

SMOOTHING-BASED CONFORMAL PREDICTION FOR BALANCING EFFICIENCY AND INTERPRETABILITY

Anonymous authors

Paper under double-blind review

ABSTRACT

Conformal Prediction (CP) is a distribution-free framework for constructing statistically rigorous prediction sets. While popular variants such as CD-split improve CP’s efficiency, they often yield prediction sets composed of multiple disconnected subintervals, which are difficult to interpret. In this paper, we propose SCD-split, which incorporates smoothing operations into the CP framework. Such smoothing operations potentially help merge the subintervals, thus leading to interpretable prediction sets. Experimental results on both synthetic and real-world datasets demonstrate that SCD-split balances the interval length and the number of disconnected subintervals. Theoretically, under specific conditions, SCD-split provably reduces the number of disconnected subintervals while maintaining comparable coverage guarantees and interval length compared with CD-split.

1 INTRODUCTION

Machine learning models have achieved remarkable success across numerous applications, including large language models (Chang et al., 2024), medical diagnosis (Marcinkevičs et al., 2022), and investment (Papasotiriou et al., 2024). Despite the impressive performance and widespread adoption, they are often sensitive to noise, model misspecification, and inference errors (Abdar et al., 2021), which undermine the prediction reliability and thus limit their practical applicability in high-stakes scenarios (Nguyen et al., 2015; Hein et al., 2019; Martino et al., 2023; Huang et al., 2025a). Consequently, this has raised interest in developing rigorous methods for *uncertainty quantification* to enhance the trustworthiness of machine learning outputs (Guo et al., 2017; Kristiadi et al., 2020).

Among various uncertainty quantification approaches, conformal prediction (CP) has emerged as a powerful and versatile framework that provides statistically rigorous uncertainty guarantees under mild assumptions (Vovk et al., 2005; Romano et al., 2019a; Angelopoulos & Bates, 2021). CP wraps around black-box predictive models and outputs prediction sets whose validity is ensured by data exchangeability, without requiring knowledge of the underlying data distribution. Owing to its strong theoretical guarantees and model-agnostic flexibility, conformal prediction has demonstrated promising and growing applicability across a variety of fields, including drug discovery (Laghvarapu et al., 2023), large language model (Gui et al., 2024), and health care (Eghbali et al., 2024).

Beyond guaranteeing the coverage, conformal prediction is expected to produce prediction sets with smaller lengths to enable more informative uncertainty quantification in practice. Consequently, many approaches have been proposed to improve length efficiency under the conformal prediction framework (Romano et al., 2019c; Teng et al., 2023; Izbicki et al., 2019). Among them, CD-split (Izbicki et al., 2021) stands out due to its strong performance in improving length efficiency, which uses the conditional density estimation as the conformity score. It approximately achieves prediction sets with minimal Lebesgue measure while maintaining valid coverage when the conditional density estimation is accurate (Izbicki et al., 2019).

Despite the strong theoretical properties of CD-split, several practical challenges arise when it is applied to real-world scenarios. When the conditional distribution is complex or highly multimodal, the prediction sets generated by CD-split often consist of many small disconnected intervals¹ (see Figure 1). The lack of connectivity makes the prediction sets difficult to interpret and thus limits their usefulness in practical tasks where clear and concise predictions are preferred.

¹We here restrict our discussions to the regression tasks (See Appendix B for more discussions).

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

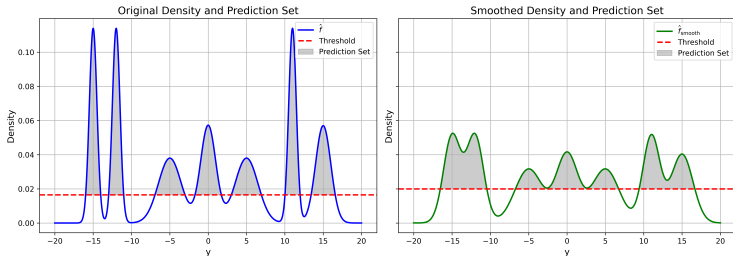


Figure 1: Illustration of Fourier smoothing on a synthetic multimodal distribution. The original density (left) contains seven sharp peaks, resulting in prediction sets composed of multiple disconnected intervals. After applying a smoothing technique (right), the number of intervals is reduced to three with a mild increase in total length, improving the interpretability of the prediction sets.

In this paper, we propose *SCD-split* to address the above challenges. *SCD-split* explicitly focuses on the interpretability of prediction sets. Specifically, interpretability measures how clearly and intuitively the prediction sets convey information to users, which is quantified using both interval length and the number of disjoint intervals (connectivity). Within *SCD-split*, users first specify a desired number of disjoint intervals that they regard as appropriate for interpretation. To meet this requirement, *SCD-split* applies a smoothing technique to the fitted conditional density function before constructing the prediction sets. This smoothing step reduces unnecessary peaks in the estimated density function and corrects distortions caused by complex noise or overfitting, making the density estimation more stable and meaningful. We further use the validation process to tune the smoothing parameter so that the final prediction set conforms to the specified number of intervals as closely as possible. As a result, the final prediction sets contain disconnected intervals matching users’ desired number, making them easier to interpret while maintaining the coverage guarantee. While smoothing has long been a standard tool, this is the first work to introduce it into conformal prediction frameworks to directly regulate the connectivity of prediction sets, thereby providing a novel and principled approach to shape their structure. We refer to Figure 1 for a visual illustration.

Theoretically, we first prove in Theorem 4.1 that the proposed smoothing procedure preserves the marginal coverage guarantee of conformal prediction. Second, under general conditions, we establish that smoothing techniques lead to controlled behaviors: the length of the prediction sets admits a provable upper bound (Theorem 4.4), and the number of disconnected intervals does not increase after smoothing (Theorem 4.2). Third, we prove that smoothing strictly reduces the number of intervals under specific structural assumptions—such as narrow-valley double peaks (Theorem 4.3), thereby improving the interpretability of the prediction sets without sacrificing coverage. **Empirically**, we evaluate our method on both synthetic and real-world datasets in Section 5. The results show that our method achieves a favorable trade-off between interval length and the number of intervals while maintaining validity, particularly under complex and multimodal distributions. Such balance leads to a notable improvement in the interpretability of the prediction sets.

In many practical problems, controlling the number of disjoint intervals in prediction sets is essential for making the results interpretable and actionable. We briefly present two motivating examples:

Health Care. In medical prognosis, when a disease’s course is highly uncertain, doctors often face diseases whose future course may branch into a few qualitatively different trajectories—for example, a fast-progressing fatal path and a long-term recovery path. For treatment planning and patient counseling, it is important to express these distinct possibilities through a manageable number of separate time ranges, rather than a single broad interval that mixes them together or an excessive number of small intervals that are difficult to interpret. Our method allows physicians to directly specify the desired number of disjoint intervals so that the resulting conformal prediction sets provide interpretable and clinically actionable uncertainty quantification.

Finance. In stock price forecasting, market conditions may be highly uncertain, and investors may face two qualitatively different outcomes: a strong upward movement or a significant decline. In such cases, they often wish to know two separate ranges with higher accuracy—one indicating how high the price may rise if the market strengthens, and another showing how low it may fall if conditions worsen. Our method allows investors to pre-specify the desired number of disjoint intervals, so that the resulting conformal prediction set clearly distinguishes these up-side and down-side scenarios and provides more concrete guidance for trading and risk management.

2 RELATED WORK

Conformal Prediction. Conformal prediction (Vovk et al., 2005; Shafer & Vovk, 2008; Barber et al., 2020) is a statistical framework that turns black-box model outputs into predictive intervals. It offers several desirable properties, including distribution-free, non-asymptotic guarantees and a user-friendly implementation (Angelopoulos & Bates, 2021). Existing research on conformal prediction mainly focuses on two aspects: interval length and coverage guarantee. Interval length is an important metric measuring the performance of conformal prediction methods (Teng et al., 2023; 2021; Angelopoulos et al., 2020; Zhou & Sesia, 2024). To minimize the predicted interval length, researchers try to build adaptive prediction intervals (Romano et al., 2019b; Lu, 2024), modify non-conformity scores (Izbicki et al., 2019; Wang & Qiao, 2025) or regard interval length as the optimization objective (Stutz et al., 2022; Kiyani et al., 2024; Bars & Humbert, 2025). For coverage guarantee, numerous works focus on improving the conditional coverage (Romano et al., 2019a; Gibbs et al., 2024; Plassier et al., 2025). Unfortunately, conditional coverage holds only on some special distributions (Vovk, 2012; Lei & Wasserman, 2014; Barber et al., 2020). Therefore, work on conditional coverage can be roughly split into two branches: (a) *local coverage* (Lei & Wasserman, 2014; Barber et al., 2020; Guan, 2023) controls the conditional coverage in a pre-selected space; (b) *asymptotic coverage* (Lei et al., 2018; Izbicki et al., 2019; Sesia & Romano, 2021) establishes conditional coverage guarantees that hold asymptotically as the sample size tends to infinity.

Conformal Prediction and Interpretability. Conformal prediction has been used to enhance the interpretability of the model (Johansson et al., 2018; Sanchez-Martin et al., 2024; Qian et al., 2024) in various fields that need interpretability and reliability, *e.g.*, medicine (Lu et al., 2022; Hirsch & Goldberger, 2024; Huang et al., 2025b) and finance (Zaffran et al., 2022). However, the interpretability of conformal prediction techniques is still underexplored, *e.g.*, confidence intervals with multiple disconnected intervals may potentially influence the interpretability of conformal prediction.

Smoothing. Smoothing methods have been used in various fields, *e.g.*, computer vision (Wang et al., 2022), statistics (Chacón et al., 2013; Ho & Walker, 2020) and numerical analysis (Pandey & Anand, 2020). In this paper, we mainly focus on Fourier smoothing and randomized smoothing. Fourier smoothing uses different frequency-domain filters, *e.g.*, ideal low-pass filter (ILPF) (Jeon et al., 2024), Gaussian low-pass filter (GLPF) (Mehrabkhani, 2019; 2022), Butterworth low-pass filter (BLPF) (Xiao & Bo, 2025) and window functions (Ohamouddou et al., 2025). Randomized smoothing has been used for constructing adversarial robustness classifiers (Cohen et al., 2019; Teng et al., 2020). Gendler et al. (2019); Yan et al. (2024) introduce randomized smoothing into conformal prediction and propose randomized smoothing conformal prediction (RSCP), which is a robust conformal prediction framework under adversaries. Unlike RSCP which applies randomized smoothing at the input level to improve robustness against adversarial perturbations, our method smooths the estimated conditional density function to improve interpretability.

3 METHODOLOGY

In this section, we propose our SCD-split framework to improve the interpretability of prediction sets constructed from conditional density estimators. We first review the classical split conformal prediction and its extension to density-based methods in Section 3.1. Then, we introduce a Fourier-based smoothing technique to regularize the estimated conditional densities in Section 3.2. Finally, we summarize the complete SCD-split procedure in Section 3.3.

3.1 CONFORMAL PREDICTION BASED ON CONDITIONAL DENSITY ESTIMATOR

Split conformal prediction is a commonly adopted method in conformal prediction, which constructs valid prediction sets through a data-splitting procedure. Specifically, the dataset is divided into two disjoint subsets: training set \mathcal{D}_{tr} and calibration set \mathcal{D}_{ca} . A predictor \hat{f} is trained on \mathcal{D}_{tr} , and non-conformity scores $V(X_i, Y_i)$ are computed on \mathcal{D}_{ca} symmetrically. This validity property relies on a mild assumption about the data: the exchangeability of the data pairs in Assumption 3.1.

Assumption 3.1 (Exchangeability). *Define $\{Z_i\}_{i=1}^n$, as the data pairs, then Z_i are exchangeable if arbitrary permutation follows the same distribution, i.e.,*

$$(Z_1, \dots, Z_n) \stackrel{d}{=} (Z_{\pi(1)}, \dots, Z_{\pi(n)}), \quad (1)$$

with arbitrary permutation π over $\{1, \dots, n\}$.

This setup ensures that the non-conformity score V_{n+1} for a new test point is exchangeable with the scores in \mathcal{D}_{ca} , which in turn implies that the rank of V_{n+1} among V_1, V_2, \dots, V_{n+1} is uniformly distributed. Consequently, a valid prediction set can be formed using a quantile-based threshold:

$$\mathcal{C}_{1-\alpha}(X_{n+1}) = \{y : V(X_{n+1}, y) \leq \text{Quantile}(1 - \alpha; \{V_i\}_{i \in \mathcal{I}_{\text{ca}}} \cup \{+\infty\})\}, \quad (2)$$

where \mathcal{I}_{ca} denotes the index set corresponding to the calibration set \mathcal{D}_{ca} . This approach guarantees marginal coverage for prediction sets:

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

CD-split. Within the framework of split conformal prediction, a notable class of methods constructs prediction sets via conditional density estimation (Izbicki et al., 2021; 2019). These methods train a conditional density estimator $\hat{f}(y | x)$ on the training set \mathcal{D}_{tr} and compute conformity scores directly as $\{\hat{f}(y_i | x_i), i \in \mathcal{I}_{\text{ca}} \cup \{n+1\}\}$ using the calibration set \mathcal{D}_{ca} and the test point. Given the exchangeability of the data points $\{Z_i\}_{i=1}^{n+1}$, these conformity scores are also exchangeable, enabling the construction of valid prediction sets under the split conformal prediction framework.

Among them, CD-split is a representative approach that exemplifies this strategy (Izbicki et al., 2021). Building on this framework, CD-split further clusters the input space and, when constructing prediction sets, uses only the calibration data that are similar to each test point. Through these mechanisms, CD-split constructs prediction sets that asymptotically converge to the oracle highest predictive density set (Proposition C.1 in Izbicki et al. (2021)). This property enables CD-split to produce smaller prediction sets compared to interval-based methods when the conditional density estimation performs well, thereby yielding improvements in efficiency. However, such mechanism may produce disconnected prediction sets under multimodal distributions, which hinders interpretability and poses challenges in practical applications. We refer to Section 5 for more details.

3.2 SMOOTHING TECHNIQUE

This section introduces the basics of smoothing techniques. Despite the efficiency advantages of CD-split, it potentially yields prediction sets with multiple disconnected intervals when the estimated conditional density is complex or highly multimodal. To address this issue, we introduce a Fourier-based smoothing technique which regularizes the estimated density function. Specifically, the Fourier transform in Definition 3.1 utilizes the powerful frequency-domain representation of functions to reduce noise and high-frequency oscillations in the estimated conditional density.

Definition 3.1 (Fourier Smoothing for Conditional Density Estimation). *Let $\hat{f}(y | x)$ be an estimated conditional density. We apply Fourier smoothing in the response variable y as follows: First compute the Fourier transform of $\hat{f}(y | x)$ with respect to y : $\mathcal{F}_y[\hat{f}](w | x) = \int_{-\infty}^{\infty} \hat{f}(y | x) e^{-2\pi i y w} dy$. Then multiply this transform by a Gaussian low-pass filter $H_\sigma(w) = e^{-2\pi^2 \sigma^2 w^2}$, where the smoothing parameter $\sigma > 0$ controls the strength of smoothing. The smoothed spectrum is $\mathcal{F}_y[\hat{f}]^{\text{FS}}(w | x) = \mathcal{F}_y[\hat{f}](w | x) H_\sigma(w)$, and the smoothed conditional density is obtained by the inverse transform*

$$\tilde{f}^{\text{FS}}(y | x) = \int_{-\infty}^{\infty} \mathcal{F}_y[\hat{f}]^{\text{FS}}(w | x) e^{2\pi i y w} dw. \quad (4)$$

The key insight is that sharp variations or spurious peaks in the density function correspond to high-frequency components in its spectral representation. By applying Fourier smoothing, we reduce the number of local modes in $\hat{f}(y | x)$, especially those arising from estimation noise. This has a direct impact on the structure of the prediction sets generated by CD-split: the number of disjoint intervals is reduced, and the resulting sets are more concise and easier to present and interpret, while still preserving valid coverage guarantees.

3.3 SCD-SPLIT ALGORITHM

The proposed algorithm integrates a smoothing technique into the CD-split framework, aiming to enhance the interpretability of prediction sets while preserving CD-split’s desirable theoretical properties, such as local conditional coverage. Each step of the pipeline is described below in detail:

Algorithm 1 SCD-split

Input: Dataset $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, confidence level $1 - \alpha \in (0, 1)$, training algorithm \mathcal{T} for conditional density, smoothing operator $s_\sigma(\cdot)$, candidate grid Σ for σ , prespecified target number of intervals K_{target} .

- 1: Randomly split \mathcal{D} into training \mathcal{D}_{tr} , validation \mathcal{D}_{val} , and calibration \mathcal{D}_{ca} .
- 2: Fit $\hat{f} = \mathcal{T}(\mathcal{D}_{\text{tr}})$ where $\hat{f}(Y_i | \mathbf{X}_i)$ is the estimated conditional density;
- 3: **for** each $\sigma \in \Sigma$ **do**
- 4: Smooth the estimator: $\tilde{f}_\sigma^{\text{FS}}(y | \mathbf{x}) \leftarrow s_\sigma(\hat{f}(y | \mathbf{x}))$.
- 5: Compute a partition \mathcal{A}_σ of \mathcal{X} by clustering training samples via profile distance (Def. C.3).
- 6: For each cell $a \in \mathcal{A}_\sigma$, form $U_\sigma(a) = \{\tilde{f}_\sigma^{\text{FS}}(Y_i | \mathbf{X}_i) : (\mathbf{X}_i, Y_i) \in \mathcal{D}_{\text{ca}}, \mathbf{X}_i \in a\}$ and compute threshold $t_\sigma^S(a) = \text{Quantile}(\alpha; U_\sigma(a))$.
- 7: For each $(\mathbf{X}_j, Y_j) \in \mathcal{D}_{\text{val}}$, find its cell $a_j \in \mathcal{A}_\sigma$, construct $\mathcal{C}_\sigma^S(\mathbf{X}_j) = \{y : \tilde{f}_\sigma^{\text{FS}}(y | \mathbf{X}_j) \geq t_\sigma^S(a_j)\}$, and record the number of disjoint intervals $N_\sigma(\mathbf{X}_j)$.
- 8: Compute $R(\sigma) = \left| \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(\mathbf{X}_j, Y_j) \in \mathcal{D}_{\text{val}}} N_\sigma(\mathbf{X}_j) - K_{\text{target}} \right|$.
- 9: **end for**
- 10: Select $\hat{\sigma} \in \arg \min_{\sigma \in \Sigma} R(\sigma)$; set $\tilde{f}^{\text{FS}} \leftarrow \tilde{f}_{\hat{\sigma}}^{\text{FS}}$ and $\mathcal{A} \leftarrow \mathcal{A}_{\hat{\sigma}}$.
- 11: Find the partition $a(\mathbf{X}_{n+1}) \in \mathcal{A}$ with $\mathbf{X}_{n+1} \in a(\mathbf{X}_{n+1})$;
- 12: Form the set $U(\mathbf{X}_{n+1}, \mathcal{D}_{\text{ca}}) = \{\tilde{f}^{\text{FS}}(Y_i | \mathbf{X}_i) : (\mathbf{X}_i, Y_i) \in \mathcal{D}_{\text{ca}}, \mathbf{X}_i \in a(\mathbf{X}_{n+1})\}$;
- 13: Compute $t^S = \text{Quantile}(\alpha; U(\mathbf{X}_{n+1}, \mathcal{D}_{\text{ca}}))$;

Output: Prediction set $\mathcal{C}_{1-\alpha}^S(\mathbf{X}_{n+1}) = \{y : \tilde{f}^{\text{FS}}(y | \mathbf{X}_{n+1}) \geq t^S\}$.

Dataset. Consider an exchangeable dataset $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. We randomly split \mathcal{D} into three parts: a training set \mathcal{D}_{tr} for fitting the conditional density, a validation set \mathcal{D}_{val} for tuning the smoothing parameter, and a calibration set \mathcal{D}_{ca} for constructing the final prediction sets. We further assume that the test pair $(\mathbf{X}_{n+1}, Y_{n+1})$ is exchangeable with \mathcal{D} .

Choice of target interval number. The target interval number K_{target} is predetermined by the user based on domain knowledge and the specific requirements of the application. This user-specified quantity serves as a way to incorporate prior understanding of the problem’s structural characteristics into the modeling process, ensuring that the resulting prediction sets are interpretable.

Training process. We utilize the machine learning model, such as random forest or neural networks, to train a model via the training set \mathcal{D}_{tr} .

Choosing the smoothing parameter σ . We select the smoothing parameter σ by evaluating all candidate values on the validation set and choosing the one whose prediction sets yield an average number of disjoint intervals closest to the user-specified target K_{target} .

First, for each candidate value σ , we smooth the estimated density \hat{f} by applying the Fourier smoothing operator s_σ , obtaining $\tilde{f}_\sigma^{\text{FS}} = s_\sigma(\hat{f})$. Based on $\tilde{f}_\sigma^{\text{FS}}$, we compute conditional CDF profiles and perform k -means++ clustering on the training covariates based on the profile distance (Definition C.3 in (Izbicki et al., 2021)) to form a partition \mathcal{A}_σ of the input space \mathcal{X} .

Second, we use the calibration set \mathcal{D}_{ca} to construct provisional conformal thresholds for each cell of \mathcal{A}_σ . These thresholds allow us to form prediction sets on the validation set. Then, we record the number of disjoint intervals for every validation point. Finally, we compute the difference between the average number of intervals and the user-specified target K_{target} , and select the σ whose prediction sets best match this target. The chosen σ determines both the final smoothed estimator \tilde{f}^{FS} and the final partition \mathcal{A} . We further discuss the choice of loss function in Appendix D.2.

Calibration and testing process. For a new test point \mathbf{X}_{n+1} , we locate the cell $a(\mathbf{X}_{n+1}) \in \mathcal{A}$ containing it and form the final conformal prediction set $\mathcal{C}_{1-\alpha}^S(\mathbf{X}_{n+1})$ using the smoothed density \tilde{f}^{FS} and the corresponding calibrated threshold.

4 THEORETICAL GUARANTEE

This section presents theoretical guarantees for SCD-split. Specifically, Theorem 4.1 establishes the finite-sample coverage guarantee of prediction sets. Theorem 4.2 demonstrates that the smoothing

operation does not increase the number of disconnected intervals, thereby improving connectivity. Theorem 4.3 provides a special structural case that the number of disconnected intervals decreases after smoothing. Lastly, Theorem 4.4 provides an upper bound on the length increase of the predicted intervals produced by the SCD-split algorithm compared with CD-split.

Theorem 4.1 (Coverage Preservation). *Let $\alpha \in (0, 1)$ and assume the data pairs $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable. For a new data point \mathbf{X}_{n+1} , let $\mathcal{C}_{1-\alpha}^S(\mathbf{X}_{n+1})$ denote the prediction sets of \mathbf{X}_{n+1} predicted by the SCD-split algorithm. Then*

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{1-\alpha}^S(\mathbf{X}_{n+1})) \geq 1 - \alpha. \quad (5)$$

The intuition for theorem 4.1 is that the smoothing operation preserves the symmetry property of the score function with respect to the calibration and test data. Consequently, the exchangeability of the data pairs $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n+1}$ naturally induces the exchangeability of the conformity scores. This key property ensures that SCD-split retains the finite-sample coverage guarantee. The detailed proof is provided in Appendix C.2.

Theorem 4.2 (Non-increasing interval count). *Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$ is bounded, measurable, and differentiable, and denote $\mathcal{B} = \{t : f(x) = t, f'(x) \neq 0\}$. Let \tilde{f}^{FS} be the function after applying Fourier smoothing to function f . For all $t \in \mathcal{B}$ such that $A_t := \{x : f(x) = t, f'(x) \neq 0\}$ is finite, i.e., $\#A_t < \infty$, it holds that*

$$\#B_t := \#\{x : \tilde{f}^{FS}(x) = t\} \leq \#A_t, \quad (6)$$

where $\#$ denotes the cardinality of the set. Namely, the number of disconnected intervals of B_t is less than or equal to that of A_t .

Intuitively, the number of disconnected intervals would not increase because Fourier smoothing merges small oscillations. Theorem 4.2 indicates that the number of intervals predicted by SCD-split does not exceed those before smoothing, thereby enhancing interpretability. We assume here that the threshold t remains unchanged for simplicity, as the smoothing operation does not significantly change its value. The detailed proof is provided in Appendix C.3.

Theorem 4.3 (Strict merging under narrow-valley structure). *Assume $f : \mathbb{R} \rightarrow \mathbb{R}$ to be bounded, measurable, and differentiable, and let \tilde{f}^{FS} denote its Fourier smoothed version. For a fixed threshold $t \in \mathcal{B}$, suppose there exist two adjacent intervals (a_1, b_1) and (a_2, b_2) , both subsets of $\{x : f(x) \geq t\}$, such that the valley region (b_1, a_2) between them satisfies:*

- $f(x) \leq t - \varepsilon$ for all $x \in (b_1, a_2)$, for some $\varepsilon > 0$;
- The gap width $\delta := a_2 - b_1$ satisfies $\int_{|u| \geq \delta/2} \phi_\sigma(u) du \geq \frac{\varepsilon}{\|f\|_\infty}$, where ϕ_σ is the Gaussian kernel used in the convolution.

Then after smoothing, the number of intervals strictly decreases:

$$\#\{x : \tilde{f}^{FS}(x) \geq t\} < \#\{x : f(x) \geq t\}. \quad (7)$$

The intuition for theorem 4.3 is that when two high regions of the function are separated by a narrow and shallow valley, the Gaussian kernel convolution has sufficient smoothing power to “fill in” the valley, effectively merging previously disconnected regions into a single connected interval. The detailed proofs are provided in Appendix C.4. We also provide another special case in Appendix C.6.

Theorem 4.4 (Interval length bound). *Define σ as the smoothing factor of Fourier smoothing, N as the number of disconnected intervals predicted by CD-split, the original estimated conditional density function as $\hat{f}(y | \mathbf{X})$ and the smoothed conditional density function as $\tilde{f}^{FS}(y | \mathbf{X})$. Assume \hat{f} and \tilde{f}^{FS} are L -Lipschitz and satisfy $|\hat{f}(y_1 | \mathbf{X}) - \hat{f}(y_2 | \mathbf{X})| \geq M|y_1 - y_2|$, $f \in \{\hat{f}, \tilde{f}^{FS}\}$. Then the difference between the original predicted interval length l and the smoothed predicted interval length \tilde{l} satisfies*

$$|\tilde{l} - l| \leq \frac{4NL\sigma}{M} \sqrt{\frac{2}{\pi}}. \quad (8)$$

The intuition for Theorem 4.4 is that the uniform bound on the pointwise deviation between the original and smoothed conditional densities leads to a bounded shift in the empirical quantile thresholds which determine the endpoints of the intervals. Consequently, the reduction in the number of intervals predicted by SCD-split given by theorem 4.2 does not result in excessively long intervals, thus preserving interpretability. We provide a detailed statement and proof in Appendix C.5.

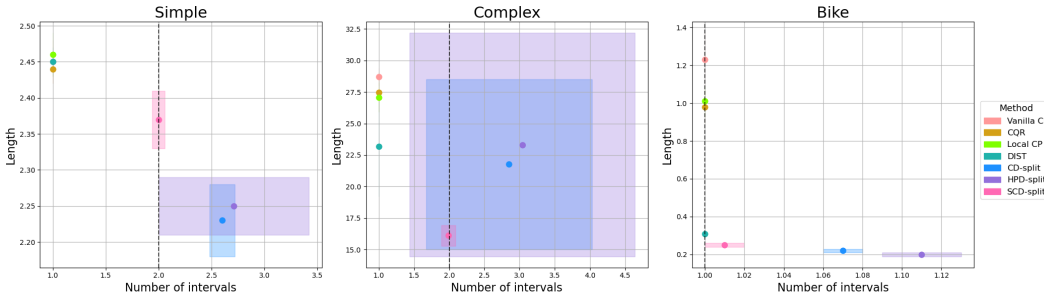


Figure 2: Length vs Number of intervals under complex synthetic and real-world data settings. Each rectangle size shows the standard deviation around the mean. Points closer to the black dashed vertical line on the x -axis (the target number of intervals) and lower on the y -axis (shorter length) indicate better performance. Our proposed SCD-split method consistently reaches the target number of intervals while maintaining shorter lengths across all tasks, demonstrating strong overall performance. We defer more related details to Appendix D.

5 EXPERIMENTS

We conduct experiments on synthetic and real-world datasets, mainly to show that SCD-split is (a) effective, *i.e.*, it constructs valid prediction sets with empirical coverage larger than $1 - \alpha$, (b) efficient, *i.e.*, it constructs prediction sets with relatively short interval length, and (c) interpretable, *i.e.*, it constructs prediction sets with interval number close to target number.

5.1 SETUP

Datasets: We evaluate our method on both *synthetic* and *real-world* datasets. The synthetic datasets include two types: a simple multimodal distribution generated by mixing three Gaussian components with identical variances, and a more complex multimodal distribution formed by mixing multiple Gaussians with varying means and variances. For real-world evaluation, we use several standard datasets commonly adopted in conformal prediction studies, such as `bio` and `bike`, covering diverse application domains and distributional characteristics.

Baselines: We compare our method against two categories of baselines. The first category consists of standard split conformal prediction methods, including vanilla conformal prediction (CP) (Vovk et al., 2005), conformalized quantile regression (CQR) (Romano et al., 2019c), and local conformal prediction (LCP) (Guan, 2023). The second category includes methods based on conditional density estimation, specifically `dist`, `CD-split`, and `HPD-split` (Izbicki et al., 2019; 2021). To ensure a fair comparison across all methods, we uniformly use random forests as the underlying predictive model.

Evaluation metrics: We evaluate all methods using three metrics. The first metric is empirical coverage, which measures the proportion of true responses captured by the prediction sets; a valid method should achieve coverage greater than or equal to $1 - \alpha$. The second metric is the interval length, where a shorter length indicates more precise predictions. The third metric is the number of disjoint subintervals, with values closer to the target number indicating better interpretability.

Synthetic data. We generate synthetic data to evaluate performance across varying levels of complexity. The covariate vector $X = (X_1, \dots, X_d)$ is sampled *i.i.d.* from $\text{Unif}(-5, 5)$ and standardized. The response variable $Y | X$ follows a flexible multi-modal mixture model:

$$Y | X \sim \sum_{k=1}^K \frac{\exp(X^\top \beta_k)}{\sum_{j=1}^K \exp(X^\top \beta_j)} \mathcal{N}(\mu_{\text{base},k} + X^\top \gamma_k, \sigma_k^2), \tag{9}$$

where the parameters are constructed to control the number, location, and shape of the modes. This setup allows us to evaluate both simple and complex structures under a unified framework.

Real-world data. We conduct experiments on several real-world datasets commonly used in the conformal prediction literature (Romano et al., 2019a; Teng et al., 2023), including the bike sharing dataset (`bike`) (Fanaee-T, 2013) and physicochemical properties of protein tertiary structure dataset (`bio`) (Rana, 2013). On these datasets, the fitted conditional density estimates are both multi-modal. We defer more related details to Appendix D.

Table 1: Different smoothing parameters on synthetic complex dataset

Method / σ	Coverage (%)	Length	Number of Intervals
CD-split ($\sigma = 0$)	91.06 ± 3.55	21.76 ± 6.74	2.85 ± 1.18
SCD-split ($\sigma = 1$)	89.38 ± 0.92	16.20 ± 0.80	2.51 ± 0.21
SCD-split ($\sigma = 1.5$)	89.23 ± 0.77	16.11 ± 0.68	1.99 ± 0.01
SCD-split ($\sigma = 2$)	89.39 ± 0.85	16.53 ± 0.36	1.74 ± 0.06
SCD-split ($\sigma = 5$)	89.42 ± 0.86	19.78 ± 1.05	1.18 ± 0.02
SCD-split ($\sigma = 10$)	89.47 ± 1.00	22.53 ± 1.10	1.00 ± 0.00

5.2 RESULT AND DISCUSSION

Validity. We summarize the empirical coverage in Table 2 and Table 3. The results show that the empirical coverage achieved by all methods matches the theoretical target $1 - \alpha$, demonstrating the effectiveness of the proposed procedures.

Efficiency. We summarize the interval lengths in Table 2, Table 3 and Figure 2. We evaluate the efficiency of each method by measuring the average length of the prediction sets. On both synthetic and real-world datasets, we find that methods based on conditional density estimation generally produce much shorter prediction intervals than standard conformal prediction methods. This is because density-based methods allow disconnected prediction sets, which makes it possible to include only the regions with high estimated probability and avoid unnecessary coverage in low-density areas. After applying the smoothing technique, we observe two different behaviors depending on the complexity of the data. When the conditional distribution is not very complex, smoothing slightly increases the interval length due to the regularization effect, but this increase is small and acceptable. However, when the distribution is highly multimodal or the data is noisy, smoothing helps remove spurious modes and reduces the influence of noise in the estimated densities. As a result, the prediction sets become shorter especially in complex settings or real-world data with large noise, which improves efficiency while maintaining valid coverage.

Connectivity and Interpretability. We evaluate connectivity by reporting the number of intervals in Table 2, Table 3, Figure 2. Different from classical conformal prediction methods which usually produce a single connected interval, density-based approaches such as CD-split and HPD-split often generate prediction sets with many disconnected components. Our results imply that applying smoothing operations allows SCD-split to accurately approach the user-specified target number of intervals, which leads to prediction sets that are both faithful to the desired structure and easier to interpret. *Furthermore*, we assess interpretability by jointly considering how close the number of intervals is to the target and how small the total interval length remains. SCD-split consistently achieves a favorable trade-off between these two metrics: compared with existing methods, it brings the number of intervals closer to the target while keeping lengths competitive.

Ablation on the smoothing parameter σ . We investigate the impact of the smoothing parameter σ on the performance of the proposed method in Table 1. The table reports test results obtained by directly fixing σ in advance and skipping the validation process, in order to directly demonstrate how different σ values affect the prediction sets on the test data. When σ is close to zero, the smoothing effect is negligible, and the results are nearly identical to those of the original CD-split method. As σ increases, the smoothing effect gradually strengthens, effectively removing spurious modes in the estimated density. When σ becomes very large (*e.g.*, $\sigma = 10$), our method degenerates to producing a single, broad prediction interval. The number of disjoint intervals decreases smoothly as σ grows. While in practice increasing the smoothing parameter does not always guarantee such a strictly monotone decrease, we find empirically that this pattern holds in most cases. Therefore, by appropriately setting the range of candidate σ values and applying our validation process, we make the resulting prediction sets match the pre-specified target number of intervals, achieving a desirable balance between efficiency and interpretability.

Ablation on smoothing techniques. Table 4 presents the experimental results obtained with different smoothing techniques. The experimental results imply that several smoothing techniques perform well under our framework and achieve prediction sets whose number of intervals is close to the target number. This demonstrates that our framework is general, allowing users to choose a smoothing technique suited to the characteristics of their specific application to obtain more interpretable

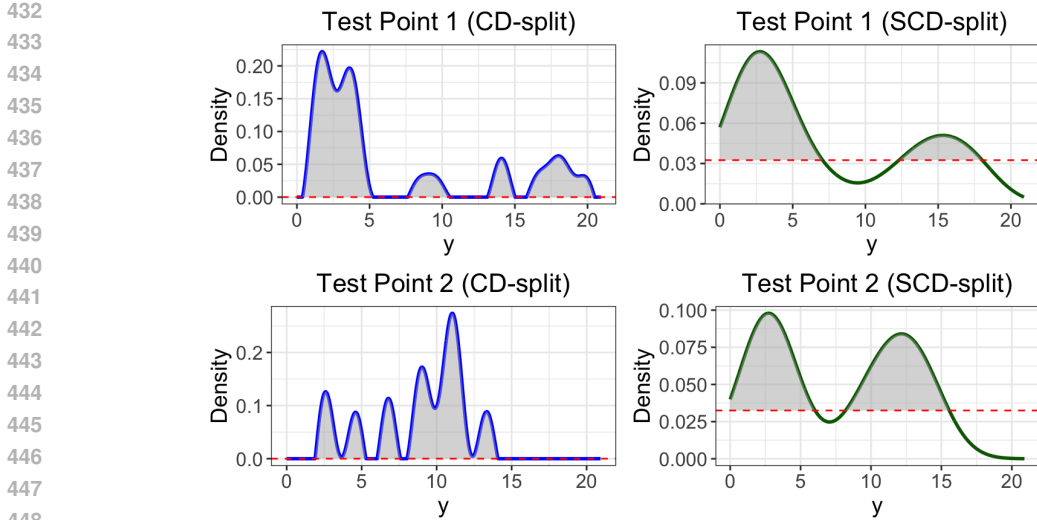


Figure 3: Illustration of prediction set construction on the first two test points from the `Bio` dataset. In each panel, the solid curve represents the estimated conditional density ($\hat{f}(y | x)$), the dashed red line indicates the threshold, and the shaded region marks the resulting prediction sets. As shown in the two left panels, $\hat{f}(y | x)$ equals zero over large regions, driving the threshold to zero and causing CD-split to return the entire output domain of y as the prediction set. By contrast, the right two panels demonstrate how SCD-split produces a smoother density function, which breaks the ties and restores coverage at the desired level $1 - \alpha$, thereby yielding more interpretable prediction sets.

prediction sets. Moreover, we observe empirically that Fourier smoothing yields prediction sets with smaller average lengths and thus better interpretability compared with other smoothing techniques when the smoothing parameter is chosen properly. Therefore, we choose Fourier smoothing as the primary technique in our framework.

Extreme case. We observe that on real-world datasets, conditional density estimation may perform poorly, and both CD-split and HPD-split are highly sensitive to such estimation errors. For instance, in `Bio` dataset, the experimental result in Table 3 implies that these errors can lead to overly large prediction sets and poor interpretability. The reason is that when the estimated density is poor, a large portion of calibration responses y_i may fall in regions where $\hat{f}(y_i | x_i) = 0$, resulting in conformity scores collapsing to zero. If the proportion of such ties exceeds α , the resulting prediction sets may degenerate into the entire output domain, yielding empirical coverage close to 100%. A common solution is to assign a small random value to break these ties and restore the coverage of $1 - \alpha$. However, this typically leads to poorer interpretability. To address this issue, our methods apply smoothing operation to \hat{f} , which effectively merges spurious zero-density regions into surrounding modes. This process breaks excessive ties and restores a more meaningful distribution of conformity scores, allowing the coverage to return to the target level of $1 - \alpha$ while keeping the prediction sets informative. In this case, smoothing may increase the interval number, which appears at first sight to conflict with our theoretical guarantee. However, this apparent discrepancy is explained by the violation of an assumption in Theorem 4.2, namely that the set $A_t := \{x : f(x) = t, f'(x) \neq 0\}$ is finite for all $t \in \mathcal{B}$. We refer to Figure 3 for illustration.

6 CONCLUSION

In this work, we propose a smoothing-based framework to enhance the interpretability of prediction sets, particularly for methods based on conditional density estimation. Theoretically, we show that smoothing preserves desirable properties of the prediction sets. Empirically, our method achieves a favorable trade-off between interval length and number across both synthetic and real-world datasets, thereby substantially improving the interpretability of the prediction sets. Consequently, this paper suggests that smoothing serves as a practical enhancement for conformal prediction methods based on conditional density estimation. It might be interesting to explore task-specific smoothing techniques that adaptively balance interval length and connectivity for interpretability in future work.

REFERENCES

- 486
487
488 Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad
489 Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir
490 Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Tech-
491 niques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535.
492 doi: <https://doi.org/10.1016/j.inffus.2021.05.008>. URL <https://www.sciencedirect.com/science/article/pii/S1566253521001081>.
493
- 494 Luca Ambrogioni, Umut Güçlü, Marcel A. J. van Gerven, and Eric Maris. The kernel mixture
495 network: A nonparametric method for conditional density estimation of continuous random vari-
496 ables, 2017. URL <https://arxiv.org/abs/1705.07111>.
- 497 Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets
498 for image classifiers using conformal prediction, 2022. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2009.14193)
499 [2009.14193](https://arxiv.org/abs/2009.14193).
- 500 Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction
501 and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2107.07511)
502 [2107.07511](https://arxiv.org/abs/2107.07511).
503
- 504 Anastasios N Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets
505 for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.
- 506 Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of
507 distribution-free conditional predictive inference. *Information and Inference: A Journal of the*
508 *IMA*, 2020.
509
- 510 Batiste Le Bars and Pierre Humbert. On Volume Minimization in Conformal Regression, February
511 2025. URL <http://arxiv.org/abs/2502.09985>. arXiv:2502.09985 [stat].
- 512 Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statis-*
513 *tics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.
- 514 José E. Chacón, Pablo Monfort, and Carlos Tenreiro. Fourier methods for smooth distribution
515 function estimation, 2013. URL <https://arxiv.org/abs/1305.2476>.
516
- 517 Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan
518 Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM*
519 *transactions on intelligent systems and technology*, 15(3):1–45, 2024.
- 520 Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via random-
521 ized smoothing, 2019. URL <https://arxiv.org/abs/1902.02918>.
522
- 523 Jan G. De Gooijer and Dawit Zerom. On conditional density estimation. *Statistica Neer-*
524 *landica*, 57(2):159–176, 2003. doi: <https://doi.org/10.1111/1467-9574.00226>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9574.00226>.
525
- 526 Niloufar Eghbali, Tuka Alhanai, and Mohammad M. Ghassemi. Distribution-free uncertainty quan-
527 tification in mechanical ventilation treatment: A conformal deep q-learning framework, 2024.
528 URL <https://arxiv.org/abs/2412.12597>.
529
- 530 Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity
531 measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996. ISSN 00063444,
532 14643510. URL <http://www.jstor.org/stable/2337441>.
- 533 Hadi Fanaee-T. Bike Sharing. UCI Machine Learning Repository, 2013. DOI:
534 <https://doi.org/10.24432/C5W894>.
- 535 Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal
536 prediction. In *International Conference on Learning Representations*, 2019.
537
- 538 Margarita Genius. Nonparametric econometrics: Theory and practice. *European Review of Agri-*
539 *cultural Economics*, 35(2):254–257, 06 2008. ISSN 0165-1587. doi: 10.1093/erae/jbn027. URL
<https://doi.org/10.1093/erae/jbn027>.

- 540 Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal Prediction With Con-
541 ditional Guarantees, September 2024. URL <http://arxiv.org/abs/2305.12616>.
542 arXiv:2305.12616 [stat].
- 543
- 544 Leying Guan. Localized conformal prediction: A generalized inference framework for conformal
545 prediction. *Biometrika*, 110(1):33–50, 2023.
- 546 Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models
547 with guarantees, 2024. URL <https://arxiv.org/abs/2405.10301>.
- 548
- 549 Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural
550 networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.
- 551 Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-
552 confidence predictions far away from the training data and how to mitigate the problem. In
553 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 41–50,
554 2019.
- 555
- 556 Roy Hirsch and Jacob Goldberger. A conformalized learning of a prediction set with applications to
557 medical imaging classification, 2024. URL <https://arxiv.org/abs/2408.05037>.
- 558
- 559 Nhat Ho and Stephen G. Walker. Multivariate smoothing via the fourier integral theorem and fourier
560 kernel, 2020. URL <https://arxiv.org/abs/2012.14482>.
- 561 Michael P. Holmes, Alexander G. Gray, and Charles Lee Isbell. Fast nonparametric conditional
562 density estimation, 2012. URL <https://arxiv.org/abs/1206.5278>.
- 563
- 564 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong
565 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language
566 models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information
567 Systems*, 43(2):1–55, 2025a.
- 568 Linhui Huang, Sayeri Lala, and Niraj K. Jha. Confine: Conformal prediction for interpretable neural
569 networks, 2025b. URL <https://arxiv.org/abs/2406.00539>.
- 570
- 571 Rob J. Hyndman, David M. Bashtannyk, and Gary K. Grunwald. Estimating and visualizing condi-
572 tional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336, 1996. ISSN
573 10618600. URL <http://www.jstor.org/stable/1390887>.
- 574
- 575 Tsuyoshi Ichimura and Daisuke Fukuda. A fast algorithm for computing least-squares cross-
576 validations for nonparametric conditional kernel density functions. *Computational Statistics
577 & Data Analysis*, 54(12):3404–3410, 2010. ISSN 0167-9473. doi: [https://doi.org/10.1016/
578 j.csda.2009.08.021](https://doi.org/10.1016/j.csda.2009.08.021). URL [https://www.sciencedirect.com/science/article/
579 pii/S0167947309003168](https://www.sciencedirect.com/science/article/pii/S0167947309003168).
- 580
- 581 Rafael Izbicki and Ann B. Lee. Converting high-dimensional regression to high-dimensional condi-
582 tional density estimation, 2017. URL <https://arxiv.org/abs/1704.08095>.
- 583
- 584 Rafael Izbicki, Gilson T. Shimizu, and Rafael B. Stern. Flexible distribution-free conditional predic-
585 tive bands using density estimators, 2019. URL <https://arxiv.org/abs/1910.05575>.
- 586
- 587 Rafael Izbicki, Gilson Shimizu, and Rafael B. Stern. Cd-split and hpd-split: efficient conformal
588 regions in high dimensions, 2021. URL <https://arxiv.org/abs/2007.12778>.
- 589
- 590 Jinsung Jeon, Hyundong Jin, Jonghyun Choi, Sanghyun Hong, Dongeun Lee, Kookjin Lee, and
591 Noseong Park. Pac-fno: Parallel-structured all-component fourier neural operators for recogniz-
592 ing low-quality images, 2024. URL <https://arxiv.org/abs/2402.12721>.
- 593
- 591 Ulf Johansson, Henrik Linusson, Tuve Ljöfström, and Henrik Boström. Interpretable regres-
592 sion trees using conformal prediction. *Expert Systems with Applications*, 97:394–404, 2018.
593 ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2017.12.041>. URL [https://www.
sciencedirect.com/science/article/pii/S0957417417308588](https://www.sciencedirect.com/science/article/pii/S0957417417308588).

- 594 Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in con-
595 formal prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Pa-
596 quet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Process-*
597 *ing Systems*, volume 37, pp. 99519–99563. Curran Associates, Inc., 2024. URL
598 [https://proceedings.neurips.cc/paper_files/paper/2024/file/
599 b41907dd4df5c60f86216b73fe0c7465-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/b41907dd4df5c60f86216b73fe0c7465-Paper-Conference.pdf).
- 600 Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction
601 and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,
602 43(11):3964–3979, 2021. doi: 10.1109/TPAMI.2020.2992934.
- 603
604 Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes
605 overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–
606 5446. PMLR, 2020.
- 607
608 Siddhartha Laghuvarapu, Zhen Lin, and Jimeng Sun. Conformal drug property prediction with den-
609 sity estimation under covariate shift, 2023. URL <https://arxiv.org/abs/2310.12033>.
- 610
611 Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression.
612 *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.
- 613
614 Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-
615 free predictive inference for regression. *Journal of the American Statistical Association*, 113
616 (523):1094–1111, 2018.
- 617
618 Charles Lu, Andreanne Lemay, Ken Chang, Katharina Hoebel, and Jayashree Kalpathy-Cramer.
619 Fair conformal predictors for applications in medical imaging, 2022. URL [https://arxiv.org/
620 abs/2109.04392](https://arxiv.org/abs/2109.04392).
- 621
622 Yuan Lu. Density-calibrated conformal quantile regression, 2024. URL [https://arxiv.org/
623 abs/2411.19523](https://arxiv.org/abs/2411.19523).
- 624
625 Ričards Marcinkevičs, Ece Ozkan, and Julia E. Vogt. Introduction to machine learning for physi-
626 cians: A survival guide for data deluge, 2022. URL [https://arxiv.org/abs/2212.
627 12303](https://arxiv.org/abs/2212.12303).
- 628
629 Ariana Martino, Michael Iannelli, and Coleen Truong. Knowledge injection to counter large lan-
630 guage model (llm) hallucination. In *European Semantic Web Conference*, pp. 182–185. Springer,
631 2023.
- 632
633 Soheil Mehrabkhani. Fourier transform approach to machine learning ii: Fourier clustering, 2019.
634 URL <https://arxiv.org/abs/1904.13241>.
- 635
636 Soheil Mehrabkhani. Fourier transform approach to machine learning iii: Fourier classification,
637 2022. URL <https://arxiv.org/abs/2001.06081>.
- 638
639 Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confi-
640 dence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer
641 vision and pattern recognition*, pp. 427–436, 2015.
- 642
643 Said Ohamouddou, Mohamed Ohamouddou, Hanaa El Afia, Abdellatif El Afia, Rafik Lasri, and
644 Raddouane Chiheb. Introducing the short-time fourier kolmogorov arnold network: A dynamic
645 graph cnn approach for tree species classification in 3d point clouds, 2025. URL [https://
646 arxiv.org/abs/2503.23647](https://arxiv.org/abs/2503.23647).
- 647
648 Ambuj Pandey and Akash Anand. Fourier smoothed pre-corrected trapezoidal rule for solution
649 of lippmann-schwinger integral equation, 2020. URL [https://arxiv.org/abs/2007.
650 06293](https://arxiv.org/abs/2007.06293).
- 651
652 Kassiani Papatotiriou, Srijan Sood, Shayleen Reynolds, and Tucker Balch. Ai in investment anal-
653 ysis: Llms for equity stock ratings. In *Proceedings of the 5th ACM International Conference on
654 AI in Finance, ICAIF ’24*, pp. 419–427. ACM, November 2024. doi: 10.1145/3677052.3698694.
655 URL <http://dx.doi.org/10.1145/3677052.3698694>.

- 648 Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim
649 Panov, and Eric Moulines. Rectifying Conformity Scores for Better Conditional Coverage, Febru-
650 ary 2025. URL <http://arxiv.org/abs/2502.16336>. arXiv:2502.16336 [stat].
651
- 652 Wei Qian, Chenxu Zhao, Yangyi Li, Fenglong Ma, Chao Zhang, and Mengdi Huai. Towards
653 modeling uncertainties of self-explaining neural networks via conformal prediction, 2024. URL
654 <https://arxiv.org/abs/2401.01549>.
- 655 Prashant Rana. Physicochemical Properties of Protein Tertiary Structure. UCI Machine Learning
656 Repository, 2013. DOI: <https://doi.org/10.24432/C5QW3H>.
- 657 Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Ad-
658 vances in neural information processing systems*, 32, 2019a.
659
- 660 Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile re-
661 gression. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence
662 d’Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Informa-
663 tion Processing Systems 32: Annual Conference on Neural Information Processing Sys-
664 tems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3538–
665 3548, 2019b. URL [https://proceedings.neurips.cc/paper/2019/hash/
666 5103c3584b063c431bd1268e9b5e76fb-Abstract.html](https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html).
- 667 Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression,
668 2019c. URL <https://arxiv.org/abs/1905.03222>.
- 669 Jonas Rothfuss, Fabio Ferreira, Simon Walther, and Maxim Ulrich. Conditional density estimation
670 with neural networks: Best practices and benchmarks, 2019. URL [https://arxiv.org/
671 abs/1903.00954](https://arxiv.org/abs/1903.00954).
- 672
- 673 Pablo Sanchez-Martin, Kinaan Aamir Khan, and Isabel Valera. Improving the interpretability of
674 gnn predictions through conformal-based graph sparsification, 2024. URL [https://arxiv.
675 org/abs/2404.12356](https://arxiv.org/abs/2404.12356).
- 676 I. J. Schoenberg. On variation-diminishing integral operators of the convolution type. *Proceedings
677 of the National Academy of Sciences*, 34(4):164–169, 1948. doi: 10.1073/pnas.34.4.164. URL
678 <https://www.pnas.org/doi/abs/10.1073/pnas.34.4.164>.
- 679 Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. In M. Ran-
680 zato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in
681 Neural Information Processing Systems*, volume 34, pp. 6304–6315. Curran Associates, Inc.,
682 2021. URL [https://proceedings.neurips.cc/paper_files/paper/2021/
683 file/31b3b31alc2f8a370206f111127c0dbd-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/31b3b31alc2f8a370206f111127c0dbd-Paper.pdf).
- 684
- 685 Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:
686 371–421, 2008. URL <https://dl.acm.org/citation.cfm?id=1390693>.
- 687 David Stutz, Krishnamurthy, Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning Op-
688 timal Conformal Classifiers, May 2022. URL <http://arxiv.org/abs/2110.09192>.
689 arXiv:2110.09192 [cs].
- 690 Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and
691 Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation.
692 In Yee Whye Teh and Mike Titterton (eds.), *Proceedings of the Thirteenth International Con-
693 ference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning
694 Research*, pp. 781–788, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL
695 <https://proceedings.mlr.press/v9/sugiyama10a.html>.
- 696
- 697 Ichiro Takeuchi, Kaname Nomura, and Takafumi Kanamori. Nonparametric conditional density
698 estimation using piecewise-linear solution path of kernel quantile regression. *Neural Compu-
699 tation*, 21(2):533–559, 02 2009. ISSN 0899-7667. doi: 10.1162/neco.2008.10-07-628. URL
700 <https://doi.org/10.1162/neco.2008.10-07-628>.
- 701 Jiaye Teng, Guang-He Lee, and Yang Yuan. 11 adversarial robustness certificates: a randomized
smoothing approach. In URL <https://openreview.net/forum>, 2020.

- 702 Jiaye Teng, Zeren Tan, and Yang Yuan. T-SCI: A two-stage conformal inference algorithm with
703 guaranteed coverage for cox-mlp. In Marina Meila and Tong Zhang (eds.), *Proceedings of the*
704 *38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event,*
705 volume 139 of *Proceedings of Machine Learning Research*, pp. 10203–10213. PMLR, 2021. URL
706 <http://proceedings.mlr.press/v139/teng21a.html>.
- 707 Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive in-
708 ference with feature conformal prediction. In *The Eleventh International Conference on Learning*
709 *Representations*, 2023. URL <https://openreview.net/forum?id=0uRm1YmFTu>.
- 710 Brian L Trippe and Richard E Turner. Conditional density estimation with bayesian normalising
711 flows, 2018. URL <https://arxiv.org/abs/1802.04908>.
- 712 Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Sci-*
713 *ence*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press,
714 2018.
- 715 Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on*
716 *machine learning*, pp. 475–490. PMLR, 2012.
- 717 Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*.
718 Springer Science & Business Media, 2005.
- 719 Baozhen Wang and Xingye Qiao. Conformal inference of individual treatment effects using condi-
720 tional density estimates, 2025. URL <https://arxiv.org/abs/2501.14933>.
- 721 Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep
722 vision transformers via the fourier domain analysis: From theory to practice, 2022. URL <https://arxiv.org/abs/2203.05962>.
- 723 Minheng Xiao and Shi Bo. Electroencephalogram emotion recognition via auc maximization, 2025.
724 URL <https://arxiv.org/abs/2408.08979>.
- 725 Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved
726 efficiency, 2024. URL <https://arxiv.org/abs/2404.19651>.
- 727 Margaux Zaffran, Aymeric Dieuleveut, Olivier Féron, Yannig Goude, and Julie Josse. Adaptive con-
728 formal predictions for time series, 2022. URL <https://arxiv.org/abs/2202.07282>.
- 729 Yanfei Zhou and Matteo Sesia. Conformal classification with equalized coverage for adaptively
730 selected groups, 2024. URL <https://arxiv.org/abs/2405.15106>.
- 731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

Appendix

In Section A, we provide some supplementary statements in this paper. In Section B, we provide several additional discussions to further contextualize and clarify the practical scope and our contributions. In Section C, we provide some detailed proofs omitted in the main context. Specifically, in Section C.1, we provide definitions of essential concepts that are repeatedly invoked in the proofs that follow. In Section C.2, we rigorously prove Theorem 4.1. In Section C.3, we rigorously prove Theorem 4.2. In Section C.4, we rigorously prove Theorem 4.3. In Section C.5, we rigorously prove Theorem 4.4. In Section C.6, we provide another special case on reducing interval number. In Section D, we provide the detailed experiment settings, results and other analysis.

A SOME SUPPLEMENTARY STATEMENTS

A.1 USE OF LARGE LANGUAGE MODELS (LLMs)

Large language models were used solely as a language-polishing tool to improve grammar, clarity, and style. They did not contribute in any way to the research ideas, theoretical developments, experimental design, data analysis, or interpretation of results. All scientific content was entirely developed by the authors.

A.2 ETHICS STATEMENT

All authors have read and comply with the ICLR Code of Ethics. Our study uses only publicly available datasets and involves no human subjects, so institutional review board (IRB) approval is not required and no privacy or security risks arise. The research introduces no potentially harmful insights or applications and poses no concerns regarding discrimination, bias, or fairness. It is free of conflicts of interest or sponsorship influences, and all data handling and experimental procedures follow accepted standards of research integrity and legal compliance. Taken together, these considerations ensure that our work satisfies the ethical requirements outlined in the ICLR Code of Ethics for data use, privacy and security, fairness, conflicts of interest, and research integrity.

A.3 REPRODUCIBILITY STATEMENT

To ensure that our results can be independently verified, we provide a complete account of the SCD-split framework and all experimental settings in the main paper (Sections 3 and 5). The appendix C contains all theoretical assumptions and full proofs of the main theorems, while the appendix D describes further experiment details including data preprocessing, parameter choices, and additional experimental results. The source code is available in the supplementary material. Together, these materials give readers all of the information and resources needed to reproduce both the theoretical and empirical findings of this work.

B ADDITIONAL DISCUSSIONS

Scope of our method: regression. Regression problems are among the most common settings in practical machine learning applications such as house price prediction, medical risk estimation, and energy demand forecasting, and classical conformal prediction methods—such as CQR (Romano et al., 2019a)—have been primarily developed for regression tasks. The method proposed in this paper is specifically tailored to regression problems, where the prediction set is continuous. In contrast, classification tasks naturally produce discrete prediction sets (i.e., subsets of labels), and thus do not suffer from the same issues of fragmented intervals or interpretability challenges that arise in regression. Therefore, the smoothing technique we introduce is meaningful primarily in the regression setting.

Why not simply merge nearby intervals? One might consider a simpler post-processing heuristic, such as merging CD-split intervals that lie within a fixed distance. However, SCD-split offers a fundamental advantage in that it fully preserves the theoretical coverage guarantees of conformal prediction. Any operation that alters prediction sets after their construction based on their geometric

810 configuration can violate the exchangeability principle between calibration and test samples, thereby
 811 invalidating the $1 - \alpha$ coverage guarantee. In contrast, SCD-split integrates the smoothing operation
 812 into the conformal procedure before the quantile computation, ensuring that all calibration and test
 813 points are treated symmetrically. This principled design provides not only empirical effectiveness
 814 but also rigorous theoretical soundness.

815 **Feasibility of the target number.** One may argue our method cannot work when the user-specified
 816 target number K_{target} exceeds the number of disjoint intervals produced by the original CD-split pro-
 817 cedure. However, in practice users typically prefer prediction sets with a relatively small number of
 818 disjoint intervals for interpretability and ease of decision making, so K_{target} is in most cases naturally
 819 modest and well below this upper bound. Moreover, if CD-split itself yields fewer disjoint intervals
 820 than K_{target} , that outcome is informative: it indicates that the underlying conditional distribution
 821 does not support as many distinct high-probability regions as the user initially expected. In such
 822 situations, our framework provides a transparent diagnostic and guidance, allowing users to simply
 823 adjust K_{target} downward so that the interpretability constraint aligns with the intrinsic complexity of
 824 the data rather than with a prior guess.

825 **Additional data split for smoothing parameter tuning.** Although reserving a small validation
 826 subset to select the smoothing parameter σ slightly reduces data efficiency, this data-driven proce-
 827 dure enables us to identify a better parameter in a principled and effective manner, leading to better
 828 overall performance. Moreover, such a design is common in conformal prediction; for instance, the
 829 RAPS method (Angelopoulos et al., 2022) similarly sets aside a small portion of data to choose its
 830 tuning parameter τ .

831 **Conditional Density Estimation (CDE).** CDE is a challenging problem in modern statistical infer-
 832 ence, especially in high-dimensional regimes (Izbicki & Lee, 2017). CDE methods can be grouped
 833 into three categories: (a) *parametric methods* assume that $p(y | x)$ follows a specific family of
 834 distributions (e.g., Gaussian, Exponential) and use maximum likelihood estimation to determine pa-
 835 rameters (Bishop, 2006); (b) *non-parametric methods* calculate the conditional density using the
 836 ratio of the joint kernel density estimate to the marginal kernel density estimate (Hyndman et al.,
 837 1996; De Gooijer & Zerom, 2003; Genius, 2008). Several works based on this method focus on
 838 using different methods to tune parameters (Ichimura & Fukuda, 2010; Holmes et al., 2012). Other
 839 approaches include different regression methods (Fan et al., 1996; Takeuchi et al., 2009; Izbicki &
 840 Lee, 2017) and least-square (Sugiyama et al., 2010); (c) *neural network based methods* combine
 841 neural networks with mixture density models called Mixture Density Networks (MDN) (Rothfuss
 842 et al., 2019) or combine neural networks with non-parametric methods called Kernel Mixture Net-
 843 works (KMN) (Ambrogioni et al., 2017). Another promising method of neural networks based CDE
 844 is normalizing flow (Trippe & Turner, 2018; Kobyzev et al., 2021).

845 **Contributions.** Our main contributions are summarized as follows:

- 846
- 847 • We introduce the number of intervals as a new evaluation metric, complementing interval
 848 length, to more comprehensively characterize the interpretability of prediction sets.
- 849
- 850 • We propose a smoothing-based method in Section 3 that regularizes the estimated condi-
 851 tional density function, reducing unnecessary peaks and ensuring that the number of dis-
 852 joint intervals in the prediction set is closer to the target number. This improves the inter-
 853 pretability and stability of the resulting prediction sets. Besides, our smoothing approach
 854 is general and can be flexibly integrated into any conformal prediction method based on
 855 conditional density estimation, including CD-split and HPD-split.
- 856
- 857 • Theoretical evidence in Section 4 shows that SCD-split is (a) valid, where the empirical
 858 coverage is larger than $1 - \alpha$, (b) efficient, where interval length is still acceptable, and (c)
 859 connective, where the number of intervals decreases under special structural cases.
- 860
- 861 • We conduct comprehensive experiments on both synthetic and real-world datasets in Sec-
 862 tion 5. The results demonstrate that our method achieves a favorable trade-off between
 863 interval length and number, leading to the better interpretability among compared methods.

C OMITTED PROOFS

C.1 SOME DEFINITIONS AND PROPOSITIONS

Definition C.1 (Randomized Smoothing). *Given base function f and input $x \in \mathbb{R}^d$, define smoothed function \tilde{f} . Specifically,*

$$\tilde{f}^{RS}(x) = \mathbb{E}(f(x + \delta)),$$

where $\delta \sim \mathcal{N}(0, \sigma^2 I_d)$. We call it σ -Randomized Smoothing. In practice, we use the Monte Carlo method to deploy (σ, n) -Randomized Smoothing as,

$$\tilde{f}_n^{RS}(x) = \frac{1}{n} \sum_{i=1}^n f(x + \delta_i), \quad \forall x \in \mathbb{R}^d,$$

where $\delta_1, \delta_2, \dots, \delta_n \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_d)$.

Definition C.2 (Fourier Smoothing with Gaussian low-pass filtering). *Define the Gaussian low-pass filtering function as*

$$H(w) = e^{-2\pi^2 \sigma^2 w^2} = e^{-\alpha w^2},$$

where the α is bandwidth. Therefore, we can formalize the Fourier smoothing with Gaussian low-pass filtering as

$$\tilde{f}^{FS}(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \left[\int_{\mathbb{R}} f(\tau) e^{-i w \tau} d\tau \right] H(w) e^{i w t} dw,$$

where \tilde{f}^{FS} is smoothed f .

Definition C.3 (Profile Distance (Izbicki et al., 2021)). *Given $\mathbf{x} \in \mathcal{X}$ and a conditional density estimator \hat{f} , we define the estimated conditional CDF*

$$\hat{H}(z | \mathbf{x}) := \int_{\{y: \hat{f}(y|\mathbf{x}) \leq z\}} \hat{f}(y | \mathbf{x}) dy.$$

The profile distance between \mathbf{x}_a and \mathbf{x}_b is the squared L^2 distance between their estimated conditional CDFs:

$$d^2(\mathbf{x}_a, \mathbf{x}_b) := \int_{-\infty}^{\infty} [\hat{H}(z | \mathbf{x}_a) - \hat{H}(z | \mathbf{x}_b)]^2 dz.$$

Proposition C.1 (Convergence to the highest predictive density set (Izbicki et al., 2021)). *The highest predictive density set, $\mathcal{C}_{1-\alpha}^*(x)$, is the region with the smallest Lebesgue measure with $1 - \alpha$ coverage:*

$$\mathcal{C}_{1-\alpha}^*(x) := \{y : f(y | x) \geq q_\alpha(x)\}, \quad \text{where } q_\alpha(x) \text{ is the } \alpha \text{ quantile of } f(Y | x).$$

A conformal prediction method converges to the highest predictive density set if:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{1-\alpha}^*(X_{n+1}) \Delta \mathcal{C}_{1-\alpha}(X_{n+1})) = o(1), \quad \text{where } A \Delta B := (A \cap B^c) \cup (B \cap A^c).$$

The CD-split method satisfies this convergence property when the estimated conditional density $\hat{f}(y | x)$ approaches the true density $f(y | x)$.

C.2 PROOF OF THEOREM 4.1

Throughout we adopt the notation of the main text. Write

$$\mathcal{D}_{n+1} := \{(\mathbf{X}_i, Y_i)\}_{i=1}^{n+1} \quad \text{and} \quad \mathcal{D}_n := \{(\mathbf{X}_i, Y_i)\}_{i=1}^n.$$

Assume \mathcal{D}_{n+1} is exchangeable. Fix a random split $\mathcal{D}_n = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{cal}}$ with $|\mathcal{D}_{\text{tr}}| = n_{\text{tr}}$ and $|\mathcal{D}_{\text{cal}}| = m := n - n_{\text{tr}}$. The SCD-split algorithm proceeds in three steps:

918 (i) **Model fitting.** Using only \mathcal{D}_{tr} we construct a conditional density estimator $\hat{f}(\cdot | \mathbf{x})$, then
 919 apply Fourier smoothing with a Gaussian kernel to obtain $\tilde{f}^{\text{FS}}(\cdot | \mathbf{x})$. Both operations are
 920 deterministic functions of \mathcal{D}_{tr} ; hence \tilde{f}^{FS} is $\sigma(\mathcal{D}_{\text{tr}})$ -measurable.

921 (ii) **Non-conformity scores.** Define the *density-level score*

$$922 S((\mathbf{x}, y); \tilde{f}^{\text{FS}}) := \tilde{f}^{\text{FS}}(y | \mathbf{x}), \quad (\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}.$$

923 Because S depends on (\mathbf{x}, y) only through the symmetric function \tilde{f}^{FS} , it is exchangeable
 924 in its arguments conditional on \mathcal{D}_{tr} . For every $(\mathbf{X}_i, Y_i) \in \mathcal{D}_{\text{cal}}$ set $S_i := S((\mathbf{X}_i, Y_i); \tilde{f}^{\text{FS}})$
 925 and for the new pair $(\mathbf{X}_{n+1}, Y_{n+1})$ set $S_{n+1} := S((\mathbf{X}_{n+1}, Y_{n+1}); \tilde{f}^{\text{FS}})$.

926 (iii) **Quantile and prediction band.** Let

$$927 q_\alpha := \text{Quantile}_\alpha(S_i : (\mathbf{X}_i, Y_i) \in \mathcal{D}_{\text{cal}}),$$

928 i.e. the $(\lceil (m+1)\alpha \rceil / (m+1))$ -th empirical quantile of the calibration scores. The SCD-
 929 split prediction band is

$$930 \mathcal{C}_{1-\alpha}^S(\mathbf{x}) := \{y \in \mathbb{R} : S((\mathbf{x}, y); \tilde{f}^{\text{FS}}) \geq q_\alpha\}.$$

931 *Proof.* Condition on the training σ -field $\mathcal{F}_{\text{tr}} := \sigma(\mathcal{D}_{\text{tr}})$. Given \mathcal{F}_{tr} , the smoothed density \tilde{f}^{FS} is
 932 fixed, while the $(m+1)$ scores

$$933 (S_i : (\mathbf{X}_i, Y_i) \in \mathcal{D}_{\text{cal}}) \quad \text{and} \quad S_{n+1}$$

934 are measurable functions of $((\mathbf{X}_i, Y_i))_{i \in \mathcal{I}_{\text{cal}} \cup \{n+1\}}$ and remain *exchangeable* because \mathcal{D}_{n+1} was
 935 exchangeable. Standard split-conformal theory (Vovk et al., 2005) then implies

$$936 \mathbb{P}(S_{n+1} \geq q_\alpha) \geq 1 - \alpha.$$

937 Equivalently,

$$938 \mathbb{P}(Y_{n+1} \in \mathcal{C}_{1-\alpha}^S(\mathbf{X}_{n+1})) \geq 1 - \alpha.$$

□

939 C.3 PROOF OF THEOREM 4.2

940 **Definition C.4** (The number of sign variations). *Let \mathcal{F} denote the space of all real-valued measurable
 941 and locally bounded functions on \mathbb{R} . For any $f \in \mathcal{F}$, we define the number of sign variations
 942 of f as*

$$943 v(f) := \sup \{ \text{VarSign}(f(x_1), f(x_2), \dots, f(x_n)) \mid n \in \mathbb{N}, x_1 < x_2 < \dots < x_n \in \mathbb{R} \},$$

944 where $\text{VarSign}(a_1, \dots, a_n)$ is defined to be the number of sign changes in the sequence (a_1, \dots, a_n)
 945 after removing all zero entries. Specifically, let $(a_{i_1}, \dots, a_{i_k})$ be the nonzero subsequence of
 946 (a_1, \dots, a_n) , and define $s_j := \text{sgn}(a_{i_j}) \in \{-1, +1\}$. Then

$$947 \text{VarSign}(a_1, \dots, a_n) := \sum_{j=1}^{k-1} \mathbb{I}(s_j \cdot s_{j+1} = -1),$$

948 where $\mathbb{I}(\cdot)$ denotes the indicator function.

949 **Definition C.5** (Variation-Diminishing Transformation). *Let $\Lambda : \mathbb{R} \rightarrow \mathbb{R}$, define operator*

$$950 T[f](x) := \int_{\mathbb{R}} \Lambda(x-t)f(t)dt.$$

951 We say T is a **variation-diminishing transformation** if $\forall f : \mathbb{R} \rightarrow \mathbb{R}$ and f is bounded and measurable,
 952 there holds

$$953 v(T[f]) \leq v(f).$$

972 **Lemma C.1** (Theorem in (Schoenberg, 1948)). *For the convolution transformation defined as*

$$973 \quad g(x) = \int_{\mathbb{R}} \Lambda(x-t)f(t)dt.$$

974 *It is variation-diminishing if Λ and f satisfies*

- 975 1. $0 < \int_{\mathbb{R}} |\Lambda(x)| dx < \infty$;
- 976 2. f is bounded and measurable.

977 *That is, if Λ and f satisfies conditions above, there holds*

$$978 \quad v(g) \leq v(f).$$

979 *Proof.* Without loss of generality, assume $t = 0$, we have

$$980 \quad M = \left\lfloor \frac{v(f) + 1}{2} \right\rfloor,$$

981 where M is the number of disconnected intervals of A_t . Since Lemma C.4 tells us that the fourier
smoothing transformation is actually Gaussian Kernel Convolution, the fourier smoothing transfor-
mation is variation-diminishing by Lemma C.1. Therefore,

$$982 \quad M' = \left\lfloor \frac{v(\tilde{f}^{\text{FS}}) + 1}{2} \right\rfloor \leq \left\lfloor \frac{v(f) + 1}{2} \right\rfloor = M.$$

983 \square

984 C.4 PROOF OF THEOREM 4.3

985 *Proof.* Without loss of generality relabel the two components as (a_1, b_1) and (a_2, b_2) with gap
 (b_1, a_2) . Let $m_{bc} := \sup_{x \in (b_1, a_2)} f(x) \leq t - \varepsilon$.

986 **Step 1: the midpoint rises above the threshold.** Set $x^* := \frac{b_1 + a_2}{2}$. By convolution,

$$987 \quad \tilde{f}^{\text{FS}}(x^*) = \int_{\mathbb{R}} \phi_{\sigma}(x^* - s)f(s) ds = \int_{(a_1, b_1) \cup (a_2, b_2)} \phi_{\sigma}(x^* - s)f(s) ds + \int_{b_1}^{a_2} \phi_{\sigma}(x^* - s)f(s) ds.$$

988 Since $f(s) \leq m_{bc}$ on the valley, we have

$$989 \quad \tilde{f}^{\text{FS}}(x^*) \geq (t - m_{bc}) \int_{|u| \geq \delta/2} \phi_{\sigma}(u) du + m_{bc}.$$

990 **Step 2: the whole valley rises above the threshold.** The kernel ϕ_{σ} is continuous, strictly positive
and unimodal, hence \tilde{f}^{FS} attains its minimum on $[b_1, a_2]$ at the endpoints. But for $s \in (a_1, b_1) \cup$
 (a_2, b_2) the integrand contribution to $\tilde{f}^{\text{FS}}(b_1)$ and $\tilde{f}^{\text{FS}}(a_2)$ is no smaller than at x^* , so

$$991 \quad \tilde{f}^{\text{FS}}(b_1) \geq t, \quad \tilde{f}^{\text{FS}}(a_2) \geq t.$$

992 Continuity of \tilde{f}^{FS} yields $\tilde{f}^{\text{FS}}(x) \geq t$ for every $x \in [b_1, a_2]$. Thus $(a_1, b_2) \subseteq B_t$ and the two
components merge.

993 **Step 3: counting components.** At least one pair of components of A_t has merged, so $M' \leq$
 $M - 1$. \square

1000 C.5 PROOF OF THEOREM 4.4

1001 C.5.1 TECHNIQUE LEMMA

1002 **Lemma C.2** (Hoeffding's Inequality, Theorem 2.6.2 in (Vershynin, 2018)). *Let X_1, X_2, \dots, X_n be*
zero-mean independent sub-Gaussian random variables. Then, for any $t > 0$, we have

$$1003 \quad \mathbb{P} \left(\left| \sum_{i=1}^n X_i \right| \geq t \right) \leq 2 \exp \left(- \frac{ct^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2} \right),$$

where $c > 0$ is an absolute constant and $\|\cdot\|_{\psi_2}$ is sub-Gaussian norm defined by:

$$\|X\|_{\psi_2} := \inf \left\{ c \geq 0 : \mathbb{E} \left(e^{X^2/c^2} \right) \leq 2 \right\}.$$

Lemma C.3. Assume $f : \mathbb{R} \rightarrow \mathbb{R}$ is L -Lipschitz. $\forall \eta \in (\frac{1}{2}, 1)$, there exist $C > 0$ an absolute constant, such that with probability $1 - \eta$,

$$\|f - \tilde{f}_n^{RS}\|_\infty < L\sigma \sqrt{\frac{\log(2/\eta)}{Cn}} + \frac{2}{\pi},$$

where \tilde{f}_n^{RS} is (σ, n) -Randomized Smoothed function.

Proof. $\forall x \in \mathbb{R}^d$, there holds

$$\begin{aligned} |f(x) - \tilde{f}_n^{RS}(x)| &\leq \frac{1}{n} \sum_{i=1}^n |f(x + \delta_i) - f(x)| \\ &\leq \frac{1}{n} \sum_{i=1}^n L |\delta_i| \\ &\leq \frac{L}{n} \left| \sum_{i=1}^n (|\delta_i| - \mathbb{E}|\delta_i|) \right| + L\mathbb{E}|\delta_i|. \end{aligned}$$

Bound the first term using Lemma C.2. $\forall \eta \in (\frac{1}{2}, 1)$, there exists $C > 0$, such that,

$$\mathbb{P} \left(\frac{L}{n} \left| \sum_{i=1}^n (|\delta_i| - \mathbb{E}|\delta_i|) \right| \geq \sqrt{\frac{L^2 \sigma^2 \log(2/\eta)}{Cn}} \right) \leq \eta.$$

Since $\mathbb{E}|\delta_i| = \sigma\sqrt{2/\pi}$ and the upper bound is consistent for all $x \in \mathbb{R}$, we have

$$\|f - \tilde{f}_n^{RS}\|_\infty < L\sigma \sqrt{\frac{\log(2/\eta)}{Cn}} + \frac{2}{\pi}$$

holds with probability $1 - \eta$. \square

Lemma C.4. Fourier smoothing with Gaussian low-pass filtering is equivalent to the case of randomized smoothing on large samples. Specifically, assume f is L -Lipschitz, if $\sigma^2 = 2\alpha$, there holds

$$\|\tilde{f}_n^{RS} - \tilde{f}^{FS}\|_\infty \xrightarrow{n \rightarrow \infty} 0,$$

where σ^2 denotes the noise variance of randomized smoothing and α denotes the bandwidth of Fourier smoothing.

Proof. Let

$$K(t - \tau) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\alpha w^2} e^{iw(t-\tau)} dw,$$

we have

$$\tilde{f}^{FS}(t) = \int_{\mathbb{R}} f(\tau) K(t - \tau) d\tau.$$

Obviously, it is a kernel function and we convert the Fourier transform and inverse Fourier transform process into a convolution form. We can calculate this kernel function. Let $s = t - \tau$

$$\begin{aligned} K(s) &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\alpha w^2} e^{iws} ds \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\alpha w^2 + iws} ds \\ &= \frac{1}{2\sqrt{\pi\alpha}} \exp \left\{ -\frac{s^2}{4\alpha} \right\}. \end{aligned}$$

Therefore, this kernel function is a Gaussian kernel with $\sigma^2 = 2\alpha$. Then let's investigate randomized smoothing.

$$\begin{aligned}\tilde{f}_n^{\text{RS}} &= \int_{\mathbb{R}} f(x + \delta)p(\delta)d\delta \\ &\stackrel{t=x+\delta}{=} \int_{\mathbb{R}} f(t)p(t-x)dt \\ &= \int_{\mathbb{R}} f(t)p(x-t)dt,\end{aligned}$$

where $p(x)$ is the density function of the Gaussian distribution with σ^2 as the variance. Therefore, we have

$$\|\tilde{f}_n^{\text{RS}} - \tilde{f}^{\text{FS}}\|_{\infty} \leq \|\tilde{f}_n^{\text{RS}} - \tilde{f}^{\text{RS}}\|_{\infty} + \|\tilde{f}^{\text{RS}} - \tilde{f}^{\text{FS}}\|_{\infty} = \|\tilde{f}_n^{\text{RS}} - \tilde{f}^{\text{RS}}\|_{\infty}.$$

Let $Z_n(x) = \tilde{f}_n^{\text{RS}}(x) - \tilde{f}^{\text{RS}}(x) = \int_{\mathbb{R}} f(x+t)(\mu_n - \mu)dt$, where $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{\delta_i}$ is an empirical measure and μ denotes the Gaussian probabilistic measure, $\forall x \in \mathbb{R}$, we have

$$\begin{aligned}Z_n(x) &= \int_{\mathbb{R}} f(x+t)(\mu_n - \mu)(dt) \\ &= \int_{\mathbb{R}} (f(x+t) - f(x))(\mu_n - \mu)(dt).\end{aligned}$$

Therefore,

$$\begin{aligned}|Z_n(x)| &\leq \int_{\mathbb{R}} |f(x+t) - f(x)|(\mu_n - \mu)(dt) \\ &\leq \int_{\mathbb{R}} L|t|(\mu_n - \mu)(dt) \\ &= \frac{1}{n} \sum_{i=1}^n |\delta_i| - \mathbb{E}|\delta_i|.\end{aligned}$$

Since the upper bound is consistent w.r.t. x , by law of large number, we have

$$\|\tilde{f}_n^{\text{RS}} - \tilde{f}^{\text{FS}}\|_{\infty} \leq \|\tilde{f}_n^{\text{RS}} - \tilde{f}^{\text{RS}}\|_{\infty} = \|Z_n\|_{\infty} \xrightarrow{n \rightarrow \infty} 0.$$

□

C.5.2 FULL PROOF OF THEOREM 4.4

Proof. For simplicity, we assume the density function is unimodal at first. Let $\{i_1, \dots, i_{n_j}\} = \{i : X_i \in A(x_{n+1})\}$, $U_l = \hat{f}(y_{i_l} | x_{i_l})$, for $l = 1, \dots, n_j$, and $U_{n_j+1} = \hat{f}(y_{n+1} | x_{n+1})$. Similarly, let $\tilde{U}_k = \tilde{f}^{\text{FS}}(y_{i_k} | x_{i_k})$, for $k = 1, \dots, n_j$, and $\tilde{U}_{n_j+1} = \tilde{f}^{\text{FS}}(y_{n+1} | x_{n+1})$. By Lemma C.3 and Lemma C.4, we have with probability $1 - \eta$,

$$|U_i - \tilde{U}_i| < L\sigma\sqrt{\frac{2}{\pi}}, \quad \forall i \in 1, \dots, n_j + 1.$$

Let t be the oracle original quantile threshold and $\varepsilon = L\sigma\sqrt{2/\pi}$, we have

$$\begin{aligned}\mathbb{P}(\tilde{U}_{n+1} \geq t - \varepsilon | x_{n+1}) &\stackrel{(i)}{\geq} \mathbb{P}(U_{n+1} \geq t | x_{n+1}) \\ &\stackrel{(ii)}{=} \mathbb{P}(U_i \geq t | x_i) \\ &\geq 1 - \alpha,\end{aligned}$$

where (i) holds since the error control of U_l and (ii) follows from the definition of the profile of the density. Therefore, let $g = f + \varepsilon$. Since the density function is unimodal, $f(y_{n+1} | x_{n+1}) = t$ only has two solution. Denote $[\ell, u] = \{y : f(y | x_{n+1}) \geq t\}$ and $[\tilde{\ell}, \tilde{u}] = \{y : g(y | x_{n+1}) \geq t - \varepsilon\}$, there holds

$$|\tilde{t} - t| \leq \left| |\tilde{\ell} - \tilde{u}| - |\ell - u| \right| \leq \left| \tilde{\ell} - \ell \right| + \left| \tilde{u} - u \right| \stackrel{(i)}{\leq} 4 \frac{|f - g|}{M} = \frac{4\varepsilon}{M} = \frac{4L\sigma}{M} \sqrt{\frac{2}{\pi}},$$

where (i) holds since $|\widehat{f}(y_1 | x) - \widehat{f}(y_2 | x)| \geq M |y_1 - y_2|$. Since the number of disconnected intervals returned by CD-split is N and the result in Theorem 4.2 demonstrates that the number of disconnected intervals doesn't increase, the result of the multimodel distribution is as follows

$$|\tilde{l} - l| \leq \frac{4NL\sigma}{M} \sqrt{\frac{2}{\pi}}. \quad (10)$$

□

C.6 ANOTHER SPECIAL CASE ON REDUCING INTERVAL NUMBER: HIGH-FREQUENCY SMALL-AMPLITUDE PERTURBATIONS

The following result shows that when the disconnected components of A_t are created *solely* by a high-frequency oscillatory perturbation, Gaussian-kernel Fourier smoothing suppresses those oscillations and thus strictly reduces the component count.

Definition C.6 ((σ, t)-HF perturbation). *Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be bounded, measurable and differentiable and fix $t \in \mathcal{T}$. Set the safety gap*

$$\Delta(t, g) := \inf_{x \in \mathbb{R}} |g(x) - t| \in (0, \infty). \quad (\dagger)$$

For parameters $\varepsilon > 0$, $k > 0$, $\sigma > 0$ we call

$$f(x) := g(x) + \varepsilon \sin(kx)$$

an (σ, t)-high-frequency perturbation of g if

- (i) $\Delta(t, g) < \varepsilon$ (the oscillation amplitude is large enough to cross the threshold);
- (ii) the attenuated amplitude after Gaussian convolution,

$$\varepsilon_\sigma := \varepsilon e^{-2\pi^2\sigma^2k^2},$$

satisfies $\varepsilon_\sigma < \Delta(t, g)$ (the residual amplitude is too small to cross t).

Theorem C.1 (Strict reduction for an HF perturbation). *Adopt the setting of Definition C.6 and write $A_t(f) = \bigsqcup_{j=1}^M (a_j, b_j)$ and $B_t(f) = \bigsqcup_{j=1}^{M'} (a'_j, b'_j)$ for $f(x) = g(x) + \varepsilon \sin(kx)$ and its Fourier-smoothed version $\tilde{f}^{\text{FS}} = T_\sigma[f] = \phi_\sigma * f$, respectively. Then*

$$M' = M(v(g)) \quad \text{and} \quad M' < M,$$

hence the number of disconnected intervals strictly decreases.

Proof. Step 1: Because $\Delta(t, g) > 0$, the function g stays uniformly away from the threshold, so $g(x) - t$ keeps a fixed sign. Consequently $v(g) = 0$ and $M(v(g)) = \lfloor (0 + 1)/2 \rfloor = 0$ or 1. Denote this value by M_g .

Step 2: creation of extra components before smoothing. Since $\varepsilon > \Delta(t, g)$, the oscillatory term produces at least two distinct roots of $f(x) - t$ within every interval of length $2\pi/k$ where $|g(x) - t| < \varepsilon$. Hence $v(f) \geq 2$ and $M \geq \lfloor (2 + 1)/2 \rfloor = 1 + M_g$; in particular $M > M_g$.

Step 3: destruction of the extra components after smoothing. Because $\tilde{f}^{\text{FS}}(x) = g(x) + \varepsilon_\sigma \sin(kx)$ and $\varepsilon_\sigma < \Delta(t, g)$, we have $\text{sgn}(\tilde{f}^{\text{FS}}(x) - t) = \text{sgn}(g(x) - t)$ for all x . Thus $v(\tilde{f}^{\text{FS}}) = v(g)$ and $M' = \lfloor (v(g) + 1)/2 \rfloor = M_g$.

Step 4: comparison. Combining Steps 2 and 3 gives $M' < M$, completing the proof. □

D EXPERIMENT DETAILS

D.1 EXPERIMENT SETTINGS AND RESULTS

Synthetic data. First we introduce the simple case. The covariate vector $X = (X_1, \dots, X_5)$ is sampled *i.i.d.* from $\text{Unif}(-5, 5)$ and standardized, and the response variable Y given X follows

$$Y | X \sim \frac{1}{3} \mathcal{N}(0 + 0.1X_1, 0.2^2) + \frac{1}{3} \mathcal{N}(1.0 + 0.1X_1, 0.2^2) + \frac{1}{3} \mathcal{N}(2.0 + 0.1X_1, 0.2^2).$$

Table 2: Results on Synthetic Datasets

Method	simple			complex		
	Cov.	Len.	Num.	Cov.	Len.	Num.
Vanilla CP	90.32 ± 0.95	2.45 ± 0.03	1.00 ± 0.00	89.98 ± 0.59	28.70 ± 0.69	1.00 ± 0.00
CQR	90.28 ± 0.76	2.44 ± 0.02	1.00 ± 0.00	90.36 ± 0.50	27.48 ± 1.55	1.00 ± 0.00
Local CP	89.79 ± 0.89	2.46 ± 0.03	1.00 ± 0.00	90.14 ± 0.69	27.06 ± 1.17	1.00 ± 0.00
DIST	89.60 ± 0.91	2.45 ± 0.02	1.00 ± 0.00	90.45 ± 2.64	23.20 ± 3.99	1.00 ± 0.00
CD-split	89.52 ± 1.02	2.23 ± 0.05	2.60 ± 0.12	91.06 ± 3.55	21.76 ± 6.74	2.85 ± 1.18
HPD-split	89.93 ± 0.85	2.25 ± 0.04	2.71 ± 0.71	92.66 ± 4.89	23.32 ± 8.86	3.04 ± 1.60
SCD-split	89.09 ± 0.90	2.37 ± 0.04	2.00 ± 0.06	89.39 ± 0.85	16.11 ± 0.82	1.99 ± 0.10

Table 3: Results on Real-world Datasets

Method	bio			bike		
	Cov.	Len.	Num.	Cov.	Len.	Num.
Vanilla CP	90.33 ± 0.85	2.12 ± 0.12	1.00 ± 0.00	90.20 ± 1.11	1.23 ± 0.14	1.00 ± 0.00
CQR	89.70 ± 0.91	1.64 ± 0.11	1.00 ± 0.00	89.87 ± 0.92	0.98 ± 0.09	1.00 ± 0.00
Local CP	90.06 ± 0.67	1.90 ± 0.11	1.00 ± 0.00	89.79 ± 0.95	1.01 ± 0.08	1.00 ± 0.00
DIST	90.23 ± 2.17	1.90 ± 0.22	1.00 ± 0.00	89.28 ± 1.41	0.31 ± 0.01	1.00 ± 0.00
CD-split	96.59 ± 2.39	2.29 ± 0.25	1.36 ± 0.55	86.76 ± 1.11	0.22 ± 0.01	1.07 ± 0.01
HPD-split	98.76 ± 3.37	2.61 ± 0.29	1.37 ± 1.19	89.18 ± 0.84	0.20 ± 0.01	1.11 ± 0.02
SCD-split	89.23 ± 0.95	1.52 ± 0.06	1.49 ± 0.05	89.00 ± 1.34	0.25 ± 0.01	1.01 ± 0.01

Second, we introduce the complex case. The covariate vector $X = (X_1, \dots, X_5)$ is sampled *i.i.d.* from $\mathcal{N}(0, 1)$ and standardized, and the response variable Y given X follows

$$Y | X \sim \sum_{k=1}^K \frac{\exp(X^\top \beta_k)}{\sum_{j=1}^K \exp(X^\top \beta_j)} \mathcal{N}(\mu_{\text{base},k} + X^\top \gamma_k, \sigma_k^2),$$

where $K = 7$ is the number of Gaussian mixture components, and the base means and standard deviations are set as

$$\mu_{\text{base}} = (-15, -10, -5, 0, 5, 10, 15), \quad \sigma = (1, 1.2, 1.5, 1, 1.5, 1.2, 1).$$

The coefficient matrices $\beta_k \in \mathbb{R}^d$ and $\gamma_k \in \mathbb{R}^d$ are randomly initialized with entries drawn from $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 0.5^2)$, respectively, and they control the conditional mixture weights and component-wise shifts in the conditional means.

Real-world data. We evaluate our method on two widely used real-world datasets. The `Bike Sharing` dataset contains 10,886 samples and 18 variables. The `Bio` dataset comprises 45,730 samples and 9 variables.

Experimental setup. For all experiments, we randomly draw 2,000 samples for conformal prediction (equally divided between training and calibration sets) and 5,000 samples for testing. All features are standardized before model fitting. We use random forest as the base model for conditional density estimation in all the experiments. Each experiment is repeated across 10 independent trials to ensure statistical reliability. All experiments are conducted using standard CPU environments without the need for GPUs, with modest runtime requirements well within a practical and reproducible range.

Table 4: Different smoothing techniques on synthetic complex dataset

Method	Coverage (%)	Length	Number of Intervals
Original	91.06 ± 3.55	21.76 ± 6.74	2.85 ± 1.18
Fourier	89.23 ± 0.77	16.11 ± 0.68	1.99 ± 0.01
Gaussian kernel	89.40 ± 0.87	16.90 ± 1.56	1.95 ± 0.05
Spline	89.30 ± 0.78	16.37 ± 0.81	1.95 ± 0.12
LOESS	89.46 ± 0.80	16.95 ± 2.11	1.98 ± 0.01

Table 5: Different loss function on synthetic complex dataset

Method / σ	Global L1	Global L2	MAE	MSE
CD-split ($\sigma = 0$)	0.80 ± 0.29	0.93 ± 1.08	1.12 ± 0.14	2.06 ± 2.02
SCD-split ($\sigma = 1$)	0.38 ± 0.06	0.21 ± 0.04	0.77 ± 0.03	1.09 ± 0.17
SCD-split ($\sigma = 1.5$)	0.15 ± 0.03	0.05 ± 0.01	0.59 ± 0.00	0.66 ± 0.00
SCD-split ($\sigma = 2$)	0.33 ± 0.02	0.13 ± 0.01	0.56 ± 0.00	0.57 ± 0.00
SCD-split ($\sigma = 5$)	0.81 ± 0.00	0.67 ± 0.00	0.81 ± 0.00	0.81 ± 0.00
SCD-split ($\sigma = 10$)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

D.2 OTHER ANALYSIS ON EXPERIMENTS

Choice of loss for validation. In Table 5, we evaluate performance on the validation set using different loss function for each candidate smoothing parameter σ . Here we consider four loss functions:

$$R_{\text{Global-L1}}(\sigma) = \left| \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(\mathbf{X}_j, Y_j) \in \mathcal{D}_{\text{val}}} N_\sigma(\mathbf{X}_j) - K_{\text{target}} \right|,$$

$$R_{\text{Global-L2}}(\sigma) = \left(\frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(\mathbf{X}_j, Y_j) \in \mathcal{D}_{\text{val}}} N_\sigma(\mathbf{X}_j) - K_{\text{target}} \right)^2,$$

$$R_{\text{MAE}}(\sigma) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(\mathbf{X}_j, Y_j) \in \mathcal{D}_{\text{val}}} |N_\sigma(\mathbf{X}_j) - K_{\text{target}}|,$$

$$R_{\text{MSE}}(\sigma) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(\mathbf{X}_j, Y_j) \in \mathcal{D}_{\text{val}}} (N_\sigma(\mathbf{X}_j) - K_{\text{target}})^2.$$

For L_1 -type and L_2 -type losses, the table reports these losses as functions of σ and demonstrates that they exhibit the same trend across all candidates, leading to the same choice of $\hat{\sigma}$. Hence, the selection of the smoothing parameter is insensitive to whether L_1 or L_2 is used. In our main algorithm, we therefore adopt the L_1 loss for validation. *Moreover*, the distinction between *Global* (outer) and *Inner* (MAE/MSE) aggregation reflects different goals: the global losses aim to match the target number of intervals in an average sense, whereas the inner losses measure how close each individual prediction set is to the target. Therefore, one can also adopt MAE or MSE and correspondingly tune the smoothing parameter if the goal is to make the number of intervals for every single test point as close as possible to K_{target} .

Stability Across Trials. We further evaluate the robustness of each method by examining the variability of performance across multiple random trials. Specifically, we measure the standard deviation of coverage, interval length, and interval number across different runs. Our results show that after applying smoothing techniques, the standard deviations of all three metrics are consistently reduced across both synthetic and real-world datasets. This indicates that smoothing not only improves the interpretability and efficiency of the prediction sets, but also enhances the stability of the predictions, making the results more reliable under different random splits or sampling variations.