by

$$\mathrm{OPT}_k(P, 1-\alpha) = \inf_{C \in \mathcal{C}_k} \{\mathrm{vol}(C) : P(C) \geq 1-\alpha\}. \tag{6}$$

Remark 2.2. We remark that we are still in a distribution-free setting, since no assumption is imposed on P. Instead, the restriction only constrains the shape of the prediction set. From a practical point of view, it is reasonable to require that $\widehat{C} \in \mathcal{C}_k$, since a more complicated prediction set would be hard to interpret. Moreover, as long as P admits a density function with at most k modes, the two notions match,

$$\mathrm{OPT}_k(P, 1-\alpha) = \mathrm{OPT}(P, 1-\alpha).$$

More generally, it can be shown that

$$\mathrm{OPT}_k(P, 1-\alpha) \leq \mathrm{OPT}(P, 1-\alpha+\varepsilon),$$

for some $\varepsilon \in (0, \alpha)$, whenever P can be approximated by a distribution with at most k modes. This, in particular, includes the situation where the density of P can be well estimated by a kernel density estimator. A rigorous statement will be given in Appendix C.

Given the observations $Y_1, \cdots, Y_n$, we define the empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$. To achieve restricted volume optimality, one can use

$$\widehat{C} = \operatorname*{argmin}_{C \in \mathcal{C}_k} \{\mathrm{vol}(C) : \mathbb{P}_n(C) \geq 1-\alpha\}. \tag{7}$$

According to its definition, the prediction set (7) satisfies both $\mathbb{P}_n(\widehat{C}) \geq 1-\alpha$ and $\mathrm{vol}(\widehat{C}) = \mathrm{OPT}_k(\mathbb{P}_n, 1-\alpha)$. The coverage and volume guarantees under P can be obtained via

$$\sup_{C \in \mathcal{C}_k} |\mathbb{P}_n(C) - P(C)| = O_P\left(\sqrt{\mathrm{VC}(\mathcal{C})/n}\right), \tag{8}$$

with $\mathrm{VC}(\mathcal{C}) = O(k)$. Therefore, approximate optimality can be achieved by (7) whenever (8) holds.

A naive exhaustive search to find (7) requires exponential computational time. We show that an efficient dynamic programming algorithm (Algorithm 1) can solve (7) approximately with some additional slack $\gamma \in [1/n, \alpha)$, which controls the trade-off between volume approximation accuracy and the computational complexity.

The dynamic programming table $DP(i, j, l)$ stores the minimum volume of i intervals that cover $l\gamma n$ points in $Y_{(1)}, \ldots, Y_{(j)}$ and the right endpoint of the rightmost interval is at $Y_{(j)}$, where $Y_{(1)}, \ldots, Y_{(n)}$ are training data points $Y_1, \ldots, Y_n$ sorted in non-decreasing order. For each state in DP table, we enumerate all possible left endpoint of the rightmost interval and the right endpoint of the previous interval (if it exists). After filling the table, a standard backtracking procedure is used to construct the final prediction set $\widehat{C}_{\mathrm{DP}}$.

Theoretical guarantees of Algorithm 1 are given in the following proposition.

Algorithm 1 Dynamic Programming Solving (7)

Input: data points $Y_1, \ldots, Y_n \in \mathbb{R}$, coverage level $1-\alpha \in (0,1)$ and slack $\gamma \in (0, \alpha)$, number of intervals k
Output: k intervals that cover $\lceil (1-\alpha)n \rceil$ points with minimum volume
Sort data in non-decreasing order $Y_{(1)} \leq \cdots \leq Y_{(n)}$
Set $DP(i, j, l) = \infty$ for all $i \in [k], j \in [n], l \in [1/\gamma]$
for $i = 1$ **to** k, $j = 1$ **to** n, $l = 1$ **to** $\lceil 1/\gamma \rceil$ **do**
  for $j' = i$ **to** j **do**
    for $j'' = i-1$ **to** $j'-1$ **do**
      Set $l' = l - \lfloor (j - j' + 1)/(\gamma n) \rfloor$
      if $l' < 0$ **and** $i = 1$ **then**
        $DP(i, j, l) = \min\{DP(i, j, l), Y_{(j)} - Y_{(j')}\}$
      end if
      if $DP(i-1, j'', l') \neq \infty$ **then**
        $DP(i, j, l) = \min\{DP(i, j, l), Y_{(j)} - Y_{(j')} + DP(i-1, j'', l')\}$
      end if
    end for
  end for
end for
Find the minimum volume among all $DP(k, j, \lceil (1-\alpha)/\gamma \rceil)$ for $j = 1, \ldots, n$ and backtrack to construct the prediction set $\widehat{C}_{\mathrm{DP}}$.
Return the set $\widehat{C}_{\mathrm{DP}}$.

Proposition 2.3. *For any $\gamma \in [1/n, \alpha)$, Algorithm 1 computes a prediction set $\widehat{C}_{\mathrm{DP}} \in \mathcal{C}_k$ by dynamic programming with time complexity $O(n^3 k/\gamma)$ such that*

1. $\mathbb{P}_n(\widehat{C}_{\mathrm{DP}}) \geq 1-\alpha$;
2. $\mathrm{vol}(\widehat{C}_{\mathrm{DP}}) \leq \mathrm{OPT}_k(\mathbb{P}_n, 1-\alpha+\gamma)$.

Together with (8), the coverage and volume guarantees of the dynamic programming can also be generalized from $\mathbb{P}_n$ to P.

2.4. Conformalizing Dynamic Programming

Having understood the generalization ability of dynamic programming, we are ready to conformalize the procedure to achieve a finite-sample coverage property. For simplicity, we will adopt the framework of split conformal prediction, though in principle full conformal prediction can also be applied here.

In the split conformal predicition framework, the data set is split into two halves. The first half is used to compute a conformity score, and the second half determines the quantile level. For convenience of notation, let us assume, from now on, that the sample size is $2n$. The split conformal procedure is outlined below.

1. Compute a score function $q(\cdot)$ using $Y_1, \cdots, Y_n$.
2. Evaluate $q(Y_{n+1}), \cdots, q(Y_{2n})$, and order them as $q_1 \leq \cdots \leq q_n$.
3. Output the prediction set

$$\widehat{C} = \left\{ y : q(y) \geq q_{\lfloor (n+1)\alpha \rfloor} \right\}. \tag{9}$$

By the exchangeability of $Y_{n+1}, \cdots, Y_{2n}, Y_{2n+1}$, the prediction set $\widehat{C}$ satisfies

$$\mathbb{P}\left(Y_{2n+1} \in \widehat{C} \right) \geq 1 - \alpha,$$

where the above probability is over the randomness of $(Y_1, \cdots, Y_n)$, that of $(Y_{n+1}, \cdots, Y_{2n})$, and that of Y_{2n+1}.

To conformalize the dynamic programming that approximately computes (7), we will first compute a nested system $S_1 \subset \cdots \subset S_m \subset \mathbb{R}$ using the data $Y_1, \cdots, Y_n$. The nested system is required to satisfy the following assumption.

Assumption 2.4. The sets $S_1 \subset \cdots \subset S_m \subset \mathbb{R}$ are measurable with respect to the σ-field generated by $Y_1, \cdots, Y_n$. Moreover, for some positive integer k, some $\alpha \in (0, 1)$ and some δ, γ such that $3\delta + \gamma + n^{-1} \leq \alpha$, we have

1. $\mathbb{P}_n(S_j) = \frac{j}{m}$ and $S_j \in \mathcal{C}_k$ for all $j \in [m]$.
2. There exists some $j^* \in [m]$, such that $\mathbb{P}_n(S_{j^*}) \geq 1 - \alpha + n^{-1} + 3\delta$ and $\mathrm{vol}(S_{j^*}) \leq \mathrm{OPT}_k(\mathbb{P}_n, 1 - \alpha + \frac{1}{n} + 3\delta + \gamma)$.

Here, $\mathbb{P}_n$ denotes the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ of the first half of the data.

To construct a nested system $\{S_j\}_{j\in[m]}$ that satisfies the above assumption, one only needs to make sure that there exists one subset S_{j^*} in the system that is computed by the dynamic programming (Algorithm 1) with confidence level $1 - \alpha + n^{-1} + 3\delta$ and slack parameter γ. The rest of the sets in the system can be constructed just to satisfy $\mathbb{P}_n(S_j) = \frac{j}{m}$. In Section 4.1, we will present a greedy expansion/contraction algorithm that satisfies Assumption 2.4.

With a nested system $\{S_j\}_{j\in[m]}$ satisfying Assumption 2.4, we can define the conformity score as

$$q(y) = \sum_{j=1}^{m} \mathbb{I}\{y \in S_j\}. \tag{10}$$

The equivalence between nested system and conformity score was advocated by (Gupta et al., 2022). Intuitively, $q(y)$ quantifies the depth of the location y. A higher score implies that y is covered by more sets in the nested system, and thus the location should be more likely to be included in the prediction set. Applying the standard split conformal framework, our prediction set based on conformalized dynamic programming is defined by (9) with the conformity score (10).

Theorem 2.5. *Consider i.i.d. observations $Y_1, \cdots, Y_{2n}, Y_{2n+1}$ generated by some distribution P on $\mathbb{R}$. Let $\widehat{C}_{\mathrm{CP-DP}}$ be the split conformal prediction set defined by the score (10) based on a nested system $\{S_j\}_{j\in[m]}$ satisfying Assumption 2.4. Suppose the parameter δ in Assumption 2.4 satisfies $\delta \gg \sqrt{\frac{k+\log n}{n}}$. Then the following properties hold.*

1. *Coverage:* $\mathbb{P}\left(Y_{2n+1} \in \widehat{C}_{\mathrm{CP-DP}} \right) \geq 1 - \alpha$.
2. *Restricted volume optimality:* $\mathrm{vol}(\widehat{C}_{\mathrm{CP-DP}}) \leq \mathrm{OPT}_k\left(P, 1 - \alpha + \frac{1}{n} + 4\delta + \gamma\right)$ *with probability at least* $1 - 2\delta$.

We emphasize that Theorem 2.5 guarantees both distribution-free coverage and distribution-free volume optimality properties. In practice, k is usually chosen to be a constant for prediction interpretability. By setting $\gamma = O\left(\sqrt{\frac{\log n}{n}}\right)$, the volume sub-optimality is at most $\frac{1}{n} + 4\delta + \gamma = O\left(\sqrt{\frac{\log n}{n}}\right)$.

3. Supervised Setting

3.1. Problem Setting

In this section, we consider conformal prediction with labeled data. Suppose data points $(X_1, Y_1), \cdots, (X_{2n}, Y_{2n}), (X_{2n+1}, Y_{2n+1})$ are $i.i.d.$ drawn from a distribution P on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \mathbb{R}$. Using the first $2n$ samples, our goal is to compute a prediction set $\widehat{C}(x)$ for each $x \in \mathcal{X}$. We will study the following properties for the prediction set.

1. *Marginal Coverage:*

$$\mathbb{P}\left(Y_{2n+1} \in \widehat{C}(X_{2n+1}) \right) \geq 1 - \alpha,$$

where the probability $\mathbb{P}$ is jointly over all $2n + 1$ pairs of observations.

2. *Conditional Coverage:*

$$\mathbb{P}\left(Y_{2n+1} \in \widehat{C}(X_{2n+1}) \mid X_{2n+1} \right) \geq 1 - \alpha, \tag{11}$$

with high probability.

It is well known that the conditional coverage property cannot be achieved without additional assumptions on P (Vovk,

2012; Lei & Wasserman, 2014; Foygel Barber et al., 2021). Therefore, some form of relaxation of (11) is necessary.

In addition to the coverage properties listed above, we will also extend the notion of restricted volume optimality (6) from the unsupervised setting to the supervised setting. Define the conditional CDF by

$$F(y \mid x) = \mathbb{P}\left(Y_{2n+1} \le y \mid X_{2n+1} = x\right).$$

The conditional restricted optimal volume is given by

$$\begin{aligned} & \mathrm{OPT}_k\left(F(\cdot \mid x), 1-\alpha\right) \\ = \; & \inf\left\{\mathrm{vol}(C) : \int_C \mathrm{d}F(\cdot \mid x) \ge 1-\alpha, C \in \mathcal{C}_k\right\}. \end{aligned}$$

With this definition, we can list the following volume requirement.

3. *Conditional Restricted Volume Optimality:*

$$\begin{aligned} & \mathrm{vol}(\widehat{C}(X_{2n+1})) \\ \le \; & \mathrm{OPT}_k\left(F(\cdot|X_{2n+1}), 1-\alpha+\varepsilon\right), \end{aligned} \quad (12)$$

with high probability, for some $\varepsilon \in (0, \alpha)$.

Similar to the conditional coverage property (11), the conditional restricted volume optimality (12) is only required for a typical value of the design point. We will show that based on an extension of distributional conformal prediction (Chernozhukov et al., 2021), these two properties can be achieved under the same assumption.

3.2. Distributional Conformal Prediction

Conformal prediction based on estimating the conditional CDF has been considered independently by (Izbicki et al., 2020; Chernozhukov et al., 2021). We will briefly review the version by Chernozhukov et al. (2021), and then extend it to serve our purpose. Suppose $\widehat{F}(y \mid x)$ is an estimator of the conditional CDF, which is computed from the first half of the data $(X_1, Y_1), \cdots, (X_n, Y_n)$. The prediction set proposed by Chernozhukov et al. (2021) is

$$\widehat{C}_{\mathrm{DCP}}(X_{2n+1}) = \left\{y \in \mathbb{R} : \left|\widehat{F}(y \mid X_{2n+1}) - \frac{1}{2}\right| \le \widehat{t}\right\},$$

where $\widehat{t}$ is an appropriate quantile of

$$\left\{\left|\widehat{F}(Y_{n+1} \mid X_{n+1}) - \frac{1}{2}\right|, \cdots, \left|\widehat{F}(Y_{2n} \mid X_{2n}) - \frac{1}{2}\right|\right\}.$$

Since $\widehat{C}_{\mathrm{DCP}}(X_{2n+1})$ is in the form of split conformal prediction, the marginal coverage property is automatically satisfied. When $\widehat{F}(y \mid x)$ is close to $F(y \mid x)$ in some appropriate sense, it was proved by (Chernozhukov et al., 2021) that asymptotic conditional coverage also holds. However, in general, $\widehat{C}_{\mathrm{DCP}}(X_{2n+1})$ is not optimal in terms of its volume. A modification was also proposed in (Chernozhukov et al., 2021) to achieve volume optimality within the class of intervals. Though not explicitly stated in (Chernozhukov et al., 2021), we believe that the DCP procedure essentially achieves (12) for $k = 1$. Our goal is to achieve the conditional restricted volume optimality for a general k by combining the ideas of DCP and dynamic programming (DP).

3.3. DCP meets DP

To achieve (12) for a general k, we will modify the DCP procedure by considering a different conformity score that generalizes (10) to the supervised setting. Recall that $\widehat{F}(y \mid x)$ is an estimator of the conditional CDF, and it is computed from the first half of the data $(X_1, Y_1), \cdots, (X_n, Y_n)$. Our first step is to construct a nested system for each $x \in \mathcal{X}$. To be specific, for each $x \in \mathcal{X}$, we will construct a collection of sets $\{S_j(x)\}_{j \in [m]}$ based on the function $\widehat{F}(\cdot \mid x)$. The requirement of the nested system is summarized as the following assumption.

Assumption 3.1. The sets $S_1(x) \subset \cdots \subset S_m(x) \subset \mathbb{R}$ are measurable with respect to the σ-field generated by $\widehat{F}(\cdot \mid x)$. Moreover, for some positive integer k, some $\alpha \in (0, 1)$ and some δ, γ such that $3\delta + \gamma + n^{-1} \le \alpha$, we have

1. $\int_{S_j(x)} \mathrm{d}\widehat{F}(\cdot \mid x) = \frac{j}{m}$ and $S_j \in \mathcal{C}_k$ for all $j \in [m]$.
2. There exists some $j^* \in [m]$, such that

$$\int_{S_{j^*}(x)} \mathrm{d}\widehat{F}(\cdot \mid x) \ge 1 - \alpha + n^{-1} + 3\delta$$

and $\mathrm{vol}(S_{j^*}) \le \mathrm{OPT}_k(\widehat{F}(\cdot \mid x), 1 - \alpha + \frac{1}{n} + 3\delta + \gamma)$.

The construction of nested systems satisfying Assumption 3.1 is similar to that in the unsupervised setting. That is, one can apply dynamic programming (Algorithm 1) and obtain $S_{j^*}(x)$, and the rest of the sets can be constructed via the greedy expansion/contraction procedure described in Section 4.1 to satisfy $\int_{S_j(x)} \mathrm{d}\widehat{F}(\cdot \mid x) = \frac{j}{m}$. The main difference here is that Algorithm 1 is directly applied to the data in the unsupervised setting, while we only have access to $\widehat{F}(\cdot \mid x)$ in the supervised setting. This issue can be easily addressed by computing quantiles $Y_1(x), \cdots, Y_L(x)$ from $\widehat{F}(\cdot \mid x)$ on a grid, and then apply Algorithm 1 with $Y_1(x), \cdots, Y_L(x)$ as input. Indeed, since the distance between $\widetilde{F}(\cdot \mid x)$ and $\widehat{F}(\cdot \mid x)$ can be controlled by the size of the grid with $\widetilde{F}(y \mid x) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{I}\{Y_l(x) \le y\}$, Assumption 3.2, which will be stated later in Section 3.4, is also satisfied by $\widetilde{F}(y \mid x)$ (with a slightly larger value of δ) by triangle inequality.

The computational cost of constructing $\{S_j(x)\}_{j\in[m]}$ for a single $x \in \mathcal{X}$ is $O(L^3k/\gamma)$. Note that there is no need to repeat the construction for each individual $x \in \mathcal{X}$. Since the split conformal framework only requires evaluating the conformity score at $(X_{n+1}, Y_{n+1}), \cdots, (X_{2n}, Y_{2n}), (X_{2n+1}, y)$, it is sufficient to compute $\{S_j(X_i)\}_{j\in[m]}$ for $i = n+1, \cdots, 2n+1$, which leads to the total computational cost $O(nL^3k/\gamma)$.

With nested systems satisfying Assumption 3.1, the conformity score in the supervised setting is defined as

$$q(y, x) = \sum_{j=1}^{m} \mathbb{I}\{y \in S_j(x)\}.$$

Let $q_1 \le \cdots \le q_n$ be the order statistics computed from the set

$$\{q(X_{n+1}, Y_{n+1}), \cdots, q(X_{2n}, Y_{2n})\}.$$

The prediction set for Y_{2n+1} is constructed as

$$\widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1}) = \left\{y : q(y, X_{2n+1}) \ge q_{\lfloor (n+1)\alpha \rfloor}\right\}.$$

3.4. Theoretical Guarantees

We will show in this section that $\widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1})$ satisfies marginal coverage. Moreover, when $\widehat{F}(y \mid x)$ is close to $F(y \mid x)$ in some appropriate sense, it also satisfies approximate conditional coverage and conditional restricted volume optimality. Given two CDFs $\widehat{F}$ and F, we define the (k, ∞) norm of the difference by

$$\|\widehat{F} - F\|_{k,\infty} = \sup_{C\in\mathcal{C}_k} \left| \int_C \mathrm{d}\widehat{F} - \int_C \mathrm{d}F \right|.$$

Assumption 3.2. The estimated conditional CDF $\widehat{F}(y \mid x)$ satisfies

$$\mathbb{P}\left(\|\widehat{F}(\cdot \mid X_{2n+1}) - F(\cdot \mid X_{2n+1})\|_{k,\infty} \le \delta\right) \ge 1 - \delta,$$

where δ takes the same value as the one in Assumption 3.1.

The theoretical properties of $\widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1})$ are given by the theorem below.

Theorem 3.3. *Consider i.i.d. observations (X_1, Y_1), $\ldots$, (X_{2n}, Y_{2n}), (X_{2n+1}, Y_{2n+1}) generated by some distribution P on $\mathcal{X} \times \mathbb{R}$. The conformal prediction set $\widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1})$ is computed from nested systems $\{S_j(\cdot)\}_{j\in[m]}$ and $\widehat{F}(\cdot \mid \cdot)$ satisfying Assumption 3.1 and Assumption 3.2. Suppose the parameter δ in the two assumptions satisfies $\delta^2 \ge \frac{\log(2\sqrt{n})}{2n}$. Then the following properties hold.*

1. *Marginal coverage,*

$$\mathbb{P}\left(Y_{2n+1} \in \widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1})\right) \ge 1 - \alpha.$$

2. *Approximate conditional coverage,*

$$\mathbb{P}\left(Y_{2n+1} \in \widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1}) | X_{2n+1}\right) \ge 1 - \alpha - 3\delta,$$

with probability at least $1 - \delta$.

3. *Conditional restricted volume optimality,*

$$\mathrm{vol}\left(\widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1})\right) \le \mathrm{OPT}_k\left(F(\cdot \mid X_{2n+1}), 1 - \alpha + \frac{1}{n} + 4\delta + \gamma\right),$$

with probability at least $1 - 2\delta$.

Note that the marginal coverage does not depend on δ, which reflects the estimation error of the conditional CDF in Assumption 3.2. Therefore, the marginal coverage guarantee always holds and does not rely on Assumption 3.2. Theorem 3.3 can be regarded as a generalization of Theorem 2.5. Indeed, when $F(\cdot \mid x)$ does not depend on x and $\widehat{F}(\cdot \mid x)$ is defined as the empirical CDF of $Y_1, \cdots, Y_n$, Theorem 3.3 recovers Theorem 2.5. Moreover, since the volume optimality is over all sets that are unions of k intervals, it also covers the length optimality of intervals considered by (Chernozhukov et al., 2021). The case $k \ge 2$ will be important if the conditional density of Y given X has multiple modes; Gaussian mixture is a leading example.

4. Numerical Experiments

We complement our theoretical guarantees with an evaluation of our methods for both the unsupervised setting of Section 2 and the supervised setting of Section 3.

4.1. Construction of Nested Systems

We first describe a procedure that generates a nested system $\{S_j\}_{j\in[m]}$ that satisfies Assumption 2.4. The construction involves the following steps:

1. *Generate S_{j^*} by Dynamic Programming.* For $j^* = \lceil (1 - \alpha + n^{-1} + 3\delta)m \rceil$, we generate S_{j^*} by applying Algorithm 1 with coverage level $1 - j^*/m$ and slack $\gamma = 1/m$. The discretization level m and statistical tolerance δ are chosen as $m = 50$ and $\delta = \sqrt{(k + \log n)/n}$,[4] where k is the number of intervals in the prediction set.

2. *Generate $S_{j^*+1}, \cdots, S_m$ by Greedy Expansion.* For each $j > j^*$, we iteratively identify the closest uncovered data point to the boundary of the current k

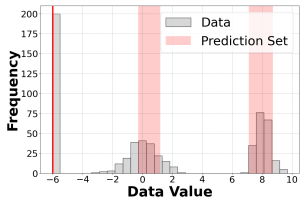
[4] We can choose $m \in (1/\alpha, n]$, where larger m yields finer discretization of the nested system. The parameter δ reflects the statistical error in estimating the (conditional) CDF.

intervals and expand the nearest interval to cover it. Once the intervals cover $\lceil jn/m \rceil$ data points, we define the union as S_j and move on to the construction of S_{j+1}.

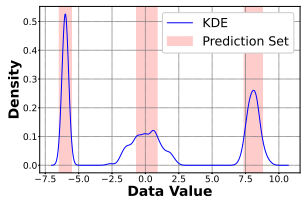
3. *Generate $S_1, \cdots, S_{j^*-1}$ by Greedy Contraction.* For each $j < j^*$, we iteratively remove a boundary point of the current k intervals that results in the maximum volume reduction. Once the intervals after contraction cover exactly $\lceil jn/m \rceil$ data points, we define the union as S_j and move on to the construction of S_{j-1}.

In the supervised setting, the above procedure will be applied to quantiles $Y_1(x), \cdots, Y_L(x)$ computed from $\widehat{F}(\cdot \mid x)$ with $L = m$ for all $x \in \{X_{n+1}, \cdots, X_{2n+1}\}$.

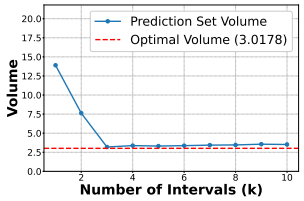
4.2. Comparison in Unsupervised Settings



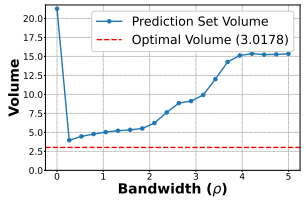
(a) The histogram of the dataset and the prediction set given by conformalized DP with $k = 3$ intervals (The first interval is at $[-6.03, -5.97]$.). The volume of the prediction sets is 3.1438.



(b) The kernel density estimation with bandwidth $\rho = 0.25$ and the prediction set given by conformalized KDE (Lei et al., 2013). The volume of the prediction sets is 3.7944.



(c) Volumes of prediction sets by conformalized DP is not sensitive to the choice of k.



(d) Volumes of prediction sets by conformalized KDE is highly sensitive to the choice of ρ.

Figure 1. Conformal prediction sets on the mixture of Gaussians data from $P = \frac{1}{3}N(-6, 0.0001) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(8, 0.25)$. The target coverage is 0.8. The theoretically optimal volume for this target coverage is 3.0178.

The algorithm in Section 2 is compared against the method based on kernel density estimation due to Lei et al. (2013) and evaluated on several different distributions. Though the original conformalized KDE was proposed in the full conformal framework, we will consider its split conformal version for a direct comparison. We believe the comparison between the full conformal versions of the two methods will lead to the same conclusion. For the conformalized DP method, the conformity score is constructed based on the nested system described in Section 4.1. The conformalized KDE is also in the form of (9), with the conformity score given by $q_{\text{KDE}}(x) = \frac{1}{n\rho}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{\rho}\right)$, where $K(\cdot)$ is the standard Gaussian kernel and ρ is the bandwidth parameter. Both methods involve a single tuning parameter, k for conformalized DP and ρ for conformalized KDE.

Figure 1, Tables 1 and 2 summarize the results using data generated from a mixture of Gaussians $\frac{1}{3}N(-6, 0.0001) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(8, 0.25)$. We report the mean and standard deviation of the results over 20 independent runs. Additional experiments on other distributions, including standard Gaussian, censored Gaussian, and ReLU-transformed Gaussian, will be presented in Appendix B.

Table 1. Conformalized KDE on the mixture of Gaussians data

Bandwidth	Volume	Coverage
0.01	3.8929 ± 0.2504	0.7899 ± 0.0258
0.25	**3.8747 ± 0.2673**	0.7963 ± 0.0231
0.5	4.3663 ± 0.3388	0.7938 ± 0.0244
0.75	4.7162 ± 0.3681	0.7931 ± 0.0224
1	4.9462 ± 0.5119	0.7933 ± 0.0231

Table 2. Conformalized DP on the mixture of Gaussians data

k	Volume	Coverage
1	13.9017 ± 0.1112	0.7947 ± 0.0309
2	7.5980 ± 0.2159	0.7883 ± 0.0373
3	**3.1888 ± 0.3403**	0.8024 ± 0.0356
4	3.3459 ± 0.3003	0.8174 ± 0.0278
5	3.4250 ± 0.3818	0.8202 ± 0.0306

The two methods are also evaluated on two real-world datasets (Acidity and Enzyme) used in density estimation literature (Richardson & Green, 1997). The experiments target a coverage level of 0.8. We report the means and standard deviations of the results over 50 independent runs. As shown in Tables 3 and 4, our conformalized DP outputs a smaller volume prediction set than the conformalized KDE with the best bandwidth for almost all $k \geq 2$. The results for the Enzyme dataset are given in Appendix B.

Table 3. Conformalized KDE on Acidity Dataset

Bandwidth	Volume	Coverage
0.1	2.4927 ± 0.1960	0.8401 ± 0.0226
0.3	**2.3934 ± 0.2044**	0.8092 ± 0.0261
0.5	2.5749 ± 0.2571	0.8133 ± 0.0315
0.7	2.7617 ± 0.2038	0.8013 ± 0.0256
0.9	2.8196 ± 0.2587	0.8021 ± 0.0294

Table 4. Conformalized DP on Acidity Dataset

k	Volume	Coverage
1	2.5999 ± 0.1937	0.8121 ± 0.0476
2	2.3627 ± 0.1755	0.8507 ± 0.0224
3	2.4099 ± 0.1832	0.8552 ± 0.0245
4	2.3615 ± 0.1452	0.8582 ± 0.0196
5	**2.2172 ± 0.1349**	0.8341 ± 0.0253

4.3. Comparison in Supervised Settings

The algorithms for the supervised setting are compared against conformalized quantile regression (CQR) (Romano et al., 2019), distributional conformal prediction methods (DCP-QR and DCP-QR*) of Chernozhukov et al. (2021), and CD-Split and HPD-Split methods (Izbicki et al., 2022) against benchmark simulated datasets in Romano et al. (2019); Izbicki et al. (2020) (Figures 2 and 3). The implementation details of all methods are given in Appendix B.

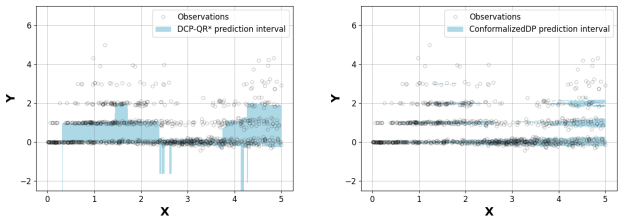


Figure 2. Results in the supervised setting on a synthetic data from Romano et al. (2019) for target coverage 0.7. The left plot shows the output of DCP-QR*, the state of the art method by Chernozhukov et al. (2021), which outputs prediction sets with average volume 1.29 and empirical coverage 0.7106. The right plot shows the output of our method with $k = 5$ intervals, which achieves a significantly improved average volume of 0.45 with empirical coverage 0.7236.

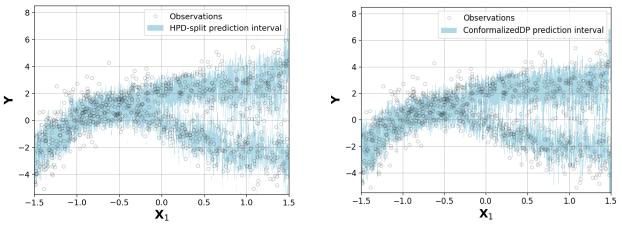


Figure 3. Results in the supervised setting on a synthetic data with 20 dimensional feature from Izbicki et al. (2020) for target coverage 0.7. The left plot shows the output of HPD-Split method by Izbicki et al. (2022), with average volume 3.60. The right plot shows the output of our method with $k = 2$ intervals, which has an average volume 3.55.

As shown in Tables 5 and 6, our method outperforms previous methods by outputting prediction sets that are unions of intervals. Among all other methods, the CD-split and HDP-split (Izbicki et al., 2020) are also able to produce unions of intervals. However, since these methods rely on consistent estimation of the conditional density, our conformalized DP still produces prediction sets with smaller volumes. The comparison is more pronounced on the first dataset (see Figure 2 and Table 5), where it would be inappropriate to assume a smooth conditional density, but our method is based on conformalizing the estimation of conditional CDF, and thus works in much more general settings. Table 6 shows results where the conditional density is smooth and can be accurately estimated, favoring density-based methods. Even in this case, the conformalized DP with $k = 2$ is still competitive against CD-split and HPD-split.

Table 5. Comparison on simulated data in Romano et al. (2019).

Method	Average Volume	Coverage
CQR	1.4237 ± 0.0743	0.7036 ± 0.0146
DCP-QR	1.4218 ± 0.0647	0.7021 ± 0.0100
DCP-QR*	1.8854 ± 0.9772	0.7054 ± 0.0124
CD-split	1.7118 ± 0.1934	0.6330 ± 0.0166
HPD-split	1.7557 ± 0.1145	0.6874 ± 0.0165
C-DP (k=1)	1.0897 ± 0.0792	0.7119 ± 0.0171
C-DP (k=5)	**0.4660 ± 0.0218**	**0.7152 ± 0.0177**

Table 6. Comparison on simulated data in Izbicki et al. (2020).

Method	Average Volume	Coverage
CQR	4.0428 ± 0.0992	0.7060 ± 0.0116
DCP-QR	3.9933 ± 0.0854	0.6987 ± 0.0160
DCP-QR*	4.0701 ± 0.1141	0.7004 ± 0.0162
CD-split	**3.6320 ± 0.1126**	0.7002 ± 0.0161
HPD-split	**3.6084 ± 0.1121**	0.7014 ± 0.0133
C-DP ($k = 1$)	4.1450 ± 0.1150	0.7126 ± 0.0165
C-DP ($k = 2$)	**3.6774 ± 0.1298**	**0.7152 ± 0.0138**

5. Conclusion

We study conformal prediction with volume optimality in both the unsupervised setting and the supervised setting, by proposing a new conformity score computed via dynamic programming. In the supervised setting, when consistent estimation of the conditional CDF is available, we prove that the proposed method not only achieves conditional coverage, but the output of prediction set also has approximate conditional volume optimality with respect to the class of unions of k intervals.

Our method is especially suitable to settings where the data generating process is multi-modal or has a mixture structure. The numerical experiments show that the performance of the method is quite insensitive to the choice of k, whenever it is not smaller than the number of modes of the distribution. For future work, it would be interesting to study restricted volume optimality in more general response settings and under other notions of coverage in conformal prediction.

Acknowledgments

This research project was supported by NSF-funded Institute for Data, Econometrics, Algorithms and Learning (IDEAL) through the grants NSF ECCS-2216970 and ECCS-2216912. The research started as part of the IDEAL special program on Reliable and Robust Data Science. Chao Gao was supported by NSF Grant DMS-2310769 and an Alfred Sloan fellowship. Vaidehi Srinivas was supported by the Northwestern Presidential Fellowship. We also acknowledge the support of the NSF-Simons SkAI institute and the NSF-Simons NITMB institute, and thank Rina Barber and Jing Lei for helpful discussions.

Impact Statement

This paper presents improved methods for quantifying uncertainty in predictions produced by a machine learning algorithm. Such conformal prediction methods can make machine learning models more reliable by providing valid confidence sets or prediction sets. This has potential for positive societal consequences when accurate estimates of uncertainty are useful in making more informed and reliable downstream decisions.

References

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4): 494–591, March 2023. ISSN 1935-8237. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Angelopoulos, A. N., Barber, R. F., and Bates, S. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Bai, Y., Mei, S., Wang, H., Zhou, Y., and Xiong, C. Efficient and differentiable conformal prediction with general function classes. In *International Conference on Learning Representations*, 2022.

Barber, R., Candès, E., Ramdas, A., and Tibshirani, R. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51, 04 2023. doi: 10.1214/23-AOS2276.

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.

Barron, A. R. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, 17(1):107–124, 1989.

Carreira-Perpinán, M. A. and Williams, C. K. On the number of modes of a gaussian mixture. In *International Conference on Scale-Space Theories in Computer Vision*, pp. 625–640. Springer, 2003.

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

Devroye, L. and Lugosi, G. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

Dhillon, G. S., Deligiannidis, G., and Rainforth, T. On the expected size of conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1549–1557. PMLR, 2024.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Gammerman, A., Vovk, V., and Vapnik, V. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pp. 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Ghosal, S. and van der Vaart, A. Posterior convergence rates of dirichlet mixtures at smooth densities. *Annals of Statistics*, 35(2):697–723, 2007.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Izbicki, R. and Lee, A. B. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11:2800–2831, 2017.

Izbicki, R., Shimizu, G., and Stern, R. Flexible distribution-free conditional predictive bands using density estimators. In *International Conference on Artificial Intelligence and Statistics*, pp. 3068–3077. PMLR, 2020.

Izbicki, R., Shimizu, G., and Stern, R. B. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87): 1–32, 2022.

Kim, B., Xu, C., and Barber, R. Predictive inference is free with the jackknife+-after-bootstrap. *Advances in Neural Information Processing Systems*, 33:4138–4149, 2020.

Kiyani, S., Pappas, G., and Hassani, H. Length optimization in conformal prediction, 2024. URL https://arxiv.org/abs/2406.18814.

Kruijer, W., Rousseau, J., and van der Vaart, A. Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

Kumar, B., Lu, C.-C., Gupta, G., Palepu, A., Bellamy, D. R., Raskar, R., and Beam, A. L. Conformal prediction with large language models for multi-choice question answering. *ArXiv*, abs/2305.18404, 2023. URL https://api.semanticscholar.org/CorpusID:258967849.

LeCam, L. and Schwartz, L. A necessary and sufficient condition for the existence of consistent estimates. *The Annals of Mathematical Statistics*, 31(1):140–150, 1960.

Lei, J. Classification with confidence. *Biometrika*, 101(4): 755–769, 2014.

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., Robins, J. M., and Wasserman, L. A. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108:278 – 287, 2013. URL https://api.semanticscholar.org/CorpusID:17499892.

Lei, J., Rinaldo, A., and Wasserman, L. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2015.

Liang, R., Zhu, W., and Barber, R. F. Conformal prediction after efficiency-oriented model selection. *arXiv preprint arXiv:2408.07066*, 2024.

Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research*, 7(35):983–999, 2006. URL http://jmlr.org/papers/v7/meinshausen06a.html.

Richardson, S. and Green, P. J. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4):731–792, 1997.

Roebroek, J. sklearn-quantile, 2023. URL https://sklearn-quantile.readthedocs.io/en/latest/index.html.

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

Sadinle, M., Lei, J., and Wasserman, L. A. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114:223 – 234, 2016. URL https://api.semanticscholar.org/CorpusID:622583.

Scott, C. D. and Nowak, R. D. Learning minimum volume sets. *J. Mach. Learn. Res.*, 7:665–704, December 2006. ISSN 1532-4435.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL http://jmlr.org/papers/v9/shafer08a.html.

Stutz, D., Dvijotham, K. D., Cemgil, A. T., and Doucet, A. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=t8O-4LKFVx.

Vovk, V. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vovk, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

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

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.

Vovk, V., Nouretdinov, I., Manokhin, V., and Gammerman, A. Cross-conformal predictive distributions. In *conformal and probabilistic prediction and applications*, pp. 37–51. PMLR, 2018.

Xie, R., Barber, R. F., and Candès, E. J. Boosted conformal prediction intervals, 2024. URL https://arxiv.org/abs/2406.07449.

Yang, Y. and Kuchibhotla, A. K. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, 120(549):435–447, 2025.

A. Related Work

Comparison to Lei et al. (2013) The influential work of Lei et al. (2013) gave the first theoretical guarantees of volume control or optimality to the best of our knowledge. In fact Lei et al. (2013) and subsequent follow-up works including Sadinle et al. (2016); Chernozhukov et al. (2021) with theoretical guarantees on volume control study a stricter quantity that corresponds to the volume of set difference $\mathrm{vol}\left(\widehat{C}\Delta C_{\mathrm{opt}}\right)$ (Lei et al., 2013; Sadinle et al., 2016; Chernozhukov et al., 2021). However, this much stronger notion requires that the optimal solution C_{opt} must not only exist but also be unique. Usually additional assumptions need to be imposed in the neighborhood of the boundary of C_{opt} in order that the set difference vanishes in the large sample limit. Specifically, the work of (Lei et al., 2013) assumes that the density is smooth, and in addition is strictly increasing or decreasing significantly. In comparison, our notion of volume optimality only requires the volume to be controlled, which can be achieved even if $\widehat{C}$ is not close to C_{opt}, or when C_{opt} does not even exist. Moreover, we do not need to make any assumptions on the smoothness of the density. In fact, the density may not even exist, and can have discrete point masses or δ functions as shown in the experiments. Indeed, from a practical point of view, any set with coverage and volume control would serve the purpose of valid prediction. Insisting the closeness to a questionable target C_{opt} comes at the cost of unnecessary assumptions on the data generating process.

Comparison to Izbicki et al. (2020; 2022) The work of Izbicki et al. (2020; 2022) provided conformal prediction methods that can produce a union of intervals in a supervised setting. Specifically, their methods, CD-split and HPD-split, are designed to leverage level sets of an estimated conditional density function. CD-split achieves local and marginal validity by partitioning the feature space adaptively but does not guarantee conditional coverage in general. In contrast, HPD-split simplifies tuning by using a conformity score based on the cumulative distribution function of the conditional density. Under certain assumptions of density estimation accuracy and the uniqueness of the optimal solution, HPD-split achieves asymptotic conditional coverage and converges to the highest predictive density set which is the smallest volume set with the desired coverage. In comparison, our method outputs a union of intervals with the smallest length from a direct estimator of the conditional CDF, which only requires the accuracy of conditional CDF estimation. Estimating the conditional CDF is statistically simpler than estimating the conditional density, which usually requires additional smoothness or regularity conditions.

Comparison to Kiyani et al. (2024) In very recent independent work, Kiyani et al. (2024) considered a min-max approach for conformal prediction in the covariate shift setting with a view towards length optimality of their intervals. They proposed a new method based on minimax optimization to optimize the average volume of prediction sets in the context of covariate shift, which generalizes the marginal or group-conditional coverage setting. Their method uses a given (predefined) conformity score, and optimizes the choice of the thresholds $h(X)$ for different covariates $X \in \mathcal{X}$ to minimize the average prediction interval length, while maintaining the marginal or group-conditional coverage. Under certain assumptions that the conformity score is consistent with a volume optimal prediction set, they show that solving their minimax optimization will give a volume-optimal solution. However the problem of finding the best threshold function $h(X)$ is a non-convex problem that may be computational inefficient in theory; but they use SGD to find a good heuristic solution in practice. This work is incomparable to this paper in multiple ways. While Kiyani et al. (2024) considers the covariate shift setting with a specific focus on marginal coverage and group coverage, we focus more on the unlabeled setting, and the conditional coverage setting of Chernozhukov et al. (2021). In contrast to their method that uses an off-the-shelf conformity score (and optimizes the thresholds $h(X)$), our method introduces a new conformity score function based on dynamic programming to find volume-optimal unions of intervals. This also suggests that our methods and the methods of Kiyani et al. (2024) may potentially be complementary. Finally, by restricting the prediction sets to unions of k intervals, we got theoretical guarantees of volume optimality and get polynomial time algorithms based on dynamic programming to achieve them. Hence, while both their work and our work try to address the important consideration of volume optimality, they are incomparable in terms of the setting, the results and the techniques.

Other Related work In the non-conformal setting, the work of (Scott & Nowak, 2006) studied the problem of finding minimal volume sets from a certain set family given samples drawn i.i.d from a distribution, with at least $1-\alpha$ fraction of probability mass. However this work mostly focused on statistical efficiency, and did not consider the conformal inference setting. In the past few years, there has been an explosion of literature in conformal inference that develops new conformal methods for various settings (see e.g., Barber et al., 2021; Stutz et al., 2022; Kumar et al., 2023; Barber et al., 2023; Xie et al., 2024, and references therein).

While our work is focused on the framework split conformal prediction, one interesting future extension is to consider various other frameworks, especially the class of cross-validation-style aggregation of split conformal sets (Vovk, 2015; Vovk et al., 2018; Barber et al., 2021; Kim et al., 2020). Another related direction to volume optimality is model selection or prediction set selection based on efficiency (Liang et al., 2024; Yang & Kuchibhotla, 2025).

B. Additional Numerical Experiments

B.1. Construction of Nested Systems

Recall the construction of the nested system described in Section 4.1. It immediately follows Proposition 2.3 that the construction satisfies Assumption 2.4 in the unsupervised setting.

In the supervised setting, the construction of the nested system is based on $\widehat{F}(y \mid x)$. For each $x \in \{X_{n+1}, \cdots, X_{2n+1}\}$, we generate $Y_1(x), \cdots, Y_L(x)$ according to the quantile level

$$Y_l(x) = \text{argmax}\{y : \widehat{F}(y \mid x) \le l/L\}.$$

Then, the greedy expansion and contraction procedure described in Section 4.1 is applied on $Y_1(x), \cdots, Y_L(x)$. Effectively, this is equivalent to using $\widetilde{F}(y \mid x) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{I}\{Y_l(x) \le y\}$ as input. By its definition, $\widetilde{F}(y \mid x)$ is a uniform approximation to $\widehat{F}(y \mid x)$ with error $1/L$. Thus, Assumption 3.1 is still satisfied for $\widetilde{F}(y \mid x)$. In all of our experiment, we set $L = m$.

Remark B.1. It is clear that the details of the greedy expansion step and the greedy contraction step do not matter much for Assumption 2.4 or 3.1 to be satisfied. However, different choices will indeed affect practical performance, especially in the supervised setting when $\widehat{F}(y \mid x)$ is not close to $F(y \mid x)$. To be more specific, sensible choices of expansion and contraction sets from the S_{j^*} generated by DP will serve as a safety net against model misspecification. We discuss this in more detail in Section B.3, see e.g. Figure 17.

B.2. Unsupervised Setting

Given i.i.d. observations $Y_1, Y_2, \ldots, Y_{2n} \in \mathbb{R}$ drawn from some distribution P, the goal is to find a prediction set $\widehat{C} = \widehat{C}(Y_1, \ldots, Y_{2n})$ such that $\mathbb{P}(Y_{2n+1} \in \widehat{C}) \ge 1 - \alpha$ for an independent future observation Y_{2n+1} drawn from the same P. We implement the proposed conformalized dynamic programming (DP) method $\widehat{C}_{\text{CP}-\text{DP}}$, and compare it with the conformalized kernel density estimation (KDE) proposed by (Lei et al., 2013) on the following synthetic datasets: (1) Gaussian; (2) Censored Gaussian; (3) Mixture of Gaussians; (4) ReLU-Transformed Gaussian.

Though the original conformalized KDE was proposed in the full conformal framework, we will consider its split conformal version for a direct comparison. We believe the comparison between the full conformal versions of the two methods will lead to the same conclusion. For the conformalized DP method, the conformity score is constructed based on the nested system described in Section B.1 with $m = 50$ and $\delta = \sqrt{(k + \log n)/n}$. The conformalized KDE is also in the form of (9), with the conformity score given by

$$q_{\text{KDE}}(x) = \frac{1}{n\rho}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{\rho}\right),$$

where $K(\cdot)$ is the standard Gaussian kernel and ρ is the bandwidth parameter. Both methods involve a single tuning parameter, k for conformalized DP and ρ for conformalized KDE.

Gaussian: Our first distribution is $P = N(0, 1)$, which is a benign example for sanity check. We consider sample size being 100, and set the coverage probability $1 - \alpha = 30\%$ for a more transparent comparison between the two methods. The conformalized DP is computed with number of intervals k ranging from 1 to 10. It turns out that the output of the prediction set is quite stable when k varies (Figure 5). Even for $k = 10$, our method still produces a single interval in this unimodal distribution.

The conformalized KDE is implemented with bandwidth ρ ranging from 0.001 to 0.005. We observe that the quality of the prediction set is quite sensitive to the choice of the bandwidth. As is shown by Figure 4, if the bandwidth of KDE is too small, the conformal prediction will output almost the entire support of the data set. This is because if the KDE overfits the training samples, the level set of the KDE will likely not cover the future observation. Therefore, a conformal procedure, which guarantees finite sample coverage, has to be conservative by outputting the entire support. Figure 5 shows that this issue will be alleviated as the bandwidth gets larger.

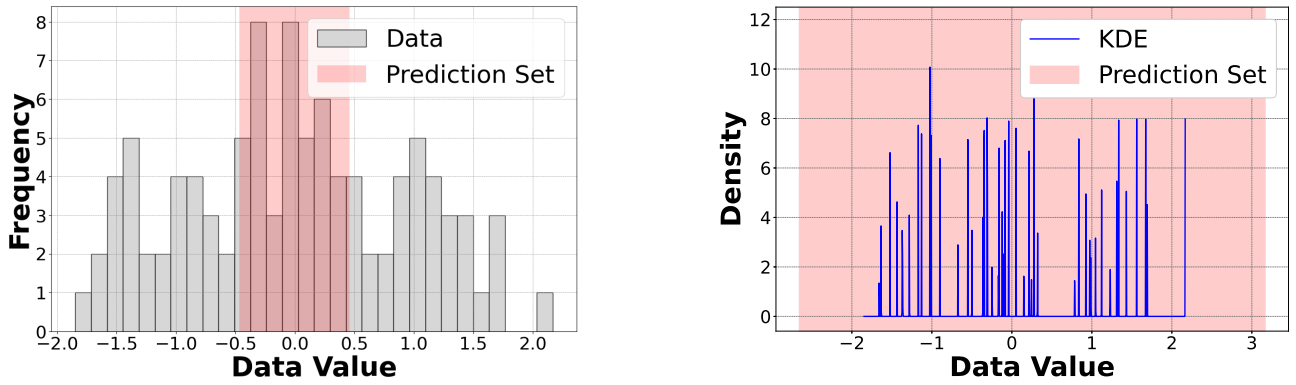


Figure 4. Conformal prediction sets on the Gaussian dataset. The left plot shows the histogram of the dataset and the prediction set produced by conformalized DP with $k = 1$; the right plot shows the kernel density estimation with bandwidth $\rho = 0.001$ and the prediction set given by the conformalized KDE.

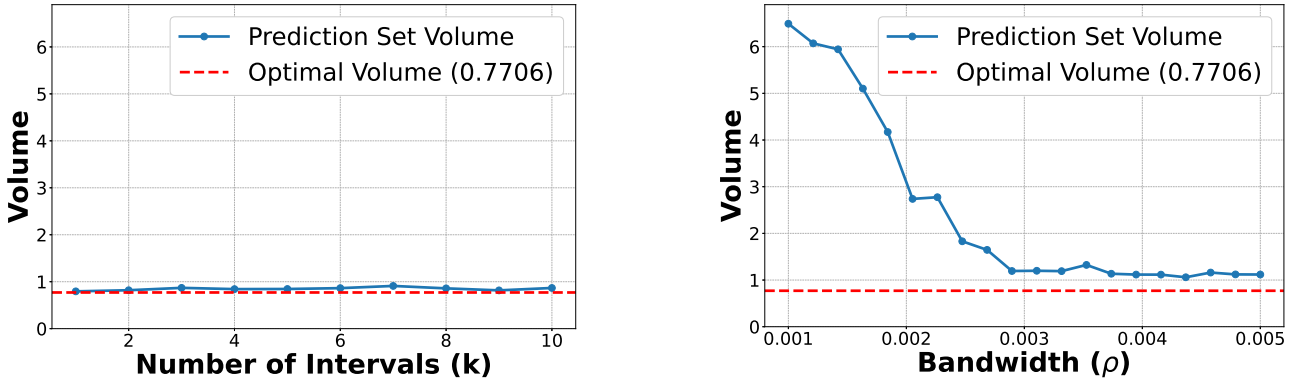


Figure 5. Volumes of prediction sets of the two methods on the Gaussian dataset (blue) and the benchmark $\mathrm{OPT}_1(N(0,1), 0.3) = 0.7706$ (red). The blue curves are computed by averaging 100 independent experiments.

Censored Gaussian: We next consider P being a censored Gaussian distribution. We take the sample size to be 100, and each sample can be generated according to $Y_i = \sigma(Z_i + 1) - \sigma(Z_i - 1)$ with $Z_i \sim N(0, 1)$ and $\sigma(t) = \max(t, 1)$ being the ReLU transform. This is equivalently a truncated Gaussian distribution, which has a standard Gaussian density on $(0, 2)$ and a point mass at 0 with probability $\mathbb{P}(Z_i \le -1)$ and another point mass at 2 with probability $\mathbb{P}(Z_i \ge 1)$. Again, for the sake of comparison, we set the coverage probability to be $1 - \alpha = 30\%$.

Since $\mathbb{P}(|Z_i| \le -1) \ge 1 - \alpha$, the population optimal volume is $\mathrm{OPT}(P, 0.3) = \mathrm{OPT}_2(P, 0.3) = 0$ due to the point masses at $\{0, 2\}$. By setting $k = 2$ for the conformalized DP procedure, the prediction set concentrates on the two point masses (Figure 6). Moreover, it produces very similar results as we increase k up to 10. Figure 7 shows that the only exception is $k = 1$, since one short interval obviosly cannot cover two points that are far away from each other.

We also run conformalized KDE with bandwidth ρ ranging from 0.001 to 1. Since the distribution does not even have a density function on the entire support, KDE is not really suitable for this setting. Not surprisingly, for a typical choice of bandwidth that is not too small, the conformalized KDE will not identify the two point masses due to smoothing (Figure 6). Figure 7 reports the volume of the prediction set as we vary bandwidth, and the volume of the prediction set is close to optimal only when the bandwidth is extremely close to 0.

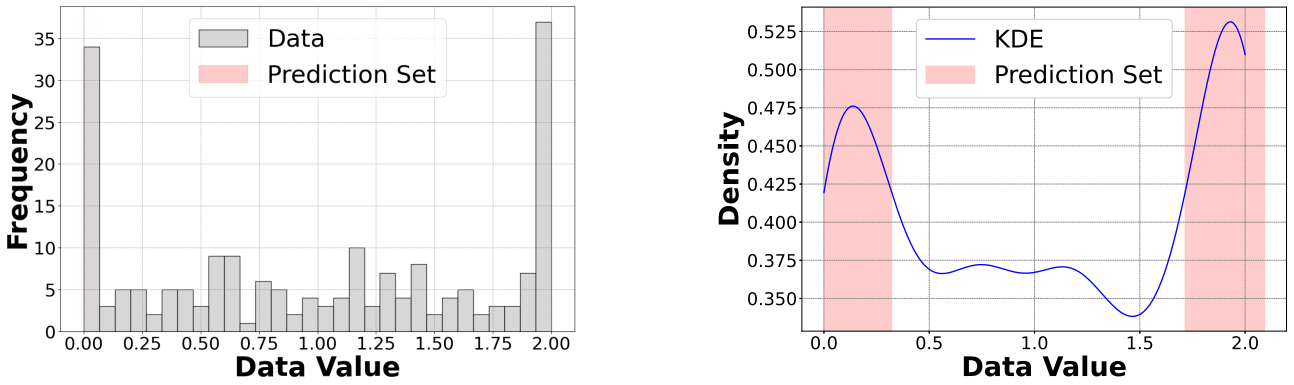


Figure 6. Conformal prediction sets on the censored Gaussian dataset. The left plot shows the histogram of the dataset and the prediction set given by conformalized DP with $k = 2$ intervals (The prediction set is two zero-length intervals at 0.0 and 2.0). The right plot shows the kernel density estimation with bandwidth $\rho = 0.2$ and the prediction set by conformalized KDE.

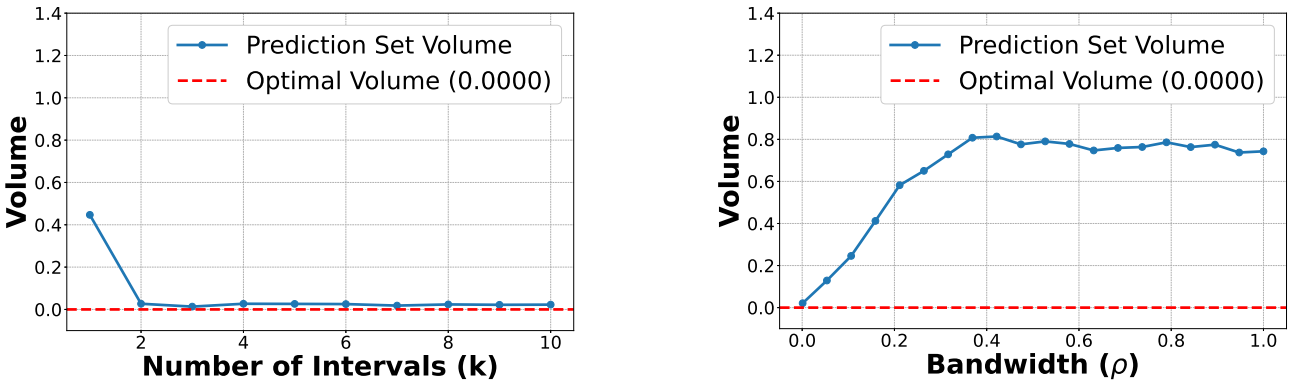
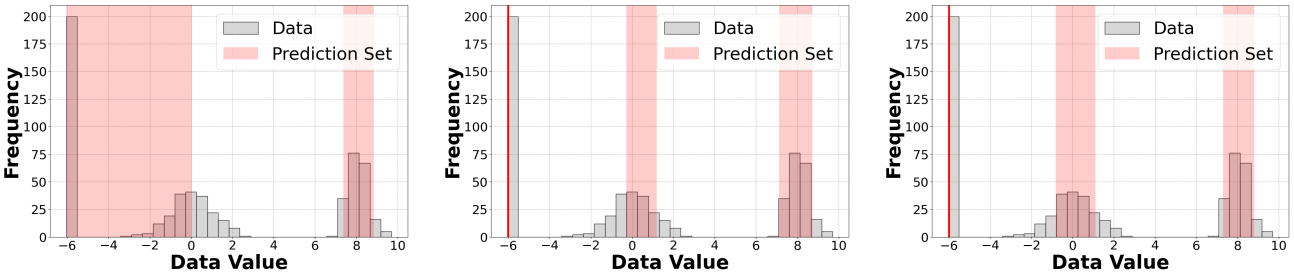


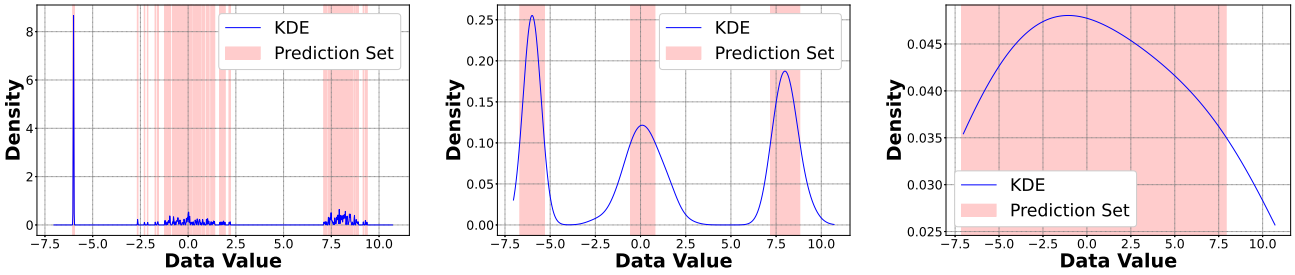
Figure 7. Volumes of prediction sets of the two methods on the censored Gaussian dataset (blue) and the optimal volume (red). The blue curves are computed by averaging 100 independent experiments.

Mixture of Gaussians: In this experiment, we consider $P = \frac{1}{3}N(-6, 0.0001) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(8, 0.25)$. The sample size and coverage probability are set as 600 and $1 - \alpha = 80\%$, respectively. The two methods are compared with k ranging from 1 to 10 in conformalized DP and bandwidth ρ ranging from 0.001 to 5 in conformalized KDE.



(a) $k = 2$ **(b)** $k = 3$ **(c)** $k = 6$

Figure 8. Prediction sets provided by the conformalized DP method with the number of intervals $k = 2, 3, 6$ on the mixture of Gaussians dataset.



(a) $\rho = 0.01$ **(b)** $\rho = 0.5$ **(c)** $\rho = 5.0$

Figure 9. Prediction sets provided by the conformalized KDE using bandwidth $\rho = 0.01, 0.5, 5.0$ on the mixture of Gaussians dataset.

We report typical results of conformalized DP with $k \in \{2, 3, 6\}$ in Figure 8 and report those of conformalized KDE with $\rho \in \{0.01, 0.5, 5.0\}$ in Figure 9. The proposed method based on DP produces similar prediction sets close to optimal as long as $k \geq 3$ (Figure 10). This is because $\mathrm{OPT}(P, 0.8) = \mathrm{OPT}_3(P, 0.8)$ with P being a Gaussian mixture of three components. In comparison, the results based on KDE are quite sensitive to the bandwidth choice, since different bandwidths lead to kernel density estimators with completely different numbers of modes. Figure 10 shows that for the optimal choice of bandwidth around 0.5, the KDE successfully identifies the three modes of the Gaussian mixture. However, even with the optimal bandwidth, the volume of the prediction set is in general still greater than that of the conformalized DP. This is partly because the three components of the Gaussian mixture do not have the same variance parameters, and thus cannot be optimally estimated by KDE with a single bandwidth.

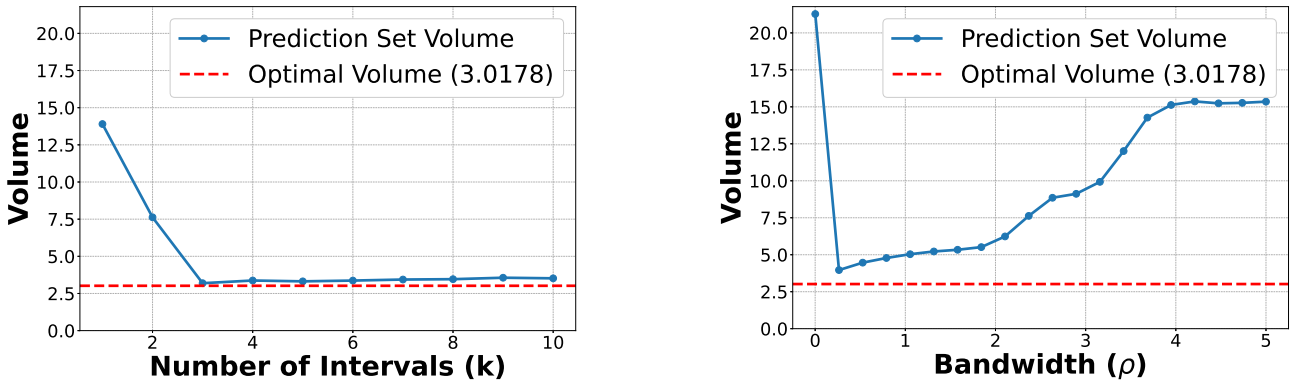
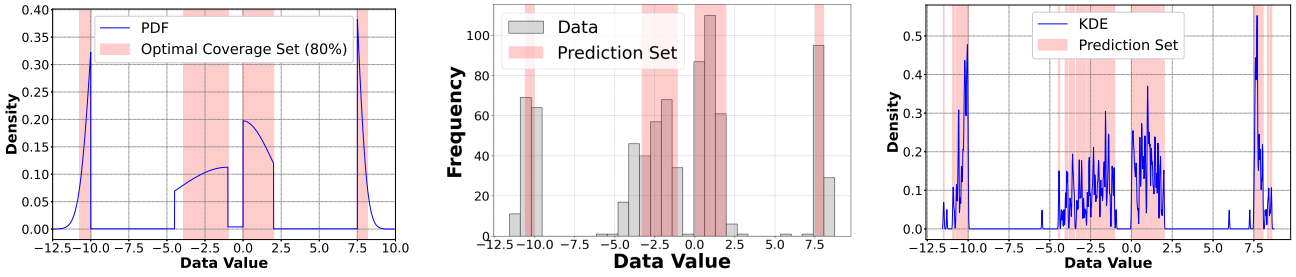


Figure 10. Volumes of prediction sets of the two methods on the mixture of Gaussians dataset (blue) and the optimal volume $\mathrm{OPT}(P, 0.8) = 3.0178$ (red). The blue curves are computed by averaging 100 independent experiments.

ReLU-Transformed Gaussian: The ReLU-Transformed Gaussian is generated according to $X_i = \sum_{j=1}^{t} a_j * \sigma(w_j * Z_i + b_j)$ with $Z_i \sim N(0, 1)$. It includes the censored Gaussian as a special case. Here, we take $t = 7$ and take a randomly generated set of coefficients. The resulting density function is plot in Figure 11 (a). The sample size and coverage probability are taken as 600 and $1 - \alpha = 80\%$, respectively.

Figure 11 also shows a typical prediction set produced by conformalized DP with $k = 4$ and one produced by conformalized KDE with bandwidth $\rho = 0.02$. Figure 12 gives a more thorough comparison. The proposed conformalized DP achieves near optimality when $k \geq 4$, since the distribution has 4 modes. The KDE solutions are sensitive to the choice of bandwidth for this complicated distribution.



(a) Optimal Coverage Set **(b)** Conformalized DP with $k = 4$ **(c)** Conformalized KDE with $\rho = 0.02$

Figure 11. (a) Density of The ReLU-Transformed Gaussian and prediction set with optimal volume; (b) Conformalized DP with $k = 4$; (c) Conformalized KDE with $\rho = 0.02$.

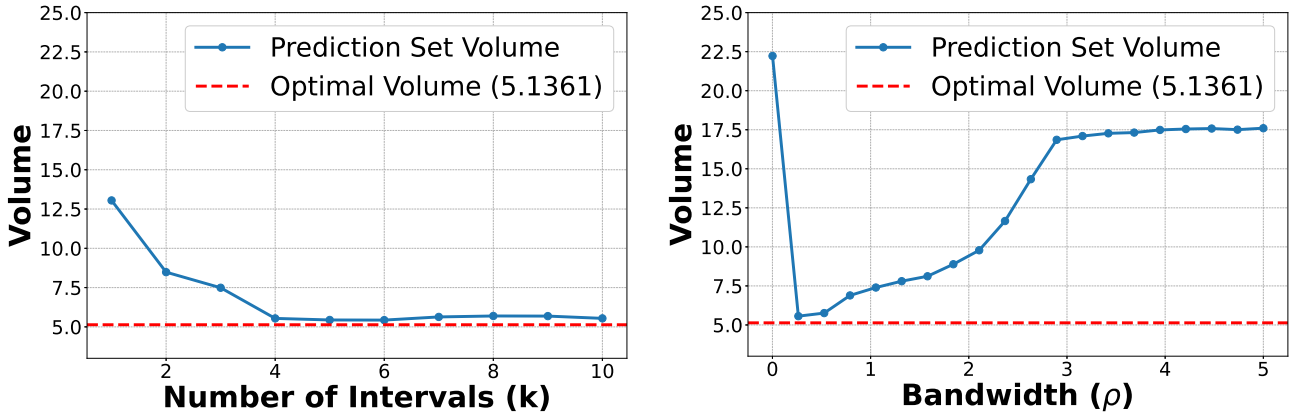


Figure 12. Volumes of prediction sets of the two methods on the ReLU-transformed Gaussian dataset (blue) and the optimal volume $\mathrm{OPT}(P, 0.8) = 5.1361$ (red). The blue curves are computed by averaging 100 independent experiments.

Effects of Sample Sizes and Coverage Probabilites: Finally, we study the effects of sample sizes and coverage probabilities for the two methods. Specifically, we examine how the volume of the prediction set decays as the sample size increases and how it varies with different coverage probabilities. The experiments will be conducted with data generated from the following two distributions:

1. $\frac{1}{3}N(-6, 0.0001) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(8, 0.25)$.

2. The ReLU-Transformed Gaussian $Y_i = \sum_{j=1}^{t} a_j * \sigma(w_j * Z_i + b_j)$ with $Z_i \sim N(0, 1)$, $t = 7$ and coefficients are the same as in the previous experiment.

For conformalized DP, we will set $k = 3$ for the first distribution and $k = 4$ for the second one to match the number of modes in the two cases. For conformalized KDE, since the method is sensitive to the choice of bandwidth, we will scan the bandwidth ρ from 0.001 to 0.2, and only report the one with the smallest volume. We also benchmark the performances of the two methods by the optimal volume and a standard split conformal procedure with conformity score $q_{\text{standard}}(y) = -\left|y - \frac{1}{n}\sum_{i=1}^{n} Y_i\right|$.

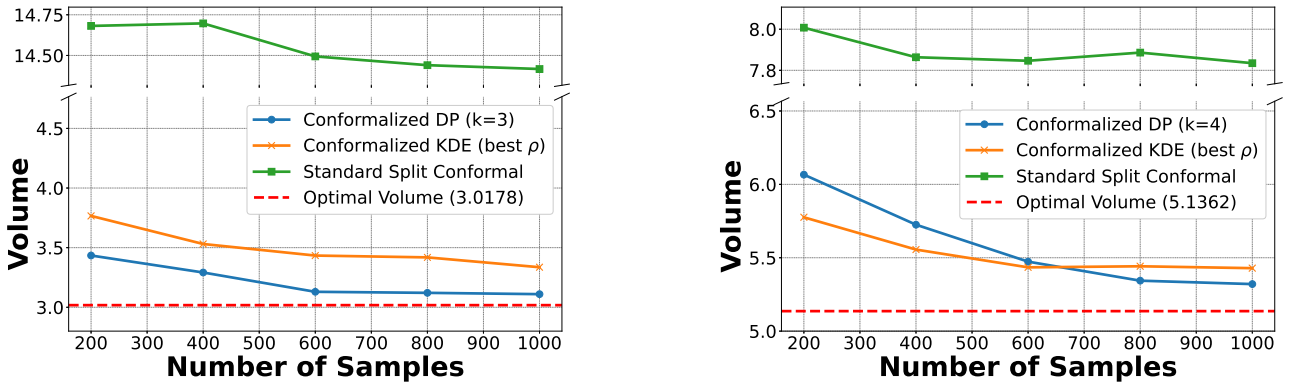


Figure 13. Volume of prediction set against sample size. Left: $\frac{1}{3}N(-6, 0.0001) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(8, 0.25)$. Right: ReLU-Transformed Gaussian. All curves are plotted by averaging results from 100 independent experiments.

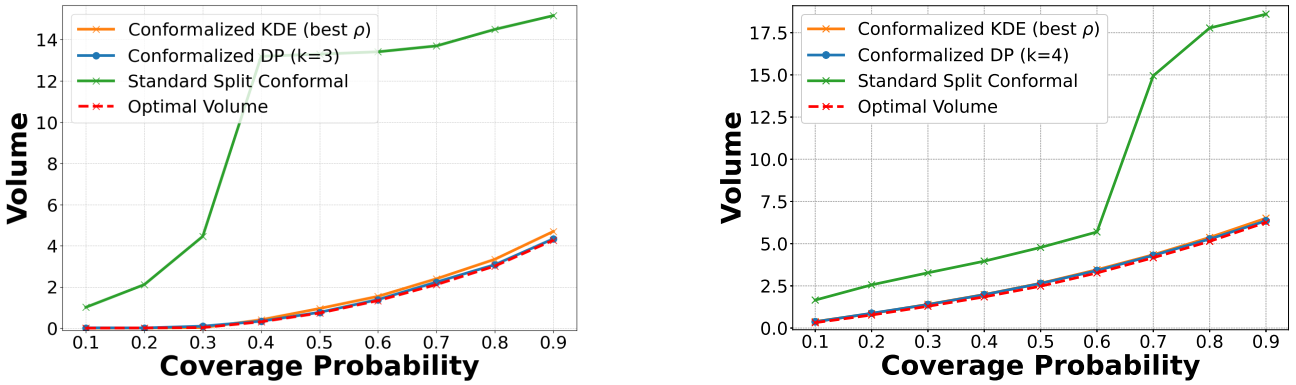


Figure 14. Volume of prediction set against coverage probability. Left: $\frac{1}{3}N(-6, 0.0001) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(8, 0.25)$. Right: ReLU-Transformed Gaussian. All curves are plotted by averaging results from 100 independent experiments.

Figure 13 shows the results with sample size ranging from 200 to 1000 with the coverage probability fixed by $1 - \alpha = 80\%$. Both conformalized DP and conformalized KDE produce smaller prediction sets as sample size increase. Even with the bandwidth optimally tuned for conformalized KDE, which is not feasible in practice, the proposed conformalized DP tends to achieve smaller volumes in most cases. In setting of the ReLU-Transformed Gaussian, we observe that the volume of conformalized KDE prediction set barely decreases after sample size 600, since in this case density estimation is very hard for KDE.

Figure 14 considers coverage probability ranging from 0.1 to 0.9, with sample size fixed at 600. The conformalized DP constantly achieves smaller volume than the conformalized KDE even though the later is computed with optimally tuned bandwidth. This demonstrates the robustness of the conformalized DP in handling varying coverage requirements while maintaining efficiency in volume.

Real-World Dataset: We compare the performance of conformalized KDE and conformalized dynamic programming (DP) methods on the Enzyme dataset in Table 7 and 8. For conformalized KDE, we report results across a range of bandwidths from 0.1 to 0.9; for conformalized DP, we vary the number of intervals $k = 1, \cdots, 5$. We output the average and standard deviation for volume and empirical coverage over 50 trials and 80% target coverage. We observe that conformalized DP achieves higher coverage with lower volume, indicating more efficient coverage compared to conformalized KDE.

Table 7. Conformalized KDE on Enzyme Dataset

Bandwidth	Volume	Coverage
0.1	**0.9826 ± 0.0969**	0.8081 ± 0.0217
0.3	1.4640 ± 0.0940	0.8056 ± 0.0231
0.5	1.5218 ± 0.1332	0.7999 ± 0.0268
0.7	1.4526 ± 0.1809	0.7962 ± 0.0272
0.9	1.3867 ± 0.1799	0.7971 ± 0.0236

Table 8. Conformalized DP on Enzyme Dataset

k	Volume	Coverage
1	1.2207 ± 0.1188	0.8101 ± 0.0461
2	**0.8997 ± 0.1414**	0.8144 ± 0.0353
3	1.0118 ± 0.1975	0.8416 ± 0.0390
4	1.0131 ± 0.1878	0.8458 ± 0.0379
5	1.0520 ± 0.2002	0.8558 ± 0.0353

B.3. Supervised Setting

In the supervised setting, we validate our results on the simulated datasets in Romano et al. (2019) and Izbicki et al. (2020). We compare against the methods of Conformalized Quantile Regression (CQR) of Romano et al. (2019) and Distributional Conformal Prediction (DCP) of Chernozhukov et al. (2021) and CD-split and HPD-split of Izbicki et al. (2022).

Simulated Dataset (Romano et al., 2019): We first describe the simulated dataset in Romano et al. (2019). In this data, each one-dimensional predictor variable X_i is sampled uniformly from the range $[0, 5]$. The response variable is then sampled according to

$$Y_i \sim \text{Pois}(\sin^2(X_i) + 0.1) + 0.03\, X_i\, \varepsilon_{1,i} + 25\, \mathbf{1}\{U_i < 0.01\}\, \varepsilon_{2,i},$$

where $\text{Pois}(\lambda)$ is the Poisson distribution with mean λ, $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ are independent standard Gaussian noise, and U_i is drawn uniformly on the interval $[0, 1]$. The first component of the distribution, $\text{Pois}(\sin^2(X_i)+0.1)$, generates a distribution that is clustered around positive integer values of Y, with variance that changes periodically in X. The second component of the distribution, $0.03\, X_i\, \varepsilon_{1,i}$, adds some additional variance to each of the integer centered clusters, where the magnitude of the variance increases with X. The final component, $25\, \mathbf{1}\{U_i < 0.01\}\, \varepsilon_{2,i}$, adds a small fraction of outliers to the distribution. We generate 2000 training examples, and 5000 test examples, as in the work of (Romano et al., 2019). The same subset of training and test examples are used in the illustration of each of these methods. The set of test examples is visualized in Figure 15, with the full range of Y values including the outliers. In the plots associated with our conformal output, we zoom in on the Y axis for readability, leaving the outliers off the chart.

Simulated Dataset (Izbicki et al., 2020): We now describe the simulated dataset in Izbicki et al. (2020). In this data, the predictor variables $X = (X_1, \ldots, X_d)$ with $d = 20$ dimensions are independently and uniformly sampled from the range $[-1.5, 1.5]$. The response variable Y is then generated according to the following bimodal conditional distribution:

$$Y \mid X \sim 0.5\mathcal{N}(f(X) - g(X), \sigma^2(X)) + 0.5\mathcal{N}(f(X) + g(X), \sigma^2(X)).$$

where the functions $f(X)$, $g(X)$, and $\sigma^2(X)$ are defined as:

$$f(X) = (X_1 - 1)^2(X_1 + 1), \quad g(X) = 2\mathbb{I}(X_1 \geq -0.5)\sqrt{X_1 + 0.5}, \quad \sigma^2(X) = \frac{1}{4} + |X_1|.$$

Here, $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2, and the indicator function $\mathbb{I}(X_1 \geq -0.5)$ accounts for the bimodal nature of the data, introducing a piecewise behavior in the response variable. The first term $f(X)$

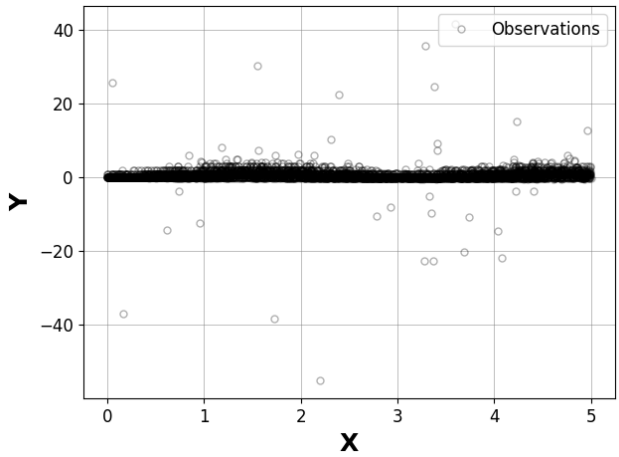


Figure 15. Simulated data of Romano et al. (2019), including outliers.

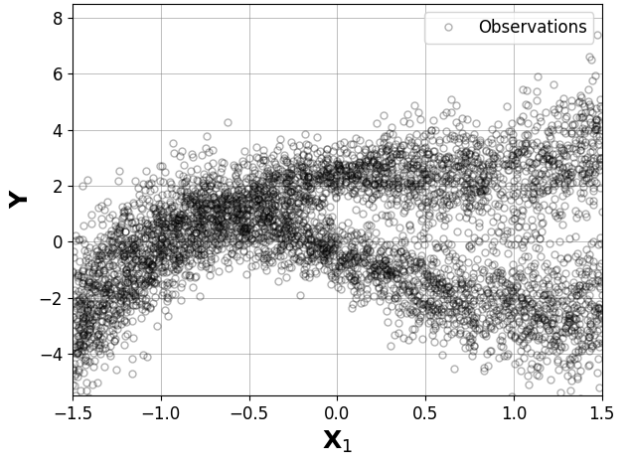


Figure 16. Simulated data from Izbicki et al. (2020), illustrating the bimodal distribution of the response variable.

captures a polynomial relationship with X_1, while the second term $g(X)$ introduces an asymmetric bimodal effect depending on the value of X_1. The variance $\sigma^2(X)$ increases linearly with $|X_1|$, adding heteroscedasticity to the distribution.

We generate 2000 training examples and 5000 test examples. The same training and test sets are used consistently across all experiments to ensure reproducibility. The test set is visualized in Figure 16, showcasing the full range of Y values, including the effects of bimodality and variance heterogeneity.

Methods: We compare our conformalized DP with the following methods: Conformalized Quantile Regression (CQR) of Romano et al. (2019) and Distributional Conformal Prediction via Quantile Regression (DCP-QR) and Optimal Distributional Conformal Prediction via Quantile Regression (DCP-QR*) of Chernozhukov et al. (2021) and CD-split and HPD-split of Izbicki et al. (2022).

We now describe the implementation of these methods. The compared methods, CQR, DCP-QR, DCP-QR*, and our conformalized DP rely on quantile regression. For simulated dataset (Romano et al., 2019) with single dimensional predictor variable, we use the package `sklearn-quantile` (Roebroek, 2023) to implement the quantile regression, which implements the method of Quantile Regression Forests, due to Meinshausen (2006). The CD-Split and HPD-Split methods require the conditional density estimation, which is achieved by the R package `FlexCoDE` (Izbicki & Lee, 2017). For simulated dataset (Izbicki et al., 2020) with high dimensional predictor variables, the quantile regression by `sklearn-quantile` is not informative. For CQR, DCP-QR, DCP-QR*, and our conformalized DP, we first use the R

package `FlexCoDE` to generate the conditional density estimation and then integrate the conditional density estimation to get quantile regression and conditional CDF. The CD-Split and HPD-Split methods again use the conditional density estimation provided by `FlexCoDE`. All methods are implemented within the split conformal framework, where the training data is randomly divided into two equal parts. Specifically, half of the data is allocated for model training, while the remaining half is used as the calibration set to ensure valid coverage guarantees.

For convenience, we will refer to the qth estimated quantile of Y given $X = x$ as $\widehat{Q}(q, x)$. Some of the following methods use quantile regression to estimate the whole conditional c.d.f. of Y given X, by estimating a set of quantiles from a fine grid. This gives us an estimate of the conditional c.d.f., which gives us access to $\widehat{F}(y \mid x)$, the inverse of $\widehat{Q}$. (That is, $\widehat{F}(y \mid x) = q$, such that $y = \widehat{Q}(q, x)$. Since we only have $\widehat{Q}$ for values of q in the grid, we set $\widehat{F}(y \mid x)$ to be the smallest q in the grid such that $y \leq \widehat{Q}(q, x)$.)

- Conformalized Quantile Regression (CQR), (Romano et al., 2019): This method fits a model to two quantiles of the data, $q_{\text{low}} = \frac{\alpha}{2}$ and $q_{\text{high}} = 1 - \frac{\alpha}{2}$. On a new test example X_{test}, CQR uses the model to estimate the low and high quantile, and the conformal procedure will output the interval

$$\left[\widehat{Q}(q_{\text{low}}, X_{\text{test}}) - b,\ \widehat{Q}(q_{\text{high}}, X_{\text{test}}) + b\right],$$

  where b is a buffer value chosen in the calibration step of the conformal procedure to guarantee coverage.

- Distributional Conformal Prediction via Quantile Regression (DCP-QR), (Chernozhukov et al., 2021): In this framework, we assume access to a model $\widehat{F}$ that can estimate the conditional c.d.f. of the distribution of Y given X, which we estimate via quantile regression. Similar to CQR, we start with $q_{\text{low}} = \frac{\alpha}{2}$ and $q_{\text{high}} = 1 - \frac{\alpha}{2}$. In DCP, instead of adding the buffer in the Y space, the buffer is added in the quantile space. That is, on a new test example X_{test}, DCP will output the interval

$$\left[\widehat{Q}(q_{\text{low}} - b, X_{\text{test}}),\ \widehat{Q}(q_{\text{high}} + b, X_{\text{test}})\right],$$

  where b is a buffer value chosen in the calibration step of the conformal procedure to guarantee coverage.

- Optimal Distributional Conformal Prediction via Quantile Regression (DCP-QR*), (Chernozhukov et al., 2021): The optimal DCP is very similar to DCP, except that q_{low} and q_{high} need not be symmetric around the median ($q = \frac{1}{2}$). Instead, they are chosen to provide the minimum volume interval that achieves the desired coverage. We note that the buffer is still applied symmetrically in the quantile space. That is, the lower quantile is lowered by some value b, and the upper quantile is raised by the same value b.

- CD-Split (Izbicki et al., 2020; 2022): This method provides prediction sets based on the conditional density estimation and a partitioning of the feature space. The conformity score in CD-split is based on a conditional density estimator, which allows the method to approximate the highest predictive density (HPD) set. The feature space is partitioned based on the profile of the conditional density estimator, and the cut-off values are computed locally within each partition. This approach enables CD-split to achieve local and asymptotic conditional validity while providing more informative prediction sets, especially for multimodal distributions.

- HPD-Split (Izbicki et al., 2022): The HPD-split method outputs prediction sets based on the highest predictive density (HPD) sets of the conditional density estimation. Unlike CD-split, which partitions the feature space, HPD-Split uses the conformity score based on the conditional CDF of the condition density estimator. Since this conditional CDF is independent of the feature variable, HPD-Split does not require the partition of the feature space and tuning parameters for that as in CD-Split. When the conditional density estimation is accurate, HPD-Split converges to the highest predictive density (HPD) sets.

- Conformalized Dynamic Programming, $k = 1$ and $k = 5$: We implement a modification of the procedure described in this work. In the unsupervised setting, we described the dynamic programming procedure that outputs the minimum volume set of k intervals that contain a desired fraction of samples. In this setting, given a new test example X, we do not have access to samples. Instead, we have access to a grid of estimated quantiles of Y given X. We implement a version of the dynamic programming procedure that operates on this quantile grid instead of a set of points, to output the minimum volume set of k intervals that cover at least the desired probability mass. We can also modify our greedy contraction and expansion procedures to provide a nested system of sets for different coverage levels.

Discussion: The results of our experiments are illustrated in Figure 18. Our experiments show that Conformalized Quantile Regression (CQR) and Distributional Conformal Prediction via Quantile Regression (DCP-QR) perform approximately as well as each other on this dataset, achieving average volume 1.42 and 1.48 respectively (see Figures 18a, 18b).

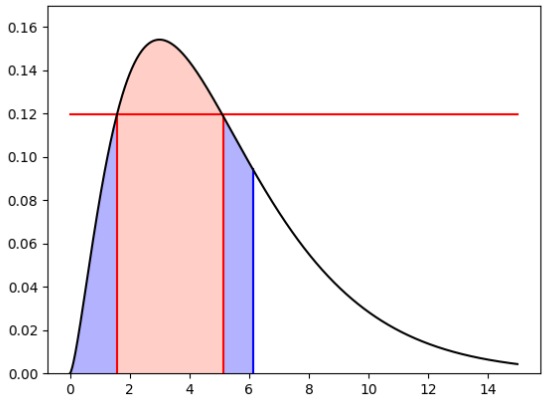
Optimal Distributional Conformal Prediction via Quantile Regression (DCP-QR*) achieves a significant improvement over DCP-QR on this data, achieving average volume 1.29 (see Figure 18c). This is due to the fact that the distribution of Y values is not symmetric around, or peaked at the median Y value. Thus, DCP-QR suffers a disadvantage, because it outputs intervals that are centered around the median in quantile space, and does not take into account the relative volumes of the quantiles in Y space. DCP-QR* on the other hand, is able to take advantage of the fact that, for this data, quantiles close to 0 have very low volume, and output intervals that use these quantiles.

While DCP-QR* uses information about the relative volume of the quantiles to choose q_{low} and q_{high}, which define the output intervals before conformalization, it does not take the volume into account during the conformalization step. Expanding the interval by a buffer value b that is small in quantile space, can lead to a large difference in Y space, increasing the volume of the output interval significantly. For example in Figure 18c, the intervals for X just larger than 4 stretch very far into the negative Y region, as a small adjustment in quantile space is a large adjustment in Y space.

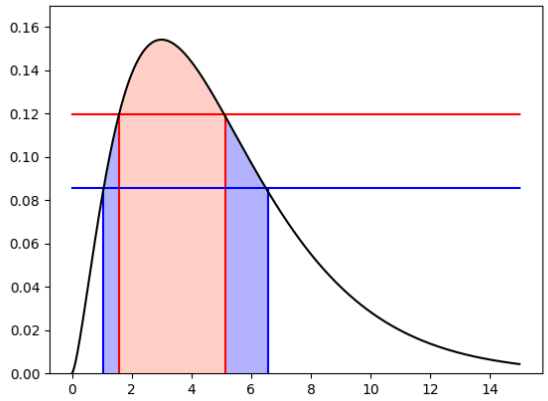
This issue is avoided by our Conformalized Dynamic Programming (Conformalized DP) method with greedy expansion and contraction for $k = 1$ interval. Before conformalization, the interval output by Conformalized DP and DCP-QR* is the same: it is the volume optimal interval that achieves a given coverage according to the estimated c.d.f.. However, our method takes the relative volume of different quantiles into account in the conformalization step, and avoids the issue of expanding the interval in quantile space in directions that add too much volume in Y space. This allows the method to achieve an improved average volume of 1.14 (see Figure 18f).

An illustration of this issue is given in Figure 17. Suppose that for a new test example X, the estimated conditional distribution of Y is skewed. (In this illustration it is $\chi^2(5)$.) Suppose that our target coverage was 0.5, and in the calibration phase we are required to expand coverage to 0.7. Both ConformalizedDP and DCP-QR* will start by calculating the minimum volume interval that captures 0.5 of the probability mass. In this case it is the red region from $x = 1.58$ to $x = 5.14$ (i.e., the set of x such that $f(x) > 0.12$, where $f(x)$ is the p.d.f. of the distribution). Then, each method must expand this interval to capture 0.7 of the probability mass. DCP-QR* does this by adding two blue regions, each of which capture an additional 0.1 probability mass. This results in expanding the interval significantly to the right, even though the density is low. ConformalizedDP takes the volume (i.e., density) into account when expanding the interval, and produces the minimum volume interval that captures 0.7 of the distribution (i.e., the set of x such that $f(x) > 0.085$), in this example. (We note that the expansion and contraction procedure of ConformalizedDP does not always result in the volume optimal prediction set for the adjusted coverage, only the original target coverage. However, in this case, since the distribution is unimodal and $k = 1$, we do indeed recover the volume optimal set even for the adjusted coverage.)

Finally, we also implement Conformalized DP with $k = 5$ intervals. This allows us to fit to the multimodal shape of the Y data, and achieve a much lower average volume of 0.45 (see Figure 18g).

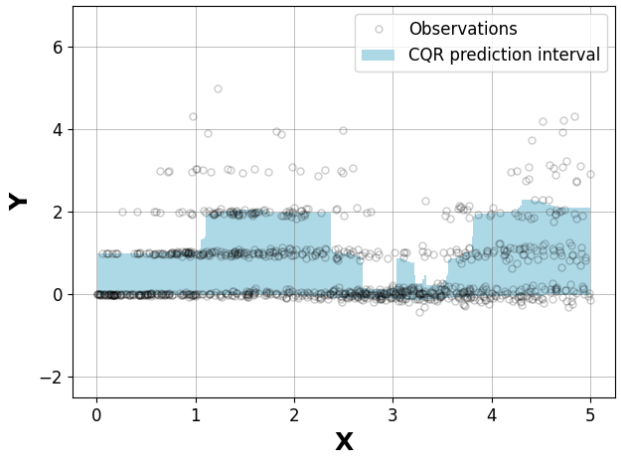


(a) When expanding from the red region, coverage 0.5, to a region of coverage 0.7, DCP-QR* chooses the blue region with additional volume 2.56.

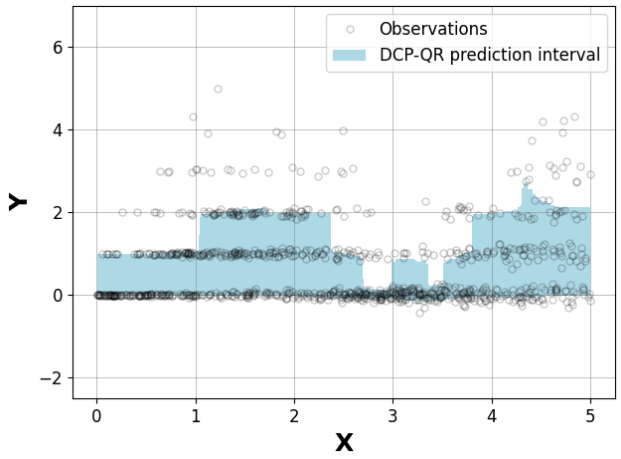


(b) When expanding from the red region, coverage 0.5, to a region of coverage 0.7, Conformalized DP chooses the blue region with additional volume 1.96.

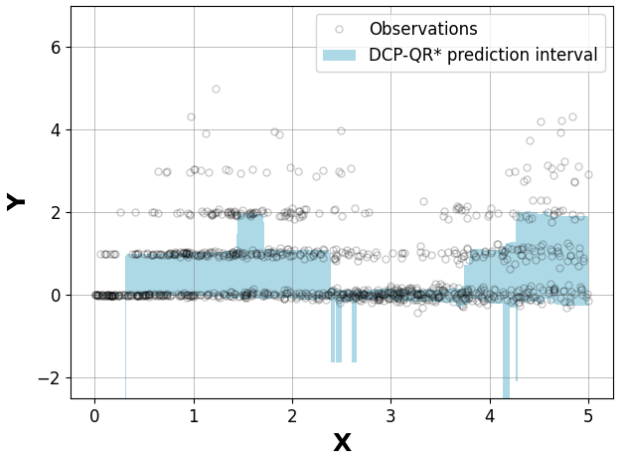
Figure 17. We illustrate the difference between DCP-QR* and Conformalized DP for $k = 1$, on the example where the estimated conditional distribution of Y for a new X_{test} is $\chi^2(5)$. We plot the intervals that are chosen by the methods against the p.d.f. of the estimated distribution.



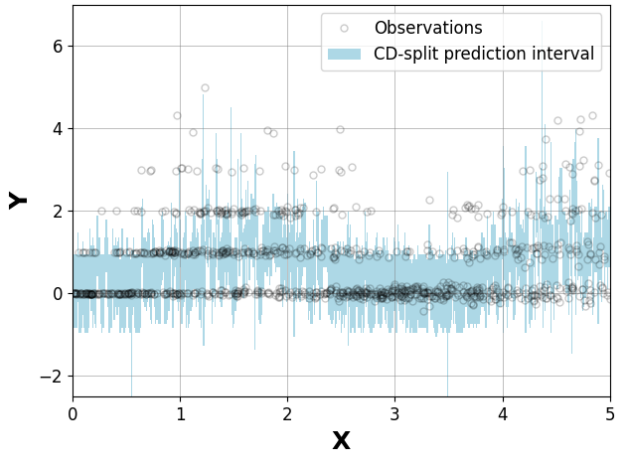
(a) Conformalized Quantile Regression (CQR), (Romano et al., 2019), achieves average volume 1.42 and empirical coverage 70.62%.



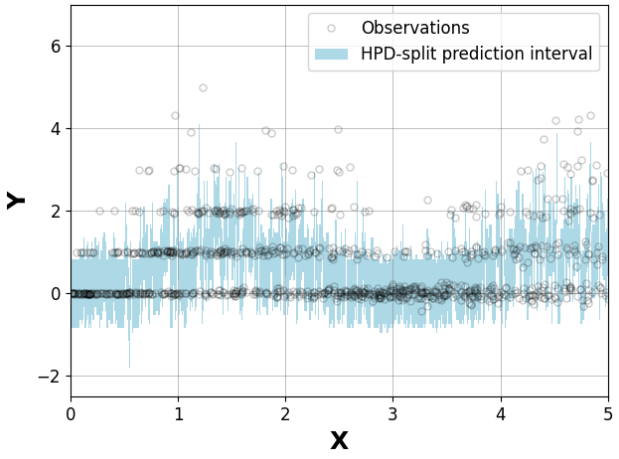
(b) Distributional Conformal Prediction (DCP), (Chernozhukov et al., 2021), achieves average volume 1.48 and empirical coverage 71.6%.



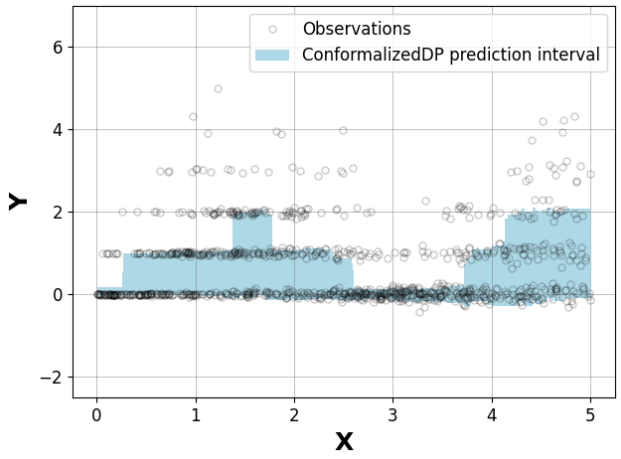
(c) Optimal Distributional Conformal Prediction (DCP-QR*), (Chernozhukov et al., 2021), achieves average volume 1.29 and empirical coverage 71.06%.



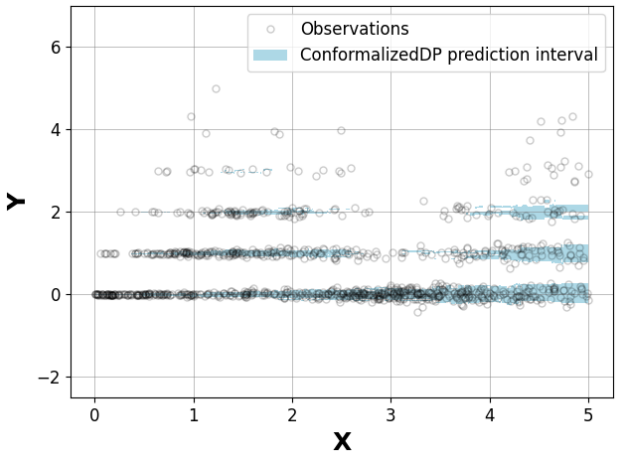
(d) CD-split Conformal Prediction, (Izbicki et al., 2022), achieves average volume 1.83 and empirical coverage 69.94%.



(e) HPD-split Conformal Prediction, (Izbicki et al., 2022), achieves average volume 1.75 and empirical coverage 69.44%.

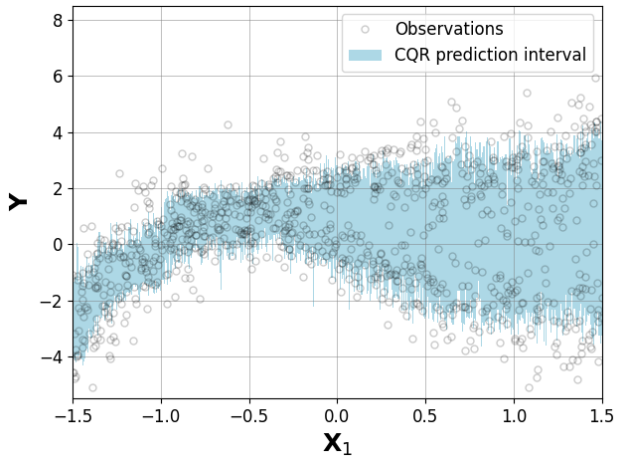


(f) Conformalized Dynamic Programming ($k = 1$), achieves average volume 1.14 and empirical coverage 74.04%.

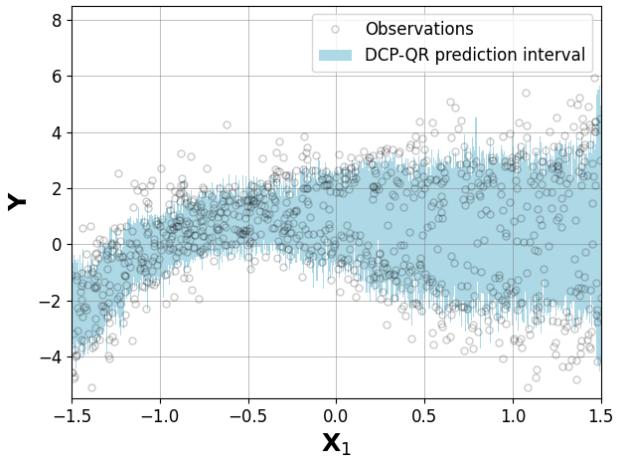


(g) Conformalized Dynamic Programming ($k = 5$), achieves average volume 0.45 and empirical coverage 72.36%.

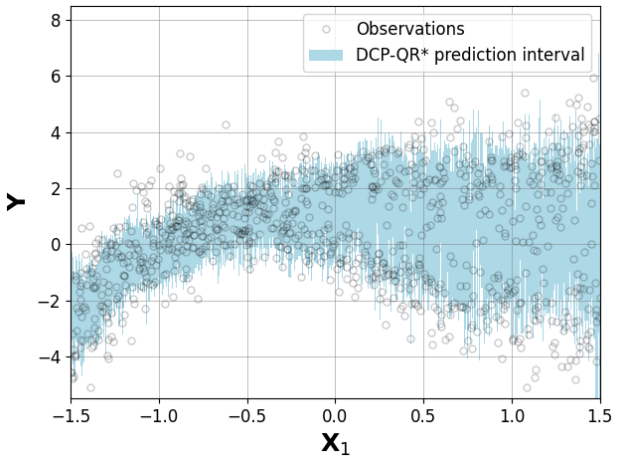
Figure 18. Comparison of supervised conformal prediction methods on simulated data from (Romano et al., 2019). All results are for a target coverage of 0.70.



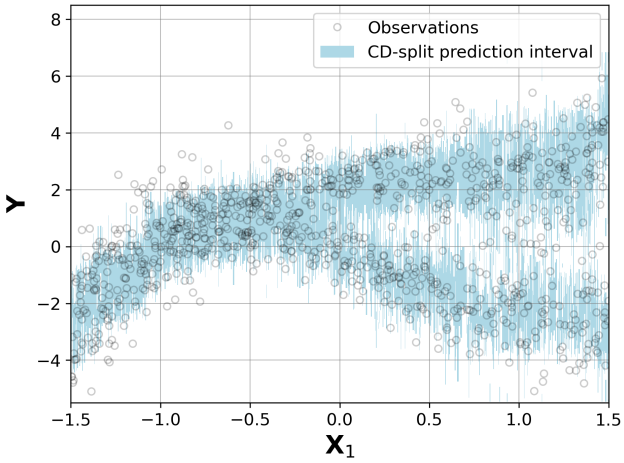
(a) CQR, achieves average volume 4.10 and empirical coverage 71.54%.



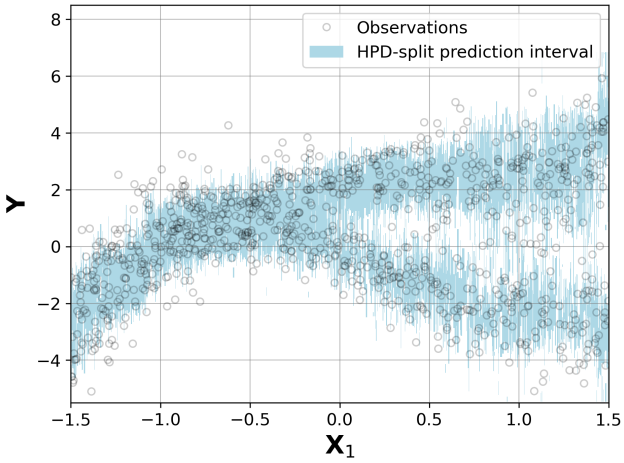
(b) DCP-QR, achieves average volume 4.04 and empirical coverage 70.85%.



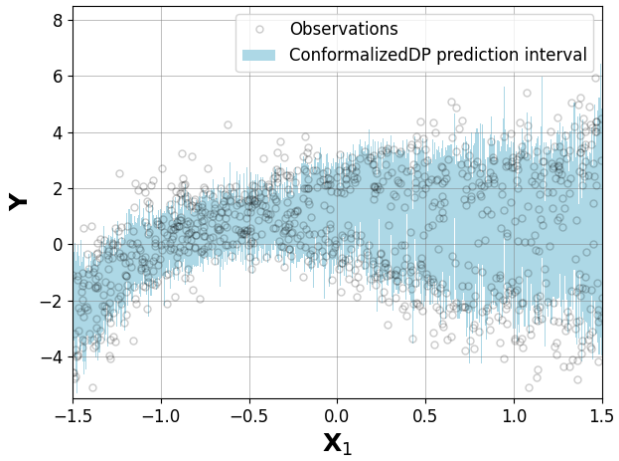
(c) DCP-QR*, achieves average volume 4.05 and empirical coverage 69.66%.



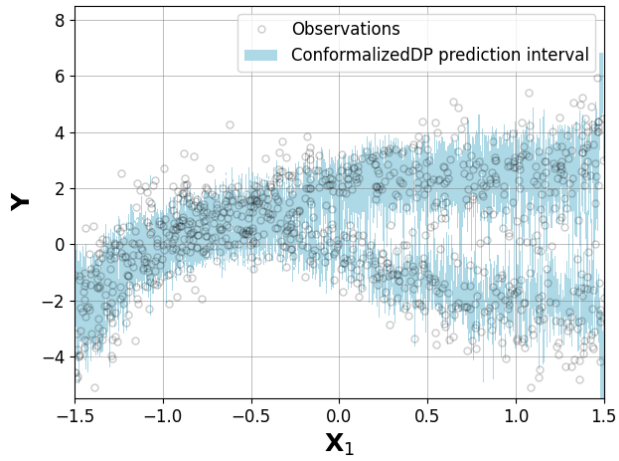
(d) CD-Split, achieves average volume 3.69 and empirical coverage 69.86%.



(e) HPD-Split, achieves average volume 3.60 and empirical coverage 69.64%.



(f) Conformalized Dynamic Programming ($k = 1$), achieves average volume 4.00 and empirical coverage 68.98%.



(g) Conformalized Dynamic Programming ($k = 2$), achieves average volume 3.55 and empirical coverage 69.42%.

Figure 19. Comparison of supervised conformal prediction methods on simulated data from (Izbicki et al., 2020). All results are for a target coverage of 0.70.

Real-World Dataset (Dhillon et al., 2024). We evaluate our method along with several baselines on two real-world regression datasets from the UCI repository: AirFoil and WineQuality. These two datasets were previously considered in the conformal prediction study by Dhillon et al. (2024). For each dataset, we compare the empirical coverage, average volume, and runtime of the prediction sets output by different methods, with a shared preprocessing time (for conditional CDF estimation using `FlexCoDE`) excluded from reported runtimes. For AirFoil and WineQuality, we repeat each experiment 20 times and report the mean and standard deviation of all results with target coverage 0.8.

Recall that our method for the supervised setting requires a conditional CDF estimator. When the feature space is high-dimensional in our experiments, this is constructed by using as a black-box the conditional density estimator of Izbicki & Lee (2017) that underlies CD-split and HPD-split. When the conditional density estimate is used as a black-box, the optimal volume set is produced by taking an appropriate level set of the conditional density as taken by CD-split and HPD-split; hence we do not expect our method to outperform CD-split and HPD-split in these experiments. However, we observe that our method is still competitive in these experiments, as described below.

Interpretation: Among the competitive methods, CQR, DCP-QR, DCP-QR* all output a single interval and are more suitable for unimodal distributions. On the other hand, CD-split and HPD-split output a union of intervals and are suitable for data generated by multimodal distributions. The experiments show that our proposed framework of Dynamic Programming is competitive against both settings with different values of k. To be specific, for the dataset in Table 9, DP with $k = 1$ is competitive with the DCP-QR and DCP-QR* which achieves the smallest volume, while in Table 10, DP with $k = 5$ is competitive with HPD-split and CD-split which achieve the smallest volume. In practice, the value of k can be tailored based on the distribution, or chosen adaptively using the ideas in Bai et al. (2022).

Table 9. Comparison of methods on the dataset AirFoil. Reported runtimes exclude a shared preprocessing time of 6.3819 ± 0.1456(s) for estimating the conditional CDF using `FlexCoDE` common to all methods.

Method	Average Volume	Empirical Coverage	Runtime (s)
CQR	7.1167 ± 0.4479	0.8061 ± 0.0266	0.3536 ± 0.0528
DCP-QR	6.4662 ± 0.4478	0.8032 ± 0.0268	0.8185 ± 0.0613
DCP-QR*	$\mathbf{6.4612 \pm 0.4078}$	0.7953 ± 0.0240	0.9003 ± 0.0555
CD-split	7.5517 ± 0.5343	0.7962 ± 0.0269	0.1311 ± 0.0110
HPD-split	7.3650 ± 0.6549	0.8013 ± 0.0278	1.8243 ± 0.0556
DP ($k = 1$)	6.8287 ± 0.3548	0.8065 ± 0.0245	32.3420 ± 0.1109
DP ($k = 5$)	7.5002 ± 0.5302	0.8032 ± 0.0211	206.3317 ± 196.6624

Table 10. Comparison of methods on the dataset WineQuality. Reported runtimes exclude a shared preprocessing time of 68.3879 ± 2.8636(s) for estimating the conditional CDF using `FlexCoDE` common to all methods.

Method	Average Volume	Empirical Coverage	Runtime (s)
CQR	1.1251 ± 0.2066	0.8058 ± 0.0135	5.0441 ± 0.5437
DCP-QR	0.9980 ± 0.0771	0.8042 ± 0.0122	7.2799 ± 0.7250
DCP-QR*	0.9514 ± 0.0721	0.8030 ± 0.0115	7.6734 ± 0.9398
CD-split	0.5066 ± 0.7077	0.7759 ± 0.0368	2.6191 ± 0.1664
HPD-split	$\mathbf{0.2694 \pm 0.0163}$	0.8010 ± 0.0154	9.0277 ± 0.0973
C-DP($k=1$)	0.9389 ± 0.0555	0.8074 ± 0.0134	191.8998 ± 213.9385
C-DP($k=5$)	0.2947 ± 0.0215	0.8048 ± 0.0127	704.8169 ± 40.4745

C. KDE Optimality Implies DP Optimality

Suppose a distribution P on $\mathbb{R}$ admits a density function p. The kernel density estimator depending on k i.i.d. samples $Z_1, \cdots, Z_k$ is defined by

$$p_k(y) = \frac{1}{k\rho} \sum_{j=1}^{k} K\left(\frac{y - Z_j}{\rho}\right), \tag{13}$$

where $K(\cdot)$ is a standard Gaussian kernel and ρ is a bandwidth parameter. The conformal prediction method by (Lei et al., 2013) is based on the idea that the level set of p_k is close to that of p as long as p_k is close to p. In this section, we will show that as long as p_k is close to p, the dynamic programming also finds a prediction set whose volume is nearly optimal compared with the level set of p. This implies that DP always requires no stronger assumption to achieve volume optimality.

The existence of a KDE close to p can be even weakened into the following assumption.

Assumption C.1. For any positive integer k, there exists some $\varepsilon_k > 0$ and some Gaussian mixture $P_k = \sum_{j=1}^{k} w_j N(\mu_j, \sigma_j^2)$ such that $\mathsf{TV}(P_k, P) \le \varepsilon_k$.

In particular, the KDE (13) based on k samples is a special case of the Gaussian mixture, given that Gaussian kernel is used. Though the characterization of the closeness between p_k and p is through ℓ_∞ norm by (Lei et al., 2013), similar error bounds also apply to the ℓ_1 norm, which is the total variation distance. For example, suppose P has bounded support and the Hölder smoothness is $\beta \in (0, 2]$. Then, one can take $\varepsilon_k = \widetilde{\Theta}\left(k^{-\frac{\beta}{2\beta+1}}\right)$ with an appropriate choice of the bandwidth, where $\widetilde{\Theta}$ hides some logarithmic factor of k.

Theorem C.2. *Consider i.i.d. observations $Y_1, \cdots, Y_n, Y_{n+1}$ generated by some distribution P on $\mathbb{R}$ that satisfies Assumption C.1. For any $\alpha, \delta, \gamma \in (0, 1)$ such that $\delta \gg \sqrt{\frac{k + \log n}{n}}$ and $\gamma + 2\delta + 2\varepsilon_k < \alpha$, let $\widehat{C}_{\mathrm{DP}} \in \mathcal{C}_k$ be the output of Algorithm 1 with coverage level $1 - \alpha + \delta$ and slack γ. Then, with probability at least $1 - \delta$, we have*

1. $\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{\mathrm{DP}} \mid Y_1, \cdots, Y_n\right) \ge 1 - \alpha$;
2. $\mathrm{vol}(\widehat{C}_{\mathrm{DP}}) \le \mathrm{OPT}(P, 1 - \alpha + \gamma + 2\delta + 2\varepsilon_k)$.

The result of Theorem C.2 can also be conformalized as in Section 2.4, so that the restricted volume optimality $\mathrm{OPT}_k(P, \cdot)$ in Theorem 2.5 can be strengthened to $\mathrm{OPT}(P, \cdot)$ without restriction whenever P satisfies Assumption C.1, which, in particular, includes the situation where the density of P can be well estimated by KDE.

The volume sub-optimality given by Theorem C.2 is $\gamma + 2\delta + 2\varepsilon_k$. When the distribution P is β-smooth, the sub-optimality is of order $\widetilde{\Theta}\left(\sqrt{\frac{k}{n}} + k^{-\frac{\beta}{2\beta+1}}\right)$ by taking $\varepsilon_k = \widetilde{\Theta}\left(k^{-\frac{\beta}{2\beta+1}}\right)$, $\delta = \widetilde{\Theta}\left(\sqrt{\frac{k}{n}}\right)$, and γ sufficiently small. Thus, optimizing this bound over k leads to the rate $\widetilde{\Theta}\left(n^{-\frac{\beta}{4\beta+1}}\right)$. In comparison, the KDE achieves a faster rate $\widetilde{\Theta}\left(n^{-\frac{\beta}{2\beta+1}}\right)$ (Lei et al., 2013) for smooth densities. This is actually a technical artifact by specializing Assumption C.1 to the KDE (13). In fact, when the density of P is β-smooth, it is well known that Assumption C.1 is satisfied with a better $\varepsilon_k = \widetilde{\Theta}(k^{-\beta})$ (Ghosal &

van der Vaart, 2007; Kruijer et al., 2010), which then leads to the volume sub-optimality $\widetilde{\Theta}\left(\sqrt{\frac{k}{n}}+k^{-\beta}\right)$ that leads to the near optimal rate $\widetilde{\Theta}\left(n^{-\frac{\beta}{2\beta+1}}\right)$ with $k=\widetilde{\Theta}(n^{\frac{1}{2\beta+1}})$.

C.1. Proof of Theorem C.2

We first state a lemma that shows that a level set of a Gaussian mixture with k components must belong to the class $\mathcal{C}_k$.

Lemma C.3. *For a Gaussian mixture $P_k=\sum_{j=1}^k w_j N(\mu_j,\sigma_j^2)$ and any $\alpha\in(0,1)$,*

$$\mathrm{OPT}_k(P_k,1-\alpha)=\mathrm{OPT}(P_k,1-\alpha).$$

The proof of the lemma will be given in Appendix D.4. Now we are ready to state the proof of Theorem C.2.

Proof of Theorem C.2. By Proposition 2.3, we know that $\widehat{C}_{\mathrm{DP}}$ satisfies $\mathbb{P}_n(\widehat{C}_{\mathrm{DP}})\geq 1-\alpha+\delta$ and $\mathrm{vol}(\widehat{C}_{\mathrm{DP}})\leq \mathrm{OPT}_k(\mathbb{P}_n,1-\alpha+\delta+\gamma)$. The condition on δ implies that $\sup_{C\in\mathcal{C}_k}|\mathbb{P}_n(C)-P(C)|\leq\delta$ with probability at least $1-\delta$ (Devroye & Lugosi, 2001). Therefore, the coverage probability is

$$\mathbb{P}\left(Y_{n+1}\in\widehat{C}_{\mathrm{DP}}\mid Y_1,\cdots,Y_n\right)\geq\mathbb{P}_n(\widehat{C}_{\mathrm{DP}})-\delta\geq 1-\alpha,$$

and the volume can be bounded by

$$\begin{aligned}
\mathrm{vol}(\widehat{C}_{\mathrm{DP}}) &\leq \mathrm{OPT}_k(\mathbb{P}_n,1-\alpha+\delta+\gamma)\\
&\leq \mathrm{OPT}_k(P,1-\alpha+2\delta+\gamma)\\
&\leq \mathrm{OPT}_k(P_k,1-\alpha+2\delta+\gamma+\varepsilon_k)\\
&= \mathrm{OPT}(P_k,1-\alpha+2\delta+\gamma+\varepsilon_k)\\
&\leq \mathrm{OPT}(P,1-\alpha+2\delta+\gamma+2\varepsilon_k),
\end{aligned}$$

where the identity above is by Lemma C.3. □

D. Additional Proofs

D.1. Proof of Theorem 2.1

The proof relies on the following technical lemma, whose proof will be given in Appendix D.4.

Lemma D.1. *For any $\delta,\varepsilon>0$ and any integer $n>0$, there exists some distribution Π supported on*

$$\mathcal{P}_\varepsilon=\{P:\mathit{supp}(P)\subset[0,1],\mathsf{TV}(P,\lambda)\geq 1-\varepsilon\},$$

such that $\mathsf{TV}\left(\lambda^n,\int P^n\mathrm{d}\Pi\right)\leq\delta$.

Now we are ready to state the proof of Theorem 2.1.

Proof of Theorem 2.1. By Lemma D.1, there exist $\Pi_{n,\delta}$ and $\Pi_{n+1,\delta}$ supported on $\mathcal{P}_\delta$, such that $\mathsf{TV}\left(\lambda^n,\int P^n\mathrm{d}\Pi_{n,\delta}\right)\leq\delta$ and $\mathsf{TV}\left(\lambda^{n+1},\int P^{n+1}\mathrm{d}\Pi_{n+1,\delta}\right)\leq\delta$.

Since $P^{n+1}(Y_{n+1}\in\widehat{C}(Y_1,\cdots,Y_n))\geq 1-\alpha$ for all P, we have

$$\int P^{n+1}(Y_{n+1}\in\widehat{C}(Y_1,\cdots,Y_n))\mathrm{d}\Pi_{n+1,\delta}\geq 1-\alpha.$$

By $\mathsf{TV}\left(\lambda^{n+1},\int P^{n+1}\mathrm{d}\Pi_{n+1,\delta}\right)\leq\delta$, we have

$$\mathop{\mathbb{E}}_{Y_1,\ldots,Y_n\sim\lambda^n}(\lambda(\widehat{C}(Y_1,\cdots,Y_n)))=\lambda^{n+1}(Y_{n+1}\in\widehat{C}(Y_1,\cdots,Y_n))\geq 1-\alpha-\delta.$$

By $\mathsf{TV}\left(\lambda^n, \int P^n \mathrm{d}\Pi_{n,\delta}\right) \leq \delta$, we have

$$\int \mathop{\mathbb{E}}_{Y_1,\ldots,Y_n \sim P^n} \lambda(\widehat{C}(Y_1, \cdots, Y_n)) \mathrm{d}\Pi_{n,\delta} \geq 1 - \alpha - 2\delta.$$

Then, there must exists some $P \in \mathrm{supp}(\Pi_{n,\delta}) \subset \mathcal{P}_\delta$, such that

$$\mathop{\mathbb{E}}_{Y_1,\ldots,Y_n \sim P^n} \lambda(\widehat{C}(Y_1, \cdots, Y_n)) \geq 1 - \alpha - 2\delta.$$

The fact that $P \in \mathcal{P}_\delta$ implies $\mathsf{TV}(P, \lambda) \geq 1 - \delta$. By the definition of total variation, there exists some set B such that $P(B) - \lambda(B) \geq 1 - \delta$, which implies $P(B) \geq 1 - \delta$ and $\lambda(B) \leq \delta$. Therefore, $\mathrm{OPT}(P, 1 - \delta) \leq \delta$.

We finally have for any $\varepsilon \in (0, \alpha)$, the expected volume of prediction set is

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y_1,\ldots,Y_n \sim P^n} \mathrm{vol}(\widehat{C}(Y_1, \cdots, Y_n)) &\geq 1 - \alpha - 2\delta \\
&\geq \delta \\
&\geq \mathrm{OPT}(P, 1 - \delta) \\
&\geq \mathrm{OPT}(P, 1 - \alpha + \varepsilon),
\end{aligned}$$

as long as δ is sufficiently small so that $\delta < \min\{(1 - \alpha)/3, \alpha - \varepsilon\}$. The proof is complete. □

D.2. Proof of Theorem 2.5

Theorem 2.5 is a special case of Theorem 3.3 in the setting where $F(\cdot \mid x)$ does not depend on x and $\widehat{F}(\cdot \mid x)$ is defined as the empirical CDF of $Y_1, \cdots, Y_n$. Then, Assumption 3.2 is automatically satisfied by a standard VC dimension bound (Devroye & Lugosi, 2001).

D.3. Proof of Theorem 3.3

We will prove the three properties of Theorem 3.3 separately. We note that the marginal coverage property holds without Assumptions 3.1 and 3.2. It is a standard consequence of applying the split conformal framework, but we still include a proof here for completeness.

Proof of Theorem 3.3 (marginal coverage). By the construction of $\widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1})$, we have

$$\mathbb{P}\left(Y_{2n+1} \in \widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1})\right) = \mathbb{P}\left(q(Y_{2n+1}, X_{2n+1}) \geq q_{\lfloor (n+1)\alpha \rfloor}\right).$$

Since the conformity score q is constructed from $\widehat{F}(\cdot \mid \cdot)$, it is independent from the second half of the data. Thus, $q(Y_{2n+1}, X_{2n+1})$ is exchangeable with $q(Y_{n+1}, X_{n+1}), \cdots, q(Y_{2n}, X_{2n})$, which implies the desired conclusion by the definition of $q_{\lfloor (n+1)\alpha \rfloor}$. □

Next, we establish the conditional coverage property. We need the following property of the conformity score that is computed based on a nested system.

Lemma D.2. *For any $j \in [m+1]$, $y \in S_j(x)$ if and only if $q(y, x) \geq m - j + 1$, where the set $S_{m+1}(x)$ is defined as $\mathbb{R}$.*

The proof of the lemma will be given in Appendix D.4.

Proof of Theorem 3.3 (approximate conditional coverage). We first note that Assumption 3.2 implies

$$\mathbb{E}\|\widehat{F}(\cdot \mid X_{2n+1}) - F(\cdot \mid X_{2n+1})\|_{k,\infty} \leq 2\delta. \tag{14}$$

We use $\mathcal{F}_{2n}$ to denote the σ-field generated by the random variables $(X_1, Y_1), \cdots, (X_{2n}, Y_{2n})$. Let $\mathbb{E}_{X_{2n+1}}$ and $\mathbb{E}_{\mathcal{F}_{2n}}$ be the expectation operators under the marginal distributions of X_{2n+1}, and of $(X_1, Y_1), \cdots, (X_{2n}, Y_{2n})$, respectively. Then, we have

$$\mathbb{P}\left(q(Y_{2n+1}, X_{2n+1}) \geq q_{\lfloor (n+1)\alpha \rfloor}\right) = \mathop{\mathbb{E}}_{\mathcal{F}_{2n}} \mathop{\mathbb{E}}_{X_{2n+1}} \mathbb{P}\left(q(Y_{2n+1}, X_{2n+1}) \geq q_{\lfloor (n+1)\alpha \rfloor} | X_{2n+1}, \mathcal{F}_{2n}\right).$$

By Lemma D.2, $q(Y_{2n+1}, X_{2n+1}) \geq q_{\lfloor (n+1)\alpha \rfloor}$ is equivalent to $Y_{2n+1} \in S_{\widehat{j}}(X_{2n+1})$ for some $\widehat{j}$ measurable with respect to $\mathcal{F}_{2n}$. Then, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{F}_{2n}} \mathop{\mathbb{E}}_{X_{2n+1}} \mathbb{P}\left(q(Y_{2n+1}, X_{2n+1}) \geq q_{\lfloor (n+1)\alpha \rfloor} | X_{2n+1}, \mathcal{F}_{2n}\right) &= \mathbb{E}_{\mathcal{F}_{2n}} \mathbb{E}_{X_{2n+1}} \mathbb{P}\left(Y_{2n+1} \in S_{\widehat{j}}(X_{2n+1}) | X_{2n+1}, \mathcal{F}_{2n}\right) \\ &= \mathbb{E}_{\mathcal{F}_{2n}} \mathbb{E}_{X_{2n+1}} \int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}F(y \mid X_{2n+1}).\end{aligned}$$

Since $S_{\widehat{j}}(X_{2n+1}) \in \mathcal{C}_k$, by (14), we have

$$\begin{aligned}\mathbb{E}_{\mathcal{F}_{2n}} \mathbb{E}_{X_{2n+1}} \int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}F(y \mid X_{2n+1}) \leq & \mathbb{E}_{\mathcal{F}_{2n}} \mathbb{E}_{X_{2n+1}} \int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}\widehat{F}(y \mid X_{2n+1}) \\ & + \mathbb{E}\|\widehat{F}(\cdot \mid X_{2n+1}) - F(\cdot \mid X_{2n+1})\|_{k,\infty} \\ \leq & \mathbb{E}_{\mathcal{F}_{2n}} \mathbb{E}_{X_{2n+1}} \int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}\widehat{F}(y \mid X_{2n+1}) + 2\delta.\end{aligned}$$

By Assumption 3.1, we have $\int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}\widehat{F}(y \mid X_{2n+1}) = \frac{\widehat{j}}{m}$, which is independent of X_{2n+1}, since $\widehat{j}$ measurable with respect to $\mathcal{F}_{2n}$. Thus, we have

$$\mathbb{E}_{\mathcal{F}_{2n}} \mathbb{E}_{X_{2n+1}} \int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}\widehat{F}(y \mid X_{2n+1}) + 2\delta = \mathbb{E}_{\mathcal{F}_{2n}} \int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}\widehat{F}(y \mid X_{2n+1}) + 2\delta.$$

By Assumption 3.2, we have with probability at least $1 - \delta$,

$$\begin{aligned}\mathbb{E}_{\mathcal{F}_{2n}} \int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}\widehat{F}(y \mid X_{2n+1}) + 2\delta \leq & \mathbb{E}_{\mathcal{F}_{2n}} \int_{S_{\widehat{j}}(X_{2n+1})} \mathrm{d}F(y \mid X_{2n+1}) + 3\delta \\ = & \mathbb{P}\left(q(Y_{2n+1}, X_{2n+1}) \geq q_{\lfloor (n+1)\alpha \rfloor} \mid X_{2n+1}\right) + 3\delta.\end{aligned}$$

Therefore, with probability at least $1 - \delta$, the approximate conditional coverage holds

$$\mathbb{P}\left(q(Y_{2n+1}, X_{2n+1}) \geq q_{\lfloor (n+1)\alpha \rfloor} \mid X_{2n+1}\right) \geq \mathbb{P}\left(q(Y_{2n+1}, X_{2n+1}) \geq q_{\lfloor (n+1)\alpha \rfloor}\right) - 3\delta \geq 1 - \alpha - 3\delta.$$

□

Finally, we prove the last property on volume optimality.

Proof of Theorem 3.3 (conditional restricted volume optimality). By Assumption 3.1, there exists some $j^* \in [m]$, such that

$$\mathbb{E} \int_{S_{j^*}(X_{2n+1})} \mathrm{d}\widehat{F}(y \mid X_{2n+1}) \geq 1 - \alpha + \frac{1}{n} + 3\delta.$$

Assumption 3.2 implies that

$$\begin{aligned}\mathbb{P}\left(Y_{2n+1} \in S_{j^*}(X_{2n+1})\right) &= \mathbb{E} \int_{S_{j^*}(X_{2n+1})} \mathrm{d}F(y \mid X_{2n+1}) \\ &\geq \mathbb{E} \int_{S_{j^*}(X_{2n+1})} \mathrm{d}\widehat{F}(y \mid X_{2n+1}) \\ &\quad - \mathbb{E}\|\widehat{F}(\cdot \mid X_{2n+1}) - F(\cdot \mid X_{2n+1})\|_{k,\infty} \\ &\geq 1 - \alpha + \frac{1}{n} + \delta.\end{aligned}$$

By Hoeffding's inequality and the condition on δ, we have

$$\frac{1}{n} \sum_{i=n+1}^{2n} \mathbb{I}\{Y_i \in S_{j^*}(X_i)\} \geq \mathbb{P}\left(Y_{2n+1} \in S_{j^*}(X_{2n+1})\right) - \delta,$$

with probability at least $1-\delta$. Combining the two inequalities above and Lemma D.2, we get

$$\frac{1}{n}\sum_{i=n+1}^{2n}\mathbb{I}\{q(Y_i,X_i)\geq m-j^*+1\}\geq 1-\alpha+\frac{1}{n},$$

with probability at least $1-\delta$. This immediately implies $q_{\lfloor n\alpha\rfloor}=q_{\lfloor n(\alpha-n^{-1})+1\rfloor}\geq m-j^*+1$ by the definition of order statistics. Therefore, the volume of the prediction set $\widehat{C}_{\mathrm{DCP-DP}}(X_{2n+1})$ is at most

$$\begin{aligned}
&\mathrm{vol}\left(\left\{y\in\mathbb{R}: q(y,X_{2n+1})\geq q_{\lfloor(n+1)\alpha\rfloor}\right\}\right)\\
\leq&\mathrm{vol}\left(\left\{y\in\mathbb{R}: q(y,X_{2n+1})\geq q_{\lfloor n\alpha\rfloor}\right\}\right)\\
\leq&\mathrm{vol}\left(\left\{y\in\mathbb{R}: q(y,X_{2n+1})\geq m-j^*+1\right\}\right)\\
=&\mathrm{vol}(S_{j^*}(X_{2n+1})),
\end{aligned}$$

where the last identiy is by Lemma D.2. The volume of $S_{j^*}(X_{2n+1})$ can be controlled by Assumption 3.1,

$$\begin{aligned}
\mathrm{vol}(S_{j^*}(X_{2n+1}))\leq&\mathrm{OPT}_k\left(\widehat{F}(\cdot\mid X_{2n+1}),1-\alpha+\frac{1}{n}+3\delta+\gamma\right)\\
\leq&\mathrm{OPT}_k\left(F(\cdot\mid X_{2n+1}),1-\alpha+\frac{1}{n}+4\delta+\gamma\right),
\end{aligned}$$

where the last inequality, which holds with probability at least $1-\delta$, is by Assumption 3.2. Combining the inequalities above with union bound, we get the conclusion. □

D.4. Proofs of Proposition 2.3, Lemma C.3, Lemma D.1 and Lemma D.2

Dynamic Programming Algorithm. The dynamic programming table $DP(i,j,l)$ stores the minimum volume of i intervals that collectively cover $l\gamma n$ points from the sorted training data $Y_{(1)},\ldots,Y_{(j)}$, where the right endpoint of the rightmost interval is fixed at $Y_{(j)}$. Here, $Y_{(1)},\ldots,Y_{(n)}$ are the training data points $Y_1,\ldots,Y_n$ sorted in non-decreasing order. For each state in the DP table, we iterate over all possible left endpoints of the rightmost interval, as well as the right endpoint of the preceding interval (if it exists). This allows us to systematically compute the optimal solution for each state by the following formula:

$$\begin{aligned}
&\text{If } i=1,\quad DP(i,j,l)=Y_{(j)}-Y_{(j-\lceil l\gamma n\rceil+1)},\\
&\text{If } i>1,\quad DP(i,j,l)=\min_{i-1\leq j''<j'\leq j}\{Y_{(j)}-Y_{(j')}+DP(i-1,j'',l')\},
\end{aligned}$$

where $l'=l-\lfloor(j-j'+1)/(\gamma n)\rfloor$. Finally, we find the minimum volume solution among all entries $DP(k,j,\lceil(1-\alpha)/\gamma\rceil)$ for all $1\leq j\leq n$. Then, we use the standard backtrack approach on the DP table to find the prediction set $\widehat{C}_{DP}$.

Proof of Proposition 2.3. Without loss of generality, we assume that $1/\gamma$ is an integer, otherwise, we can decrease γ to make this hold. For any $i\in[k], j\in[n], l\in[1/\gamma]$, we use the dynamic programming table entry $DP(i,j,l)$ to store the minimum volume of i intervals that cover $\lceil l\cdot\gamma n\rceil$ points in $Y_{(1)},\ldots,Y_{(j)}$ and the right endpoint of the rightmost interval is at $Y_{(j)}$. If there is no feasible solution for this subproblem, we set $DP(i,j,l)=\infty$. This dynamic programming is shown in Algorithm 1. This dynamic programming runs in time $O(n^3k/\gamma)$.

We then find the solution with the minimum volume among all subproblems $DP(k,j,\lceil(1-\alpha)/\gamma\rceil)$ for $1\leq j\leq n$. It is easy to see that there exists a feasible solution. Let $\widehat{C}_{\mathrm{DP}}\in\mathcal{C}_k$ be a union of k intervals in this solution. This solution covers at least $\lceil(1-\alpha)/\gamma\cdot(n\gamma)\rceil=\lceil(1-\alpha)n\rceil$ points in $X_1,\ldots,X_n$. Thus, we have

$$\mathbb{P}_n(\widehat{C}_{\mathrm{DP}})\geq 1-\alpha.$$

If the restricted optimal volume $\mathrm{OPT}_k(\mathbb{P}_n,((1-\alpha)/\gamma+1)\cdot(n\gamma)/n)$ is smaller than the volume of $\widehat{C}_{\mathrm{DP}}$, then this solution cannot have the minimum volume among all subproblems $DP(k,j,\lceil(1-\alpha)/\gamma\rceil)$ for $1\leq j\leq n$. Thus, the volume of this solution must satisfy

$$\begin{aligned}
\mathrm{vol}(\widehat{C}_{\mathrm{DP}})\leq&\mathrm{OPT}_k(\mathbb{P}_n,((1-\alpha)/\gamma+1)\cdot(n\gamma)/n)\\
=&\mathrm{OPT}_k(\mathbb{P}_n,1-\alpha+\gamma).
\end{aligned}$$

The proof is thus complete. □

Proof of Lemma C.3. By (Carreira-Perpinán & Williams, 2003), there are $k' \leq k$ local maxima for the density function p_k. We will use k' intervals and define the rest of the intervals to be empty. Suppose $u_1 \leq u_2 \leq \cdots \leq u_{k'} \in \mathbb{R}$ are the local maxima of p_k. The density $p_k(y)$ is differentiable, and its local minima and local maxima have to alternate. Hence there are exactly $k'-1$ local minima, denoted by $\ell_1, \ell_2, \ldots, \ell_{k'-1}$ with $\ell_j \in [u_j, u_{j+1}]$ for all $j \in \{1, 2, \ldots, k'-1\}$. (Note that there are no local minima less than u_1 or greater than $u_{k'}$ since $p_k(y) \to 0$ as $y \to \pm\infty$). For notational convenience let $\ell_0 = -\infty, \ell_{k'} = \infty$. Let $C^* \subset \mathbb{R}$ satisfies $P_k(C^*) \geq 1-\alpha$ and $\mathrm{vol}(C^*) = \mathrm{OPT}(P_k, 1-\alpha)$.

We now show that there exist $I_1, \cdots, I_{k'} \in \mathcal{C}_1$ such that $u_j \in I_j \subset [\ell_{j-1}, \ell_j]$ for all $j \in [k']$, $P_k(\cup_{j=1}^{k'} I_j) \geq 1-\alpha$, and $\mathrm{vol}(\cup_{j=1}^{k'} I_j) \leq \mathrm{vol}(C^*)$. This would imply the desired conclusion. Consider $S_j = S^* \cap [\ell_{j-1}, \ell_j]$ for all $j \in [k']$. Next we observe for all $j \in [k']$, p_k is monotonically increasing in the interval $[\ell_{j-1}, u_j]$ and is monotonically decreasing in the intervals $[u_j, \ell_j]$, with a local maximum at u_j. Hence if S_j comprises multiple disjoint intervals within $[\ell_{j-1}, \ell_j]$, we can pick one interval with the same volume within $[\ell_{j-1}, \ell_j]$ that also includes u_j and covers at least as much probability mass. This establishes the property of $\cup_{j=1}^{k'} I_j$, and hence the lemma. □

Proof of Lemma D.1. First, we construct a family of distributions supported on subsets of $[0,1]$. Consider an integer m. We partition the interval $[0,1]$ into m intervals with length $1/m$ each and define the subinterval $A_j = \left[\frac{j-1}{m}, \frac{j}{m}\right)$ for $j \in [m]$. We next define a family of distributions supported on these subintervals. For any vector $Z \in \{0,1\}^m$, let $A_Z = \bigcup_{j:Z_j=1} A_j$ denote the union of intervals corresponding to the indices where $Z_j = 1$. Then, we define the density function

$$p_Z(y) = \frac{\mathbb{I}\{y \in A_Z\}}{\frac{1}{m}\sum_{j=1}^m Z_j},$$

where $\mathbb{I}$ is the indicator function. Let P_Z be the corresponding distribution.

We then construct the weight distribution Π. Given any $\varepsilon > 0$, we now provide a restricted set of vectors $Z \in \{0,1\}^m$ such that the distribution P_Z has at least $1-\varepsilon$ total variation distance to the uniform distribution λ. We pick a parameter $\beta \in (0,1)$ depending on ε and m. Then, we define a set of vectors Z with A_Z covering approximately β fraction of $[0,1]$,

$$\mathcal{Z} = \left\{ Z \in \{0,1\}^m : \left| \frac{1}{m}\sum_{j=1}^m Z_j - \beta \right| \leq \left(\frac{\beta}{m}\right)^{1/3} \right\}.$$

For any $Z \in \mathcal{Z}$, we have the total variation distance

$$\mathsf{TV}(P_Z, \lambda) \geq \lambda(A_Z^c) = 1 - \frac{1}{m}\sum_{j=1}^m Z_j \geq 1 - \beta - \left(\frac{\beta}{m}\right)^{1/3}.$$

Therefore, as long as $\beta + \left(\frac{\beta}{m}\right)^{1/3} \leq \varepsilon$, we have $P_Z \in \mathcal{P}_\varepsilon$ for all $Z \in \mathcal{Z}$. We construct the weight distribution Π supported on the $\{P_Z : Z \in \mathcal{Z}\}$. Let $\tilde{\Pi} = \otimes_{j=1}^m \text{Bernoulli}(\beta)$ be the product distribution on $\{0,1\}^m$ such that each coordinate is 1 with probability β. Then, we define Π to be the distribution $\tilde{\Pi}$ conditioning on $\mathcal{Z}$, $\Pi(Z) = \frac{\tilde{\Pi}(Z \cap \mathcal{Z})}{\tilde{\Pi}(\mathcal{Z})}$.

We bound $\mathsf{TV}(\lambda^n, \int P_Z^n d\Pi(Z))$ by chi-squared divergence,

$$\frac{1}{2}\mathsf{TV}\left(\lambda^n, \int P_Z^n \mathrm{d}\Pi(Z)\right)^2 \leq \int_{[0,1]^n} \left(\int_Z p_Z^n \mathrm{d}\Pi(Z)\right)^2 \mathrm{d}x - 1.$$

Since $\int P_Z^n d\Pi(Z)$ is a mixture of product distributions over $\{P_Z^n : Z \in \mathcal{Z}\}$ with weights Π, we can expand the first term in the right-hand side as

$$\int_{[0,1]^n} \left(\int_Z p_Z^n \mathrm{d}\Pi(Z)\right)^2 \mathrm{d}x = \mathbb{E}_{Z,Z' \overset{iid}{\sim} \Pi}\left(\int p_Z(y) p_{Z'}(y) \mathrm{d}x\right)^n.$$

By taking the density $p_Z(y) = \frac{\mathbb{I}\{x \in A_Z\}}{\frac{1}{m}\sum_{j=1}^m Z_j}$ into the equation, we have

$$\mathbb{E}_{Z,Z' \overset{iid}{\sim} \Pi}\left(\int p_Z(y) p_{Z'}(y) \mathrm{d}x\right)^n = \mathbb{E}_{Z,Z' \overset{iid}{\sim} \Pi}\left(\frac{1}{\frac{1}{m}\sum_{j=1}^m Z_j} \cdot \frac{1}{\frac{1}{m}\sum_{j=1}^m Z'_j} \cdot \frac{1}{m}\sum_{j=1}^m Z_j Z'_j\right)^n.$$

According to the construction of $\mathcal{Z}$, for any vector $Z \in \mathcal{Z}$, we have $\frac{1}{m}\sum_{j=1}^{m} Z_j \geq \beta - \left(\frac{\beta}{m}\right)^{1/3}$. Since $\tilde{\Pi}$ is the product of Bernoulli distribution with probability β, by the Chernoff bound, we have $\tilde{\Pi}(Z \in \mathcal{Z}) \geq 1 - (\beta/m)^{1/3}$. Thus, we have

$$\begin{aligned}
& \mathbb{E}_{Z,Z' \overset{iid}{\sim} \Pi} \left(\int p_Z(y) p_{Z'}(y) \mathrm{d}x \right)^n \\
\leq & \left(\beta - \left(\frac{\beta}{m}\right)^{1/3} \right)^{-2n} \mathbb{E}_{Z,Z' \overset{iid}{\sim} \Pi} \left(\frac{1}{m} \sum_{j=1}^{m} Z_j Z'_j \right)^n \\
= & \left(\beta - \left(\frac{\beta}{m}\right)^{1/3} \right)^{-2n} \frac{\mathbb{E}_{Z,Z' \overset{iid}{\sim} \tilde{\Pi}} \left(\frac{1}{m} \sum_{j=1}^{m} Z_j Z'_j \right)^n \mathbb{I}\{Z \in \mathcal{Z}\} \mathbb{I}\{Z' \in \mathcal{Z}\}}{\tilde{\Pi}(Z \in \mathcal{Z})^2} \\
\leq & \left(\beta - \left(\frac{\beta}{m}\right)^{1/3} \right)^{-2n} \left(1 - \left(\frac{\beta}{m}\right)^{1/3} \right)^{-2} \mathbb{E}_{Z,Z' \overset{iid}{\sim} \tilde{\Pi}} \left(\frac{1}{m} \sum_{j=1}^{m} Z_j Z'_j \right)^n .
\end{aligned}$$

The last term on the right-hand side can be bounded by

$$\begin{aligned}
\mathbb{E}_{Z,Z' \overset{iid}{\sim} \tilde{\Pi}} \left(\frac{1}{m} \sum_{j=1}^{m} Z_j Z'_j \right)^n & \leq \left(\beta^2 + \left(\frac{\beta^2}{m}\right)^{1/3} \right)^n + \tilde{\Pi} \left(\frac{1}{m} \sum_{j=1}^{m} Z_j Z'_j > \beta^2 + \left(\frac{\beta^2}{m}\right)^{1/3} \right) \\
& \leq \left(\beta^2 + \left(\frac{\beta^2}{m}\right)^{1/3} \right)^n + \left(\frac{\beta^2}{m}\right)^{1/3},
\end{aligned}$$

where the first inequality is using $\frac{1}{m}\sum_{j=1}^{m} Z_j Z'_j \leq 1$ for $\frac{1}{m}\sum_{j=1}^{m} Z_j Z'_j > \beta^2 + \left(\frac{\beta^2}{m}\right)^{1/3}$ and the second inequality is from the Chernoff bound on the sum of m Bernoulli variables with probability β^2. Combining all bounds above, we have

$$\frac{1}{2}\mathsf{TV}\left(\lambda^n, \int P_Z^n \mathrm{d}\Pi(Z) \right)^2 \leq \frac{\left(\beta^2 + \left(\frac{\beta^2}{m}\right)^{1/3} \right)^n + \left(\frac{\beta^2}{m}\right)^{1/3}}{\left(\beta - \left(\frac{\beta}{m}\right)^{1/3} \right)^{2n} \left(1 - \left(\frac{\beta}{m}\right)^{1/3} \right)^2} - 1.$$

It is clear that the above bound tends to zero when $m \to \infty$. Therefore, for any $\delta > 0$, we have $\mathsf{TV}\left(\lambda^n, \int P_Z^n \mathrm{d}\Pi(Z)\right) \leq \delta$ for a sufficiently large m. □

Proof of Lemma D.2. The result is a direct consequence of the nested property of $\{S_j(x)\}_{j \in [m]}$. □