SPLIT CONFORMAL PREDICTION WITH TUNABLE PREDICTION SET SIZE

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal Prediction (CP) is a model-agnostic framework for constructing prediction sets with finite-sample coverage guarantees and has become a key tool for reliable decision-making in high-risk applications. However, a limitation of CP is that the size of the prediction set is implicitly determined by the target coverage level, which can result in sets that are too large to be actionable. This is particularly problematic in applications such as medical diagnosis, where there are many possible conditions and decisions must be made under time constraints. To address this limitation, recent work studies CP procedures in the *Tunable Prediction Set Size* (TPSS) setting, wherein the user specifies a maximum allowable prediction set size, and the goal is to provide valid coverage guarantees under this constraint. While Split CP (sCP) is the most widely used variant of CP, enforcing such size control within this framework is challenging due to its quantile-based calibration procedure. As a result, existing approaches for TPSS rely on the more flexible Conformal e-Prediction (eCP) framework, which enables data-dependent size control but inherits limitations such as weaker structural properties and less accurate coverage approximation. Motivated by this gap, we develop *Tunable-sCP (TsCP)*, a method to address the TPSS problem within the sCP framework, enabling explicit control over prediction set size while preserving the key structural properties of sCP. We further show that our leave-one-out construction yields a substantially more accurate coverage estimate than eCP for the same calibration size, reducing conservativeness and making it more reliable in practice. Empirically, we demonstrate on synthetic and real-world datasets that TsCP consistently improves the empirical coverage–efficiency trade-off compared to prior approaches.

1 INTRODUCTION

Conformal Prediction (CP) (Papadopoulos et al., 2002; Vovk et al., 2005; Lei & Wasserman, 2014; Thurin et al., 2025; Narteni et al., 2025; Patel et al., 2025) is a widely used framework for uncertainty quantification that operates under minimal assumptions, being both model-agnostic and distribution-free. At a high level, CP transforms the predictions of any underlying model into a set of plausible outcomes by calibrating prediction errors using past data. This procedure yields prediction sets with rigorous finite-sample guarantees, ensuring that, over repeated samples, the true outcome is contained in the set with probability at least a user-specified coverage level. CP has been successfully applied across a broad range of domains, including medical imaging (Lu et al., 2022a;b; Kutiel et al., 2023), classical machine learning (Balasubramanian et al., 2014; Johansson et al., 2014), deep learning (Papadopoulos et al., 2007; Wieslander et al., 2020; Zhang et al., 2021; Zargarbashi et al., 2023), diffusion models (Teneggi et al., 2023; Sun et al., 2023; Wang et al., 2024; Hu et al., 2025b), large language models (Cherian et al., 2024; Vishwakarma et al., 2025), image segmentation (Gade et al., 2024; Mossina & Friedrich, 2025; Chen et al., 2025; Cheung et al., 2025), and federated learning (Lu & Kalpathy-Cramer, 2021; Lu et al., 2023).

A widely used variant of CP is *Split Conformal Prediction* (sCP), which is standard in practice due to its simplicity and computational efficiency (Papadopoulos, 2008; Angelopoulos & Bates, 2021). sCP partitions the data into a training set and a calibration set, using the former to train the model and the latter to quantify prediction errors and calibrate uncertainty, thereby producing prediction sets that are easy to compute while preserving finite-sample coverage guarantees. A closely related variant is *Conformal e-value Prediction* (eCP), which follows a similar data-splitting strategy but

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

replaces the p-value–based construction with e-values, enabling additional flexibility such as handling data-dependent miscoverage levels up to a first-order Taylor approximation (Vovk, 2025; Gauthier et al., 2025b). However, this flexibility comes at the cost of departing from the sCP framework: the resulting guarantees are less directly interpretable as coverage probabilities, and the procedures no longer retain key structural properties of classical conformal methods, such as the locality of predictions and the independence behavior of prediction errors (Vovk, 2025; 2002).

Despite the advantages of classical CP methods, a key limitation is that the size of the prediction set is implicitly determined by the chosen coverage level. In many applications, this can lead to prediction sets that are too large to be used in downstream tasks. For instance, in *medical diagnosis*, overly large prediction sets may offer little practical guidance for decision-making under time constraints (Kononenko, 2001), while in *demand forecasting*, overly large prediction sets can be difficult to translate into actionable decisions, limiting their usefulness for planning and resource allocation (Carbonneau et al., 2008). To address these limitations, recent work has focused on developing CP procedures that allow users to explicitly control the maximum size of prediction sets while retaining a coverage guarantee under this constraint. We refer to this setting as the *Tunable Prediction Set Size* (TPSS) problem. Due to the inherent difficulty of controlling prediction set size within the sCP framework, prior work has addressed TPSS using the more flexible eCP variant. In particular, (Gauthier et al., 2025b;a) propose *Backward Conformal Prediction*, which leverages the data-dependent nature of eCP to directly enforce size constraints and achieve the TPSS objective. However, the resulting procedure also inherits the limitations of eCP, leading to less favorable average coverage and efficiency, as well as substantially looser bounds under the leave-one-out construction across calibration sizes.

Contribution. Motivated by this gap, we develop *Tunable-sCP*, a method to address the TPSS problem within the sCP framework, enabling explicit control over prediction set size while preserving the structural properties of sCP. Leveraging these properties, we introduce a leave-one-out approximation of the expected-value terms in the coverage guarantee and show that, within our framework, it is both consistent and has decaying variance, ultimately yielding substantially more accurate empirical coverage estimates than existing eCP approaches. The paper is organized as follows:

1. In Section 3, we briefly review how eCP can be naturally adapted to the TPSS setting. We then show that this approach does not extend to the standard sCP framework.
2. In Section 4, we present our method, *Tunable-sCP*, for addressing the TPSS problem within the sCP framework. Under mild assumptions, we show that the resulting prediction sets achieve coverage guarantees while satisfying the prescribed size constraint.
3. In Section 5, we introduce a LOO procedure to approximate the expected-value terms in the coverage guarantee. We show that, by exploiting the properties of sCP, the resulting approximation is consistent and has decaying variance.
4. In Section 6, we present empirical results on both synthetic and real-world datasets, for both continuous and discrete settings, to evaluate the empirical behavior associated with our theoretical results, including coverage, size control, and the convergence and variance properties of the LOO estimator. We also compare our method with existing approaches for the TPSS problem and conduct ablation studies.

Section 2 reviews background on CP and its variants, as well as the TPSS problem. A table of notation, additional background on CP, comparisons with the eCP approach for the TPSS problem, theoretical details, and further experimental results are deferred to the Appendix.

2 PRELIMINARIES

We review Split Conformal Prediction (sCP), Conformal e-value Prediction (eCP), and the Tunable Prediction Set Size (TPSS) problem. *Additional details are provided in Appendix A.*

Split Conformal Prediction (sCP). Let $\hat{f} : \mathcal{X} \to \mathcal{Y}$ be a *prediction model*, and let $\alpha \in (0, 1)$ denote a user-specified *miscoverage level*. The goal of all Conformal Prediction (CP) procedures is to construct a *set-valued predictor* $C : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, where $\mathcal{P}(\mathcal{Y})$ denotes the collection of all subsets of $\mathcal{Y}$, such that for a test pair $(X_0, Y_0) \in \mathcal{X} \times \mathcal{Y}$, the following *coverage guarantee* holds:

$$\mathbb{P}\left(Y_0 \in C(X_0)\right) \geq 1 - \alpha, \tag{1}$$

sCP achieves this by partitioning the data into a training and a *calibration set* $\mathcal{C}$ (Papadopoulos, 2008). The model $\hat{f}$ is first fitted on the training data, and the calibration set is then used for CP calibration.

Formally, let n denote the size of the calibration set $\mathcal{C}$ and let $\mathcal{C} = \{(X_i, Y_i)\}_{i=1}^n$. The calibration step defines a *nonconformity score* function $s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, which quantifies the discrepancy between predictions and responses and evaluates the calibration scores $s_i := s\big(\hat{f}(X_i), Y_i\big)$, for $i \in \{1, \ldots, n\}$, where the choice of the score function s is determined by the application of interest (Papadopoulos et al., 2007; Angelopoulos & Bates, 2021; Andéol et al., 2023). A data-dependent threshold is then obtained from the empirical distribution of the calibration scores by taking $\hat{q} := \text{the } \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\text{-quantile of } s_i{}_{i=1}^n$, or equivalently, the $\lceil (n+1)(1-\alpha) \rceil$-th order statistic of the calibration scores, where, for any collection of $m \in \mathbb{N}$ scores, we adopt the convention that the $(m+1)$-st order statistic is $+\infty$. The corresponding prediction set is defined as

$$C(x) := \left\{ y \in \mathcal{Y} : s(\hat{f}(x), y) \leq \hat{q} \right\}.$$

Under this construction, sCP admits the following finite-sample coverage guarantee.

Theorem 2.1 (Coverage guarantee of sCP (Angelopoulos & Bates, 2021))**.** *Assume that* $\{(X_i, Y_i)\}_{i=0}^n$ *are exchangeable (See Definition A.1). Then the following inequality holds:*

$$1 - \alpha \;\leq\; \mathbb{P}(Y_0 \in C(X_0))$$

where the probability is taken over the joint distribution of $(X_0, Y_0), \ldots, (X_n, Y_n)$.

Conformal e-value Prediction (eCP). *eCP* follows the same data-splitting procedure as sCP: the data are partitioned into a training set and a calibration set $\mathcal{C} = \{(X_i, Y_i)\}_{i=1}^n$, the former is used to fit $\hat{f}$, and the latter is used to compute the nonconformity scores $\{s_i\}_{i=1}^n$. The difference between eCP and sCP lies in how these scores are aggregated. Instead of computing a quantile, eCP evaluates the test score $s_0 := s\big(\hat{f}(X_0), Y_0\big)$ for a test pair (X_0, Y_0) through the following e-value function:

$$E(X_0, Y_0) := \frac{s_0}{\frac{1}{n+1}\left(s_0 + \sum_{i=1}^n s_i\right)}.$$

Under exchangeability and the assumption that the score function is nonnegative, $E(X_0, Y_0)$ is nonnegative and satisfies $\mathbb{E}[E(X_0, Y_0)] \leq 1$. Using this function, for a user-specified miscoverage level $\alpha \in (0, 1)$, the prediction set is defined as

$$C(X_0) := \left\{ y \in \mathcal{Y} : E(X_0, y) < \frac{1}{\alpha} \right\}. \tag{2}$$

Recalling the definition of exchangeability in Definition A.1, eCP achieves the following guarantee:

Theorem 2.2 (Coverage guarantee of eCP (Vovk, 2025))**.** *Assume that* $\{(X_i, Y_i)\}_{i=0}^n$ *are exchangeable. Then,*

$$\mathbb{P}(Y_0 \in C(X_0)) \geq 1 - \alpha.$$

Tunable Prediction Set Size (TPSS). As before, we denote the input and output spaces by $\mathcal{X}$ and $\mathcal{Y}$, respectively. The TPSS setting characterizes CP methods that simultaneously provide a valid coverage guarantee and explicit control over the size of prediction sets. In particular, a CP method in the TPSS setting is required to satisfy two properties: (i) the size of the prediction set at each test point does not exceed a user-specified budget; and (ii) for the specified budget, it admits a coverage guarantee of the form in equation 1 for some miscoverage level $\alpha \in (0, 1)$.

Formally, let $\mathcal{S} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a user-specified budget function that assigns to each input $x \in \mathcal{X}$ a maximum allowable size for the prediction set. We will assume that the budget function $\mathcal{S}$ is lower bounded by a constant $\zeta > 0$ to avoid degenerate cases. TPSS then requires

$$|C(X_0)| \leq \mathcal{S}(X_0), \tag{3}$$

for the test input $X_0 \in \mathcal{X}$, while admitting a coverage guarantee of the form equation 1 for some $\alpha \in (0, 1)$. Here, $|\cdot|$ denotes a user-specified notion of size on subsets of $\mathcal{Y}$. In particular, when $\mathcal{Y}$ is discrete, $|\cdot|$ can be chosen to correspond to cardinality, while when $\mathcal{Y} \subset \mathbb{R}^d$, it can be chosen to

correspond to the Lebesgue measure. Additionally, instead of enforcing equation 3, we consider a relaxed formulation in which the constraint holds asymptotically as the calibration size n increases:

$$\mathbb{P}\big(|C(X_0)| \leq \mathcal{S}(X_0)\big) \to 1 \quad \text{as } n \to \infty. \tag{4}$$

In Appendix D.4, we show that while TsCP is only theoretically guaranteed to satisfy equation 4, it satisfies equation 3 with high probability empirically, even for small to moderate calibration set sizes n. *More details on motivation and application of the TPSS problem are given in Appendix A.3.*

3 Extension of eCP and sCP to the TPSS problem

In this section, we briefly review how TPSS can be naturally addressed within the eCP framework, following Gauthier et al. (2025b). We then provide a high-level explanation of why this approach does not directly extend to sCP, with further discussion deferred to Appendix C.1.

Why eCP extends to TPSS naturally. The key observation is that the size of the prediction set in the construction of eCP is directly controlled by the choice of α, since α directly controls the inclusion threshold in equation 2. Consequently, the TPSS constraint in equation 3 can be enforced by selecting α so that the resulting prediction set does not exceed the prescribed budget as follows

$$\tilde{\alpha} := \inf\left\{\alpha \in (0,1) : \left|\left\{y : E(X_0, y) < \frac{1}{\alpha}\right\}\right| \leq \mathcal{S}(X_0)\right\}. \tag{5}$$

The choice $\alpha = \tilde{\alpha}$ in equation 5 is therefore data-dependent and, by construction, satisfies the size constraint in equation 3. At a high level, eCP extends naturally to the TPSS setting because its validity can accommodate this data-dependent choice of the miscoverage level α. As shown in Gauthier et al. (2025b), even with data-dependent α, eCP continues to yield a coverage guarantee, up to a first-order Taylor approximation, of the form

$$\mathbb{P}\left(Y_0 \in C_{\tilde{\alpha}}(X_0)\right) \geq 1 - \mathbb{E}[\tilde{\alpha}],$$

where $C_\alpha(X)$ denotes the prediction set construction given in equation 2 for a miscoverage level α.

Why this extension fails in sCP. In the sCP framework, the miscoverage level α determines the quantile threshold and thus controls the size of the prediction set. One may therefore attempt to mimic the above construction by choosing a data-dependent α at each test X_0. In particular, denote

$$C_\alpha(X_0) = \{\, y \in \mathcal{Y} : s(\hat{f}(X_0), y) \leq \hat{q}_\alpha \,\},$$

where $\hat{q}_\alpha$ is the empirical $(1-\alpha)$-quantile of the calibration scores. To enforce the cardinality constraint $|C_\alpha(X_0)| \leq \mathcal{S}(X_0)$, one may define a data-dependent miscoverage level α as

$$\tilde{\alpha}(X_0) := \inf\big\{\alpha \in (0,1) : |C_\alpha(X_0)| \leq \mathcal{S}(X_0)\big\}.$$

To see where this approach fails, recall the sCP argument. Let $s_1, \ldots, s_n$ and s_0 denote the scores of the calibration and test points, respectively. Under exchangeability and no ties, the rank $R := 1 + \sum_{i=1}^{n} \mathbf{1}\{s_i \leq s_0\}$ is uniformly distributed over $\{1, \ldots, n+1\}$. Hence, for any fixed $\alpha \in (0,1)$,

$$\mathbb{P}(R > \lceil (n+1)(1-\alpha) \rceil) \leq 1 - \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \leq \alpha,$$

which implies the standard coverage guarantee $\mathbb{P}(Y_0 \notin C_\alpha(X_0)) \leq \alpha$. However, when α is chosen adaptively as $\tilde{\alpha}(X_0)$, the corresponding event becomes

$$Y_0 \notin C_{\tilde{\alpha}(X_0)}(X_0) \Leftrightarrow R > \lceil (n+1)(1-\tilde{\alpha}(X_0)) \rceil.$$

Unlike the fixed-α case, the threshold $\lceil (n+1)(1-\tilde{\alpha}(X_0)) \rceil$ is now a random quantity that depends on X_0 and, hence, is no longer independent of the rank R. Consequently, $\{R > \lceil (n+1)(1-\tilde{\alpha}(X_0)) \rceil\}$ involves a data-dependent cutoff and cannot be analyzed using the uniformity of R alone. This implies that the extension of eCP to enforce TPSS does not apply in the sCP setting.

4 Tunable-sCP: sCP with Tunable Prediction Set Size

We develop *Tunable-sCP (TsCP)*, a novel method for solving the TPSS problem within the sCP framework. We also provide a theoretical analysis showing that, under mild assumptions, TsCP achieves a coverage guarantee while satisfying the soft size constraint in equation 4. Pseudocode for our proposed method is given in Algorithm 1 in the Appendix.

4.1 METHODOLOGY

Following the procedure of Split Conformal Prediction (sCP), we partition the available data into two disjoint subsets: a training set and a calibration set $\mathcal{C}$ of size $2n$. The model $\hat{f}$ is then trained using only the training data. In contrast to classical sCP, where the entire calibration set is used for a single calibration step, we further partition the calibration set $\mathcal{C}$ into two disjoint subsets of equal size,

$$\mathcal{C}_a = \{(X_1^a, Y_1^a), \ldots, (X_n^a, Y_n^a)\}, \qquad \mathcal{C}_c = \{(X_1^c, Y_1^c), \ldots, (X_n^c, Y_n^c)\}.$$

Similar to sCP, we define a score function $s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that measures the discrepancy between a predicted output and a candidate response. We then compute the following scores:

$$s_i^{(a)} := s\big(\hat{f}(X_i^a), Y_i^a\big), \qquad s_i^{(c)} := s\big(\hat{f}(X_i^c), Y_i^c\big), \qquad i = 1, \ldots, n.$$

For $\alpha \in (0,1)$, let $\hat{q}_a(\alpha)$, $\hat{q}_c(\alpha)$ denote the $(1-\alpha)$-quantiles of $\{s_i^{(a)}\}_{i=1}^n$ and $\{s_i^{(c)}\}_{i=1}^n$, respectively.

As discussed in Section 3, controlling the size of the prediction set requires selecting an appropriate quantile level. However, if this level is chosen adaptively using the same calibration data that is subsequently used for conformal calibration, the exchangeability-based arguments underlying sCP no longer apply. To address this issue, we will use $\mathcal{C}_a$ to determine a data-dependent quantile level, and apply this level to the independent set $\mathcal{C}_c$ for conformal calibration.

For a user-specified slack $\delta \in (0, \zeta]$ and a fixed truncation parameter $\varepsilon \in (0, 1/2)$, we define a data-dependent miscoverage level $\alpha_\delta(X)$ as the smallest value in $[\varepsilon, 1-\varepsilon]$ such that the resulting prediction set, constructed using $\hat{q}_a(\alpha)$, satisfies the size constraint up to a slack of δ. Formally,

$$\alpha_\delta(X) := \inf\left\{\alpha \in [\varepsilon, 1-\varepsilon] : \left|\left\{y \in \mathcal{Y} : s\big(\hat{f}(X), y\big) \le \hat{q}_a(\alpha)\right\}\right| \le \mathcal{S}(X) - \delta\right\}. \tag{6}$$

Here, if the set defining $\alpha_\delta(X_0)$ is empty, we adopt the convention $\alpha_\delta(X_0) = 1 - \varepsilon$. The same convention is used for all LOO and sample-size-indexed adaptive levels defined below. Finally, for a chosen slack $\delta \in (0, \zeta]$, we construct the prediction set for a test input X_0 as

$$C_\delta(X_0) := \left\{y \in \mathcal{Y} : s\big(\hat{f}(X_0), y\big) \le \hat{q}_c(\alpha_\delta(X_0))\right\}. \tag{7}$$

The choice of the slack parameter δ reflects the discrepancy induced by using two independent quantile estimates from $\mathcal{C}_a$ and $\mathcal{C}_c$. In particular, smaller values of δ lead to less conservative prediction sets and potentially improved coverage, while larger values of δ provide a more robust control of the size constraint. The truncation parameter $\varepsilon \in (0, 1/2)$ is introduced to avoid degenerate edge cases in the quantile estimation. In both theory and practice, ε can be chosen *arbitrarily small*, e.g., 10^{-10}, and thus has a negligible impact on the resulting prediction sets. Further discussion and ablation studies on the effect of δ and ε can be found in Appendix D.3 and Appendix C.2, respectively.

4.2 THEORETICAL ANALYSIS

We first state the following assumption, which will be used throughout the theoretical analysis.

Assumption 4.1. *The training, calibration, and test samples are i.i.d. draws from a common distribution, and the nonconformity score $s(\hat{f}(X), Y)$ admits no ties almost surely.*

Discussion of Assumption 4.1. In the classical conformal prediction literature, exchangeability of the data is sufficient to establish coverage guarantees. Nevertheless, the assumption of i.i.d. samples remains standard in most machine learning and deep learning settings (Liu et al., 2021). The assumption that the score $s(\hat{f}(X), Y)$ admits no ties is typically satisfied when the distribution of $s(\hat{f}(X), Y)$ is continuous. In cases where ties do occur, they can be handled using standard tie-breaking techniques from the literature (Jin & Candès, 2023; Vovk et al., 1999; 2005; Bates et al., 2023), such as adding a small random perturbation to the scores. Taken together, these conditions are standard and impose only mild restrictions in practical applications.

We now establish a coverage guarantee for TsCP through the following theorem.

Theorem 4.2 (Coverage guarantee of TsCP)**.** *Let $F(\cdot)$ denote the marginal CDF of the nonconformity score s, and let $\mathrm{BinTail}_n$ be defined as in the table of notation (Appendix). For $\delta \in (0, \zeta]$, define*

$$\Delta_{\delta,n}(\mathcal{C}_a) := \mathbb{E}\left[\mathrm{BinTail}_n\left(\lceil (1-\alpha_\delta(X_0))(n+1)\rceil, F\left(s\left(\widehat{f}(X_0), Y_0\right)\right)\right) - \alpha_\delta(X_0)\middle|\mathcal{C}_a\right],$$

Suppose Assumption 4.1 holds, the set $C_\delta(X_0)$ satisfies

$$\mathbb{P}\left(Y_0 \in C_\delta(X_0)\right) = 1 - \mathbb{E}\left[\alpha_\delta(X_0)\right] - \mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right] = 1 - \mathbb{E}\left[\alpha_\delta(X_0) + \Delta_{\delta,n}(\mathcal{C}_a)\right]. \tag{8}$$

Here, the marginal probability and expectations are taken over $\mathcal{C}_a$, $\mathcal{C}_c$, and (X_0, Y_0).

The proof of Theorem 4.2 is deferred to Appendix B.1. Theorem 4.2 establishes a marginal coverage guarantee analogous to classical CP results. Unlike standard methods with a fixed miscoverage level, the exact coverage expression in equation 8 depends on the data-adaptive quantity $\mathbb{E}[\alpha_\delta(X_0)]$ and the error $\mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right]$. In Section 5, we introduce a leave-one-out estimator for these terms and establish its consistency together with a decaying variance, yielding a computable approximation of the coverage level.

We now show that TsCP satisfies the soft size constraint in equation 4. For each $x \in \mathcal{X}$, we define

$$N_x(t) := \left|\left\{y \in \mathcal{Y} : s\big(\hat{f}(x), y\big) \le t\right\}\right|,$$

which represents the size of the set obtained by thresholding the score function at level t.

At a high level, the key observation behind our method is the fact that the quantile level is selected using $\mathcal{C}_a$, whereas the final threshold is evaluated using the independent set $\mathcal{C}_c$. Consequently, the corresponding empirical quantiles $\hat{q}_a(\alpha)$ and $\hat{q}_c(\alpha)$ at the same level α may differ slightly, which in turn induces a discrepancy between the sizes $N_x\big(\hat{q}_a(\alpha)\big)$ and $N_x\big(\hat{q}_c(\alpha)\big)$.

Under sufficient conditions, these empirical quantiles converge to the same population quantile as the number of calibration samples increases. As a result, the difference between $\hat{q}_a(\alpha)$ and $\hat{q}_c(\alpha)$ becomes small, and it is therefore sufficient to control how the function $N_x(t)$ behaves locally with respect to t. This motivates imposing a regularity condition on $N_x(t)$, stated below.

Assumption 4.3. *There exist constants $L, \eta_0 > 0$ such that, for all $x \in \mathcal{X}$, $t \in \mathbb{R}$, and $0 < \eta < \eta_0$,*

$$N_x(t + \eta) - N_x(t) \le L.$$

We show that, under this assumption, TsCP satisfies the soft size constraint in equation 4.

Theorem 4.4. *Under Assumptions 4.1 and 4.3, suppose that the quantile function $Q(u) := F^{-1}(u)$ is continuous on $(\varepsilon/2, 1 - \varepsilon/2)$, and that the set defining $\alpha_\delta(X_0)$ in equation 6 is nonempty almost surely for sufficiently large n. Then, for any $L \le \delta \le \zeta$, the prediction set satisfies*

$$\mathbb{P}\big(|C_\delta(X_0)| \le \mathcal{S}(X_0)\big) \to 1 \quad \text{as } n \to \infty.$$

The proof of Theorem 4.4 is given in Appendix B.2. While our theoretical results establish only equation 4, we provide additional empirical evidence in Appendix D.4 showing that the method also satisfies equation 3 in practice, even for moderate calibration sample sizes n. Assumption 4.3 is crucial for establishing that our method satisfies equation 4. We discuss this assumption further below.

Discussion of Assumption 4.3. While this assumption may appear restrictive, it is in fact satisfied in a broad range of standard CP settings under commonly used choices of $\mathcal{X}$, $\mathcal{Y}$, and score functions, typically with a small constant L. We now illustrate this through several representative settings.

In the *regression setting* where $\mathcal{Y} = \mathbb{R}$, for standard scores such as the absolute residual $s(\hat{f}(x), y) = |\hat{f}(x) - y|$ and quantile-regression score $s(\hat{f}(x), y) = \max\{q_L(x) - y, y - q_U(x)\}$ (Angelopoulos & Bates, 2021), the induced level sets are intervals whose lengths scale linearly with t. Consequently, Assumption 4.3 holds with a constant L that can be made arbitrarily small. For more complicated scores, such as the normalized residual scores of the form $s(\hat{f}(x), y) = \frac{|\hat{f}(x) - y|}{\hat{\sigma}(x)}$ (Angelopoulos & Bates, 2021), Assumption 4.3 holds whenever $0 < \underline{\sigma} \le \hat{\sigma}(x) \le \overline{\sigma} < \infty$ uniformly over x. Here, the positive lower bound ensures that the normalized score is well-defined, while the finite uniform upper bound controls the size increment of its level sets. Such bounded scale estimates can, for example, be enforced in practice through clipping or other bounded parameterizations.

In higher-dimensional settings such as *multivariate regression*, where $\mathcal{Y} \subset \mathbb{R}^d$ and $s(\hat{f}(x), y) = |\hat{f}(x) - y|_d$, the level sets are Euclidean balls whose volumes vary smoothly with t. As a result, Assumption 4.3 holds for some constant L when $\mathcal{Y}$ is bounded.

In the *classification setting*, where $\mathcal{Y}$ is finite, Assumption 4.3 always holds with the coarse choice $L = |\mathcal{Y}|$. A smaller L requires additional control on the candidate-label scores. For example, if there exists $\eta_0 > 0$ such that, uniformly over $x \in \mathcal{X}$, any interval of length less than η_0 contains at most L candidate-label scores, then Assumption 4.3 holds with this L. In particular, if the candidate-label scores are uniformly separated by at least η_0, then the assumption holds with $L = 1$.

Taken together, these cases show that Assumption 4.3 is satisfied under standard regression settings and under appropriate local multiplicity conditions in classification.

Increase in Calibration Set Size. TsCP requires an additional approximation set $\mathcal{C}_a$ beyond the standard calibration set $\mathcal{C}_c$ to estimate the size-aware quantile level, leading to a modest increase in calibration data; we discuss this cost and its empirical impact further in Appendix D.4.

5 Coverage Estimation via Leave-One-Out

While Theorem 4.2 gives a coverage guarantee for TsCP, equation 8 involves data-dependent terms and is therefore not directly computable. Following Gauthier et al. (2025a), we develop an LOO approximation for these quantities within the sCP framework and, by exploiting its structure, establish consistency and decaying variance guarantees. The procedure is given in Algorithm 2 (Appendix).

5.1 Methodology

At a high level, we will treat each point (X_k^a, Y_k^a) in $\mathcal{C}_a$ as a proxy for a test sample and construct its associated level using the remaining data. In this way, each term is evaluated in a setting that mimics test-time prediction, where the point of interest is not used in the construction of the quantile. Averaging over all such LOO evaluations yields a natural estimator of the target expectations.

Formally, recall the quantile approximation set, $\mathcal{C}_a = \{(X_i^a, Y_i^a)\}_{i=1}^n$. For each $k \in [n]$, define the corresponding LOO sample $\mathcal{C}_a^k := \mathcal{C}_a \setminus \{(X_k^a, Y_k^a)\}$ and let $\left\{ s_i^{a,k} := s(\hat{f}(X_i^a), Y_i^a) : i \in [n],\, i \neq k \right\}$ denote the scores associated with $\mathcal{C}_a^k$. Denote by $\hat{q}_a^k(\alpha)$ the empirical $(1-\alpha)$-quantile of these scores.

For a fixed slack parameter $\delta \in (0, \zeta]$, we then define, for each $k \in [n]$, the *pseudo-level*

$$\alpha_\delta^k(X_k^a) := \inf\left\{ \alpha \in [\varepsilon, 1-\varepsilon] : \left| \left\{ y \in \mathcal{Y} : s\big(\widehat{f}(X_k^a), y\big) \leq \widehat{q}_a^k(\alpha) \right\} \right| \leq \mathcal{S}(X_k^a) - \delta \right\}.$$

The quantities $\{s_i^{a,k}\}_{i \neq k}$, $\widehat{q}_a^k(\alpha)$, and $\alpha_\delta^k(X_k^a)$ collectively form the LOO analogues of the conformity scores, quantile function, and level $\alpha_\delta(X)$, respectively, all constructed by excluding the k-th sample from $\mathcal{C}_a$. Additionally, for each $k \in [n]$, define the associated pseudo-error contribution

$$\Delta_{\delta,n}^k = \mathbf{1}\left\{ s\big(\widehat{f}(X_k^a), Y_k^a\big) > \widehat{q}_c\big(\alpha_\delta^k(X_k^a)\big) \right\} - \alpha_\delta^k(X_k^a).$$

Finally, we define the LOO averages

$$\mathbb{E}\big[\alpha_\delta(X_0)\big] \approx \frac{1}{n}\sum_{k=1}^n \alpha_\delta^k(X_k^a) =: \hat{\alpha}_\delta, \qquad \mathbb{E}\big[\Delta_{\delta,n}(\mathcal{C}_a)\big] \approx \frac{1}{n}\sum_{k=1}^n \Delta_{\delta,n}^k =: \hat{\Delta}_{\delta,n}.$$

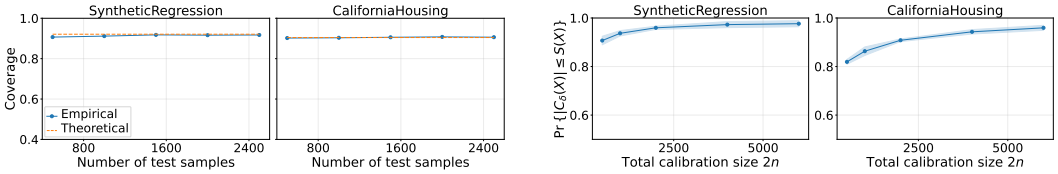
Since each level $\alpha_\delta^k(X_k^a)$ is constructed from the LOO sample $\mathcal{C}_a^k$, while the corresponding error contribution $\Delta_{\delta,n}^k$ is evaluated at the held-out point (X_k^a, Y_k^a) using the independent calibration quantile $\widehat{q}_c$, these estimators closely mimic the test-time construction of the prediction set in equation 7 and, therefore, provide natural and principled approximations to the two target expectations.

5.2 Theoretical Analysis

We now establish the theoretical properties of the proposed LOO estimators $\hat{\alpha}_\delta$ and $\hat{\Delta}_{\delta,n}$. To this end, we introduce assumption 5.1. A detailed discussion and comparison with assumptions in the existing literature is provided in Appendix C.3.

Assumption 5.1. *Let μ denote the marginal distribution of $s(\hat{f}(X), Y)$, and let μ_x denote its conditional distribution given $X = x$. Assume that there exists a constant $K < \infty$ such that, for P_X-almost every $x \in \mathcal{X}$, $\mu_x \ll \mu$ and $d\mu_x/d\mu \leq K$ μ-a.e.*

Under Assumption 5.1, we first show that $\hat{\alpha}_\delta$ and $\hat{\Delta}_{\delta,n}$ provide a consistent approximation to the target expectations with convergence rate $\mathcal{O}_P(n^{-1/2})$.



(a) Empirical validation of Theorem 4.2.

(b) Empirical validation of Theorem 4.4.

Figure 1: Empirical validation of the theoretical results in Section 4 on the **SyntheticRegression** and **California Housing** datasets. See Subsection 6.1 for a detailed description of the experimental setup.

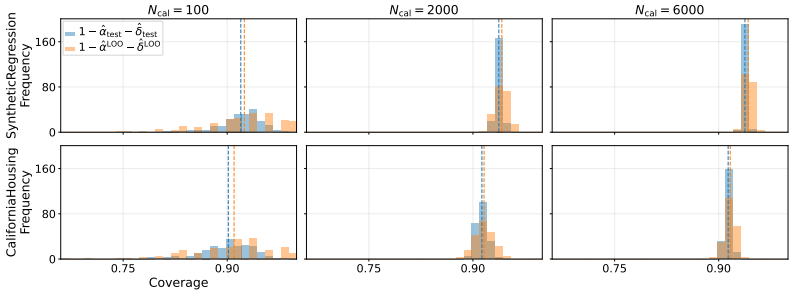


Figure 2: Empirical behavior of the LOO estimator corresponding to the convergence and variance properties in Theorems 5.2 and 5.3 on the **SyntheticRegression** and **California Housing** datasets.

Theorem 5.2 (Convergence Rates)**.** *Fix $0 < \delta < \zeta$. Under Assumptions 4.1 and 5.1, we have*

$$\left|\hat{\alpha}_\delta - \mathbb{E}[\alpha_\delta(X_0)]\right| = \mathcal{O}_P\left(n^{-1/2}\right), \quad \left|\hat{\Delta}_{\delta,n} - \mathbb{E}[\Delta_{\delta,n}(\mathcal{C}_a)]\right| = \mathcal{O}_P\left(n^{-1/2}\right).$$

Consequently, $\left|\hat{\alpha}_\delta + \hat{\Delta}_{\delta,n} - \mathbb{E}[\alpha_\delta(X_0) + \Delta_{\delta,n}(\mathcal{C}_a)]\right| = \mathcal{O}_P\left(n^{-1/2}\right)$.

The proof of Theorem 5.2 is deferred to Appendix B.3. We next characterize the variability of the estimators, showing that its variance scales as $\mathcal{O}(n^{-1})$, through the following theorem.

Theorem 5.3 (Variance)**.** *Fix $0 < \delta < \zeta$. Under Assumptions 4.1 and 5.1, we have*

$$\mathrm{Var}(\hat{\alpha}_\delta) = \mathcal{O}\left(n^{-1}\right), \quad \mathrm{Var}(\hat{\Delta}_{\delta,n}) = \mathcal{O}\left(n^{-1}\right).$$

Consequently, $\mathrm{Var}\left(\hat{\alpha}_\delta + \hat{\Delta}_{\delta,n}\right) = \mathcal{O}\left(n^{-1}\right)$.

The proof of Theorem 5.3 is deferred to Appendix B.4.

6 Experimental Results

We empirically evaluate TsCP by first validating its theoretical guarantees and then comparing it with existing TPSS methods. In particular, we show that TsCP achieves **(A)** smaller prediction sets at comparable empirical coverage and **(B)** substantially tighter LOO-based empirical coverage bounds than existing eCP approaches under the same calibration budget.

Overall Setup. We evaluate our method in both regression and classification settings, using both synthetic and real-world datasets. For regression, we use Random Forest (Rigatti, 2017) and evaluate on a synthetic linear-Gaussian dataset, California Housing (Pace & Barry, 1997), and Superconductivity (Hamidieh, 2018). For classification, we use Logistic Regression (LaValley, 2008) and consider a synthetic multi-class logistic dataset with informative and correlated redundant features, along with MNIST (LeCun et al., 2002) and Covertype (Blackard & Dean, 1999). Due to space constraints, we present only the regression results in the main text, while the classification results are deferred to Appendix D.2. We will also only consider a constant budget function $\mathcal{S}(X) = c$, for some $c > 0$, in the main text, and investigate more general choices of the budget function in Appendix D.5.

Ablation Study. We conduct ablation studies on the *slack parameter* δ, *type of size constraint*, *base model* $\hat{f}$, and *inference time*. Results are provided in Appendices D.3, D.4, D.6, and D.7, respectively.

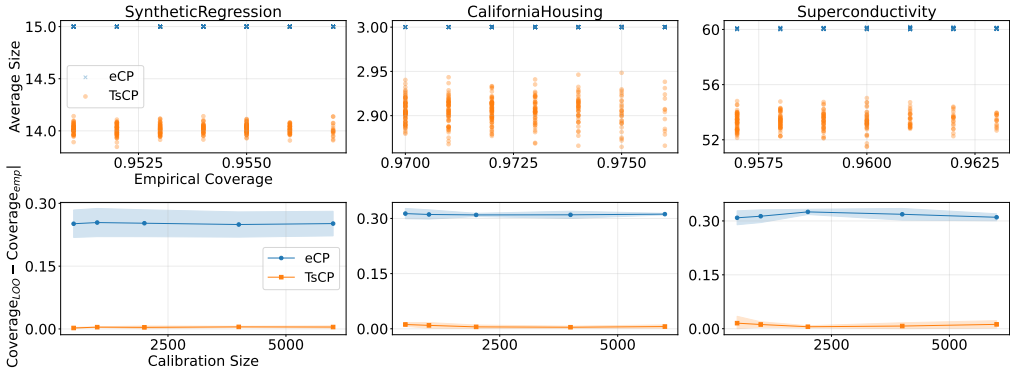


Figure 3: Comparison of TsCP with baselines. The first and second rows correspond to Task **(a)** and Task **(b)**, respectively, while the columns correspond to **SyntheticRegression**, **California Housing**, and **Superconductivity**, respectively.

6.1 Validation of Theoretical Guarantees

Setup. To validate our theoretical guarantees, we fix the budget c and slack δ. For Theorem 4.2, we fix the calibration size while varying the number of test points, and compare the empirical coverage with the exact coverage quantity in equation 8. For Theorem 4.4, we vary the calibration size and evaluate the probability that the prediction-set size satisfies the prescribed budget. To evaluate Theorems 5.2 and 5.3, following (Gauthier et al., 2025a), we repeat the experiment R times for each calibration size and compare the distributions of $1-\mathbb{E}[\alpha_\delta(X_0)+\Delta_{\delta,n}(\mathcal{C}_a)]$ and $1-\widehat{\alpha}_\delta^{\mathrm{LOO}}-\widehat{\Delta}_{\delta,n}^{\mathrm{LOO}}$, tracking the alignment of their means and the reduction in estimator variability as the calibration size increases. Further details are provided in Appendix D.1.1.

Results. Empirical validations of Theorems 4.2 and 4.4 are shown in Figures 1 and 4, while Figures 2 and 5 illustrate empirical behavior consistent with Theorems 5.2 and 5.3. Across datasets, the empirical coverage achieved by TsCP closely matches the theoretical coverage quantity, while the prediction-set size satisfies the prescribed budget c with high probability. Moreover, the empirical mean of $1-\widehat{\alpha}_\delta^{\mathrm{LOO}}-\widehat{\Delta}_{\delta,n}^{\mathrm{LOO}}$ closely tracks that of $1-\mathbb{E}[\alpha_\delta(X_0)+\Delta_{\delta,n}(\mathcal{C}_a)]$, supporting consistency. As the calibration size increases, the distribution of the LOO-based estimate also becomes increasingly concentrated, consistent with the stability guarantee.

6.2 Comparison with Existing Methods

Setup. We compare TsCP with the eCP approach of (Gauthier et al., 2025a) on two tasks. **(a)** We fix the budget level c and evaluate both methods across multiple random seeds, reporting average test coverage and prediction set size to assess the coverage–efficiency trade-off. **(b)** We vary the calibration size n and compare the LOO coverage estimate of TsCP and the LOO-estimated bound of eCP with their corresponding empirical test coverage. Further implementation details are provided in Appendix D.1.2.

Results. The results for Tasks **(a)** and **(b)** are shown in the first and second rows of Figure 3, respectively. For Task **(a)**, TsCP achieves smaller prediction sets at comparable empirical coverage across all three datasets. For Task **(b)**, its LOO-based bound remains substantially closer to empirical coverage across all datasets and calibration sizes, while eCP exhibits a larger persistent gap as n increases. These results suggest that, by retaining the structure of sCP, TsCP provides a more accurate and less conservative practical coverage estimate under LOO estimation.

7 Conclusion

We show that extending sCP to TPSS is non-trivial and propose a data-splitting approach that constructs a size-aware quantile level on an auxiliary set and applies it to an independent calibration set, preserving validity while controlling prediction-set size. We also develop a consistent and stable LOO estimator for the coverage quantity. Experiments on synthetic and real-world regression and classification tasks exhibit behavior consistent with our theoretical findings and show strong performance against existing methods. A limitation is the i.i.d. assumption, which may fail under covariate shift; extending TsCP to this setting is an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on our experiments and implementing methods. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology, and interpreting results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for modifying scientific figures, suggesting experimental parameters, editing software code, editing our research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. In particular, we discussed and evaluated within our group the AI-generated suggestions for experimental design. We also carefully assessed AI suggestions for method implementation and wrote the final code ourselves. In addition, we reviewed all AI-generated or AI-edited figures and manuscript revisions, and carefully read and verified the literature suggested by AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Given the nature of the work, we do not foresee any negative societal and ethical impacts of our work.

REPRODUCIBILITY STATEMENT

Source codes for our experiments are provided in the supplementary materials of the paper. The details of our experimental settings are given in Section 6 and the Appendix. All datasets that we used in the paper are published, and they are easy to access in the Internet.

REFERENCES

Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In *Conformal and Probabilistic Prediction with Applications*, pp. 36–55. PMLR, 2023.

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

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.

Real Carbonneau, Kevin Laframboise, and Rustam Vahidov. Application of machine learning techniques for supply chain demand forecasting. *European journal of operational research*, 184 (3):1140–1154, 2008.

Danhui Chen, Ziquan Liu, Chuxi Yang, Dan Wang, Yan Yan, Yi Xu, and Xiangyang Ji. Conformalsam: Unlocking the potential of foundational segmentation models in semi-supervised semantic segmentation with conformal prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24045–24055, 2025.

John Cherian, Isaac Gibbs, and Emmanuel Candes. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

Matt Y Cheung, Ashok Veeraraghavan, and Guha Balakrishnan. Compass: Robust feature conformal prediction for medical segmentation metrics. *arXiv preprint arXiv:2509.22240*, 2025.

Marius Gade, Kevin Mekhaphan Nguyen, Sol Gedde, and Alvaro Fernandez-Quilez. Impact of uncertainty quantification through conformal prediction on volume assessment from deep learning-based mri prostate segmentation. *Insights into Imaging*, 15(1):286, 2024.

Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. *arXiv preprint arXiv:1301.7375*, 2013.

Etienne Gauthier, Francis Bach, and Michael I Jordan. Backward conformal prediction. *arXiv preprint arXiv:2505.13732*, 2025a.

Etienne Gauthier, Francis Bach, and Michael I Jordan. E-values expand the scope of conformal prediction. *arXiv preprint arXiv:2503.13050*, 2025b.

Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. In *2020 Information theory and applications workshop (ITA)*, pp. 1–54. IEEE, 2020.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Dongjian Hu, Junxi Wu, Shu-Tao Xia, and Changliang Zou. Distribution-informed online conformal prediction. *arXiv preprint arXiv:2512.07770*, 2025a.

Peiyan Hu, Xiaowei Qian, Wenhao Deng, Rui Wang, Haodong Feng, Ruiqi Feng, Tao Zhang, Long Wei, Yue Wang, Zhi-Ming Ma, et al. From uncertain to safe: Conformal adaptation of diffusion models for safe pde control. In *Forty-second International Conference on Machine Learning*, 2025b.

Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine learning*, 97(1):155–176, 2014.

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. Conformal prediction masks: Visualizing uncertainty in medical imaging. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pp. 163–176. Springer, 2023.

Michael P LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

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

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Charles Lu and Jayasheree Kalpathy-Cramer. Distribution-free federated learning with conformal predictions. *arXiv preprint arXiv:2110.07661*, 2021.

Charles Lu, Anastasios N Angelopoulos, and Stuart Pomerantz. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *International conference on medical image computing and computer-assisted intervention*, pp. 545–554. Springer, 2022a.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 12008–12016, 2022b.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pp. 22942–22964. PMLR, 2023.

Luca Mossina and Corentin Friedrich. Conformal prediction for image segmentation using morphological prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 78–88. Springer, 2025.

Sara Narteni, Alberto Carlevaro, Fabrizio Dabbene, Marco Muselli, and Maurizio Mongelli. A novel score function for conformal prediction in rule-based binary classification. *Pattern Recognition*, pp. 112219, 2025.

Ilia Nouretdinov, Thomas Melluish, and Volodya Vovk. Ridge regression confidence machine. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 385–392, 2001.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, 1997.

Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pp. 388–395. IEEE, 2007.

Yash Patel, Eduardo Ochoa Rivera, and Ambuj Tewari. Conformal prediction for ensembles: Improving efficiency via score-based aggregation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Steven J Rigatti. Random forest. *Journal of insurance medicine*, 47(1):31–39, 2017.

Glenn Shafer. The language of betting as a strategy for statistical and scientific communication. *arXiv preprint arXiv:1903.06991*, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Jiankai Sun, Yiqi Jiang, Jianing Qiu, Parth Nobel, Mykel J Kochenderfer, and Mac Schwager. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. *Advances in Neural Information Processing Systems*, 36:80324–80337, 2023.

Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *International Conference on Machine Learning*, pp. 33940–33960. PMLR, 2023.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *arXiv preprint arXiv:2501.18991*, 2025.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Harit Vishwakarma, Alan Mishler, Thomas Cook, Niccolo Dalmasso, Natraj Raman, and Sumitra Ganesh. Prune'n predict: Optimizing llm decision-making with conformal prediction. In *Forty-second International Conference on Machine Learning*, 2025.

Vladimir Vovk. On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pp. 187–196. IEEE, 2002.

Vladimir Vovk. Conformal e-prediction. *Pattern Recognition*, 166:111674, 2025.

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

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 444–453, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Chunjiang Wang, Yunjie Gu, Pengbo Liu, Shang Zhao, and S Kevin Zhou. E-dm: Evaluating diffusion model by conformal prediction. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. Moving to a world beyond "p< 0.05", 2019.

Håkan Wieslander, Philip J Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan Karlsson, Ola Spjuth, and Carolina Wählby. Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. *IEEE journal of biomedical and health informatics*, 25(2):371–380, 2020.

Soroush H Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In *International Conference on Machine Learning*, pp. 12292–12318. PMLR, 2023.

Jin Zhang, Ulf Norinder, and Fredrik Svensson. Deep learning-based conformal prediction of toxicity. *Journal of chemical information and modeling*, 61(6):2648–2657, 2021.

TABLE OF NOTATION

$\mathbb{R}^d$	d-dimensional Euclidean space
$\mathbb{R}_{\geq 0}$	Set of non-negative real numbers
$\|\cdot\|_d$	Euclidean norm in $\mathbb{R}^d$
$\inf\{\cdot\}$	Infimum (greatest lower bound)
$\mathbb{E}[\cdot]$	Expectation of a Random Variable
$\text{Var}(\cdot)$	Variance of a Random Variable
$\mathcal{O}(\cdot)$	Big-O notation
$\mathcal{O}_P(\cdot)$	Big-O notation in probability
$\mathcal{P}(X)$	The collection of all subsets of X for some set X
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space
$\hat{f}$	Base prediction model
α	Miscoverage level
$C(\cdot)$	Set-valued predictor
$s(\cdot)$	Score function
$\mathcal{S}(\cdot)$	Budget function
ζ	Lower bound of the Budget function
$\|\cdot\|$	User-specified notion of size of a set
$\mathcal{C}_a$	Approximation set
$\mathcal{C}_c$	Calibration set
$\mathcal{C}_a^k$	$\mathcal{C}_a \setminus \{(X_k^a, Y_k^a)\}$, for $k \in \{1, \ldots, n\}$
n	Size of the Approximation set and the Calibration set
$s_i^{(a)}$	Score of the i-th sample in the Approximation set
$s_i^{(c)}$	Score of the i-th sample in the Calibration set
$s_i^{a,k}$	Score of the i-th sample in $\mathcal{C}_a^k$
$\hat{q}_a(\alpha)$	$\lceil (n+1)(1-\alpha) \rceil$-th order statistic of $\{s_i^{(a)}\}_{i=1}^n$
$\hat{q}_c(\alpha)$	$\lceil (n+1)(1-\alpha) \rceil$-th order statistic of $\{s_i^{(c)}\}_{i=1}^n$
$\hat{q}_a^k(\alpha)$	$\lceil n(1-\alpha) \rceil$-th order statistic of $\{s_i^{a,k}\}_{i \in \{1,\ldots,n\}}^{i \neq k}$
δ	Slack Parameter
ε	Truncation Parameter
$N_x(t)$	Size of $\left\{ y \in \mathcal{Y} : s\big(\hat{f}(x), y\big) \leq t \right\}$, for some $x \in \mathcal{X}$
L	Bounding constant given in Assumption 4.3
$\hat{\alpha}_\delta$	Leave-one-out Estimator
$\text{BinTail}_n(k, p)$	Binomial upper-tail probability $\mathbb{P}(Z \geq k)$ for $Z \sim \text{Bin}(n, p)$.
$\alpha_\delta(X)$	Data-dependent miscoverage level used to satisfy the prediction-set size constraint.
$\Delta_{\delta,n}(C_a)$	Conditional coverage correction term
$\widehat{\Delta}_{\delta,n}$	LOO estimator of $\mathbb{E}[\Delta_{\delta,n}(C_a)]$.
F	Marginal CDF of the nonconformity score.
μ, μ_x	Marginal and conditional distributions of the nonconformity score, respectively.

Appendix of "Split Conformal Prediction with Tunable Prediction Set Size"

Table of Contents

A Background **16**

A.1 Split Conformal Prediction 16

A.2 Conformal e-value Prediction 21

A.3 The Tunable Prediction Set Size problem 24

B Theoretical Proofs **26**

B.1 Proof of the Coverage Guarantee - Theorem 4.2 26

B.2 Proof of Soft Size Control- Theorem 4.4 27

B.3 Proof of LOO Estimator Convergence Rate - Theorem 5.2 28

B.4 Proof of LOO Estimator Stability - Theorem 5.3 37

C Further Discussion and Analysis **39**

C.1 Comparison between eCP and sCP for the Tunable Prediction Set Size problem . . 39

C.1.1 Why eCP admits a more direct extension to TPSS than sCP 39

C.1.2 Advantages of TsCP over eCP-Based TPSS Methods 40

C.2 Effect of the Truncation Parameter ε 41

C.2.1 Why ε is necessary 42

C.2.2 Why ε does not weaken the proposed method 42

C.3 Discussion of Assumption 5.1 42

D Experimental Details **46**

D.1 Experimental Setup for Regression Setting 46

D.1.1 Validation of Theoretical Guarantees 47

D.1.2 Comparison with Existing Methods 49

D.2 Experiments on the Classification Setting 50

D.2.1 Validation of Theoretical Guarantees 50

D.2.2 Comparison with Existing Methods 52

D.3 Ablation on the slack parameter δ 53

D.4 Ablation on Size Constraint Variants and Calibration Set Size 55

D.5 Ablation on Budget Function $\mathcal{S}$ 57

D.6 Ablation on the model $\hat{f}$ 58

D.7 Ablation on Inference Time 59

D.7.1 Computational Overhead of TsCP compared to sCP 59

D.7.2 Empirical Comparison with standard Split Conformal Prediction 59

D.8 Supplementary Statements 60

Algorithm 1 Our proposed method for sCP under Tunable Prediction Set Size setting

Require: Training set $\mathcal{D}_{\text{train}}$, Calibration set $\mathcal{C}$, Test input X_0, Size budget function $\mathcal{S}(\cdot)$, slack $\delta \in (0, \zeta]$, truncation parameter $\varepsilon \in (0, 1/2)$, Nonconformity score $s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

1: Split $\mathcal{C}$ into two disjoint subsets of equal size:

$$\mathcal{C}_a = \{(X_1^a, Y_1^a), \ldots, (X_n^a, Y_n^a)\}, \qquad \mathcal{C}_c = \{(X_1^c, Y_1^c), \ldots, (X_n^c, Y_n^c)\}.$$

2: Train the prediction model $\hat{f}$ on the training data.

3: **for** $i = 1, \ldots, n$ **do**

4: Compute

$$s_i^{(a)} \leftarrow s\big(\hat{f}(X_i^a), Y_i^a\big), \qquad s_i^{(c)} \leftarrow s\big(\hat{f}(X_i^c), Y_i^c\big).$$

5: **end for**

6: Define $\hat{q}_a(\alpha) \leftarrow$ the empirical $(1-\alpha)$-quantiles of $\{s_i^{(a)}\}_{i=1}^n$.

7: Define $\hat{q}_c(\alpha) \leftarrow$ the empirical $(1-\alpha)$-quantiles of $\{s_i^{(c)}\}_{i=1}^n$.

8: Compute the data-dependent level

$$\alpha_\delta(X_0) \leftarrow \inf\Big\{\alpha \in [\varepsilon, 1-\varepsilon] : \big|\{y \in \mathcal{Y} : s(\hat{f}(X_0), y) \le \hat{q}_a(\alpha)\}\big| \le \mathcal{S}(X_0) - \delta\Big\}.$$

if the admissible set is empty, set $\alpha_\delta(X_0) = 1 - \varepsilon$.

9: Output the prediction set

$$C_\delta(X_0) \leftarrow \{y \in \mathcal{Y} : s(\hat{f}(X_0), y) \le \hat{q}_c(\alpha_\delta(X_0))\}.$$

A Background

A.1 Split Conformal Prediction

Motivation. Recent advances in machine learning have enabled highly accurate predictions across a diverse range of domains, including structured data analysis, large-scale recommendation systems, and scientific modeling. Despite these successes, the outputs of such models remain inherently uncertain: they typically provide single-point predictions without any intrinsic guarantee of correctness. In many practical scenarios, this lack of reliability poses a significant concern. For instance, in decision-support systems for healthcare, automated trading strategies, or safety monitoring pipelines, an erroneous prediction presented with high confidence may lead to severe downstream consequences.

A key difficulty lies in the fact that standard confidence measures produced by modern models are not designed to reflect true predictive uncertainty. These scores are often sensitive to model misspecification and can behave unpredictably when the input deviates from the training distribution. As a result, practitioners face a fundamental challenge: how to augment arbitrary predictive models with uncertainty quantification that remains reliable even in finite samples and under minimal assumptions. Conformal Prediction addresses this issue by constructing set-valued predictions that are guaranteed to contain the true outcome with a prescribed probability level. Rather than relying on model-specific uncertainty estimates, conformal methods leverage the relative conformity of new observations with respect to a reference dataset to calibrate prediction sets. This approach yields finite-sample validity without requiring a parametric model or distributional specification. Its only assumption is the exchangeability of the calibration and test data, which we recall below.

Definition A.1 (Exchangeability)**.** Random variables $X_1, \ldots, X_n$ are called exchangeable if their joint distribution is invariant under any permutation of the indices. That is, for every permutation π of $\{1, \ldots, n\}$,

$$(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)}). \tag{9}$$

Consequently, Conformal Prediction provides a flexible and model-agnostic framework for uncertainty quantification, making it particularly appealing for modern machine learning systems where robustness and reliability are essential.

Algorithm 2 LOO Estimators for $\mathbb{E}[\alpha_\delta(X_0)]$ and $\mathbb{E}[\Delta_{\delta,n}(\mathcal{C}_a)]$

Require: Approximation subset $\mathcal{C}_a = \{(X_1^a, Y_1^a), \ldots, (X_n^a, Y_n^a)\}$, calibration subset $\mathcal{C}_c = \{(X_1^c, Y_1^c), \ldots, (X_n^c, Y_n^c)\}$, size budget function $\mathcal{S}(\cdot)$, slack $\delta \in (0, \zeta]$, truncation parameter $\varepsilon \in (0, 1/2)$, nonconformity score s.

Compute calibration scores:

$$s_i^{(c)} \leftarrow s(\hat{f}(X_i^c), Y_i^c), \qquad i = 1, \ldots, n.$$

Compute empirical $(1-\alpha)$-quantiles $\hat{q}_c(\alpha)$ from $\{s_i^{(c)}\}_{i=1}^n$.

for $k = 1, \ldots, n$ **do**

Define leave-one-out set:

$$\mathcal{C}_a^k \leftarrow \mathcal{C}_a \setminus \{(X_k^a, Y_k^a)\}.$$

Compute scores:

$$s_i^{a,k} \leftarrow s(\hat{f}(X_i^a), Y_i^a), \qquad i \neq k.$$

Compute empirical $(1-\alpha)$-quantiles $\hat{q}_a^k(\alpha)$ from $\{s_i^{a,k}\}_{i \neq k}$.

Compute pseudo-level:

$$\alpha_\delta^k(X_k^a) \leftarrow \inf\Big\{\alpha \in [\varepsilon, 1-\varepsilon] : \big|\{y \in \mathcal{Y} : s(\hat{f}(X_k^a), y) \leq \hat{q}_a^k(\alpha)\}\big| \leq \mathcal{S}(X_k^a) - \delta\Big\}.$$

if the admissible set is empty, set $\alpha_\delta^k(X_k^a) = 1 - \varepsilon$.

Compute pseudo-error contribution:

$$\Delta_{\delta,n}^k \leftarrow \mathbf{1}\Big\{s(\hat{f}(X_k^a), Y_k^a) > \hat{q}_c\big(\alpha_\delta^k(X_k^a)\big)\Big\} - \alpha_\delta^k(X_k^a).$$

end for

Output estimators:

$$\hat{\alpha}_\delta \leftarrow \frac{1}{n}\sum_{k=1}^n \alpha_\delta^k(X_k^a),$$

$$\widehat{\Delta}_{\delta,n} \leftarrow \frac{1}{n}\sum_{k=1}^n \Delta_{\delta,n}^k.$$

Full (Transductive) Conformal Prediction (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021). Conformal Prediction begins with the transductive (full) formulation, which provides the conceptual foundation for many of its modern variants. Consider an exchangeable collection of $n+1$ observations

$$\{(X_1, Y_1), \ldots, (X_n, Y_n), (X_0, Y_0)\}, \tag{10}$$

where the learner observes the first n labeled samples together with the test input X_0, and seeks to infer the unknown response Y_0.

The defining feature of this framework is that candidate responses are evaluated individually. For each $y \in \mathcal{Y}$, one forms an augmented dataset by appending (X_0, y) to the observed sample, and trains a predictive rule $\hat{f}^y$ on

$$\{(X_1, Y_1), \ldots, (X_n, Y_n), (X_0, y)\}. \tag{11}$$

To ensure validity, the underlying learning procedure is required to be invariant to permutations of the training data.

Given a score function $s(\cdot, \cdot; \cdot)$ measuring the degree of agreement between inputs, outputs, and the fitted model, define the augmented set of scores

$$s_i^y := s\big(X_i, Y_i; \hat{f}^y\big), \quad i = 1, \ldots, n, \tag{12}$$

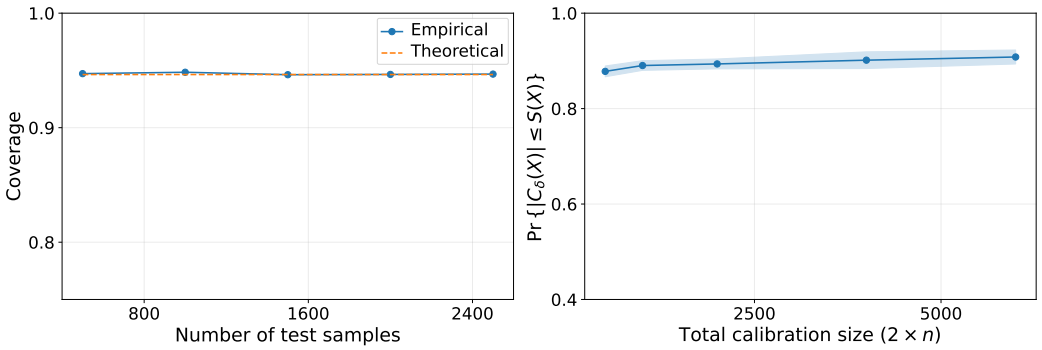


Figure 4: Empirical validation of Theorem 4.2 (Left) and Theorem 4.4 (Right) on the **Superconductivity** dataset. See Subsection 6.1 for a detailed description of the experimental setup.

and define

$$s_0^y := s\big(X_0, y; \hat{f}^y\big). \tag{13}$$

Let

$$k := \lceil (n+1)(1-\alpha) \rceil, \tag{14}$$

and denote by $\widehat{q}^y$ the k-th smallest value among $\{s_0^y, s_1^y, \ldots, s_n^y\}$. The conformal prediction set is then defined as

$$\mathcal{C}(X_0) = \Big\{ y \in \mathcal{Y} : s_0^y \leq \widehat{q}^y \Big\}. \tag{15}$$

This construction retains precisely those candidate outputs whose associated score is not unusually large relative to the augmented sample. In other words, (X_0, y) is accepted whenever it behaves comparably to the observed data under the model fitted with that same candidate.

Theorem A.2 (Full conformal coverage guarantee (Angelopoulos & Bates, 2021))**.** *Suppose that $\{(X_i, Y_i)\}_{i=0}^n$ are exchangeable, and that the learning algorithm used to construct $\hat{f}^y$ is permutation-invariant. Then the resulting prediction set satisfies*

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0)\right) \geq 1 - \alpha. \tag{16}$$

The guarantee follows from the symmetry of the augmented scores, which ensures that the rank of s_0^y among $\{s_i^y\}_{i=0}^n$ is uniformly distributed.

Despite its conceptual simplicity and strong finite-sample validity, this procedure is often impractical for large-scale applications, as it requires retraining the model separately for each candidate y. This computational burden motivates alternative constructions, such as split conformal methods, which approximate the same guarantee while significantly reducing the cost of implementation.

Split (Inductive) Conformal Prediction. While full conformal prediction provides a conceptually simple construction, its computational cost can be prohibitive in practice. Split Conformal Prediction offers a scalable alternative by decoupling model fitting from uncertainty quantification. Instead of retraining the model for each candidate output, the predictive rule is learned once and then evaluated using a separate calibration sample. Let $\{(X_i, Y_i)\}_{i=1}^n$ be exchangeable observations from an unknown distribution on $\mathcal{X} \times \mathcal{Y}$. Given a target miscoverage level $\alpha \in (0, 1)$, the objective is to construct a set-valued predictor $\mathcal{C} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ such that

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0)\right) \geq 1 - \alpha, \tag{17}$$

where (X_0, Y_0) is an independent test point drawn from the same distribution.

The split conformal procedure begins by partitioning the data into two disjoint subsets: a training set and a calibration set. A predictive model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ is trained exclusively on the training portion and is thereafter kept fixed. The calibration set is used solely to quantify the typical variability of the prediction errors.

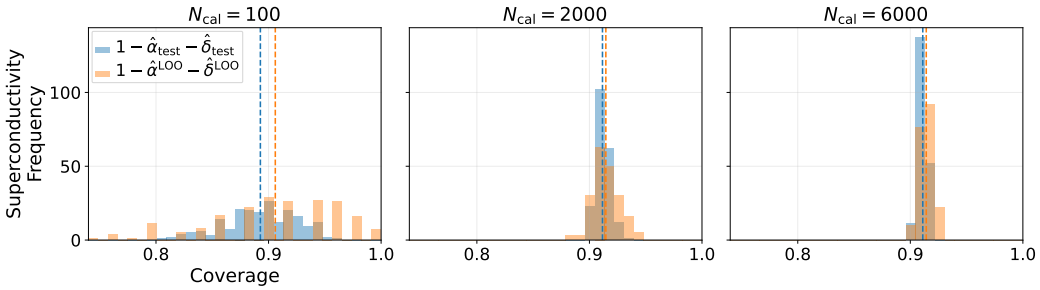


Figure 5: Empirical behavior of the LOO estimator corresponding to the convergence and variance properties in Theorems 5.2 and 5.3 on the **Superconductivity** dataset. See Subsection 6.1 for a detailed description of the experimental setup.

To this end, let $\mathcal{I}_{\text{cal}} \subset \{1, \ldots, n\}$ denote the indices of the calibration points, and define the calibration scores via a user-specified nonconformity function

$$s_i := s\big(\hat{f}(X_i), Y_i\big), \quad i \in \mathcal{I}_{\text{cal}}. \tag{18}$$

The function s measures the discrepancy between the predicted and observed responses, and its choice depends on the underlying task.

An empirical threshold is obtained from the upper quantiles of the calibration scores. Specifically, let $n_{\text{cal}} := |\mathcal{I}_{\text{cal}}|$, and define

$$\hat{q} := \text{the } \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil \text{-th smallest value among } \{s_i\}_{i \in \mathcal{I}_{\text{cal}}}. \tag{19}$$

The resulting prediction set for a new input x is given by

$$\mathcal{C}(x) = \big\{ y \in \mathcal{Y} : s\big(\hat{f}(x), y\big) \leq \hat{q} \big\}. \tag{20}$$

Theorem A.3 (Split conformal coverage guarantee)**.** *Conditional on the training set used to construct $\hat{f}$, if the calibration observations $(X_i, Y_i)_{i \in \mathcal{I}_{\text{cal}}}$ and the test point (X_0, Y_0) are exchangeable, then the prediction set $\mathcal{C}$ defined above satisfies*

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0)\right) \geq 1 - \alpha, \tag{21}$$

for any underlying data distribution and any choice of score function s.

The validity of this procedure follows from the exchangeability between the test point (X_0, Y_0) and the calibration samples, which ensures that the rank of the test score among the calibration scores is uniformly distributed. Unlike the full conformal construction, this argument does not require retraining the model for each candidate output.

Conceptual interpretation. Split Conformal Prediction can be viewed as a constrained version of the full conformal procedure, in which the predictive model is trained once and subsequently held fixed. Conformity is then evaluated using a separate calibration sample, rather than recomputing scores under a model refit on each augmented dataset. This modification breaks the full symmetry across all $n + 1$ observations, but preserves the exchangeability between the test point and the calibration data, which is sufficient to ensure marginal coverage.

Computational considerations. This reformulation yields a significant reduction in computational complexity. Full conformal prediction requires retraining the model for each candidate $y \in \mathcal{Y}$, which quickly becomes infeasible when the output space is large, continuous, or structured. In contrast, Split Conformal Prediction performs model fitting only once, and thus scales well to modern settings involving high-dimensional outputs or complex architectures. While this simplification may reduce statistical efficiency relative to the full method, it retains the distribution-free coverage guarantee and is therefore the standard choice in practice. **Extensions of Conformal Prediction.** Recent developments have pushed Conformal Prediction beyond its classical setting, enabling valid uncertainty quantification in a variety of more challenging and structured learning scenarios. Many of these works aim to refine the standard marginal coverage guarantee to better align with application-specific requirements or to accommodate departures from ideal assumptions. TsCP,

Tunable Prediction Set Size (TPSS), belongs to this broader line of research. Before introducing our approach, we briefly review several representative extensions that illustrate how the conformal framework can be adapted to more complex settings.

Group-wise validity (Angelopoulos & Bates, 2021). In many applications, it is not sufficient to ensure coverage only on average over the entire population; instead, reliability may be required within predefined subgroups. This leads to group-balanced conformal methods, where the goal is to guarantee

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0) \mid X_0 \in g\right) \geq 1-\alpha, \quad \forall g \in \{1, \ldots, G\}. \tag{22}$$

To achieve this, calibration scores are partitioned according to group membership, and separate quantile thresholds are computed within each group. At inference time, the prediction set is constructed using the threshold corresponding to the group of the test input. This ensures coverage guarantees hold uniformly across all groups under exchangeability.

Label-conditional guarantees (Angelopoulos & Bates, 2021). In classification tasks, one may instead seek guarantees that are conditioned on the true class label. Formally, for $\mathcal{Y} = \{1, \ldots, K\}$, the objective becomes

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0) \mid Y_0 = k\right) \geq 1-\alpha, \quad \forall k \in \mathcal{Y}. \tag{23}$$

Since the true label is unknown at test time, these methods construct class-specific thresholds using calibration data stratified by class. A candidate label is included in the prediction set if its associated score satisfies the corresponding class-level threshold. This yields valid class-conditional coverage while preserving the conformal framework.

Risk Control (Angelopoulos et al., 2022). More recent work generalizes conformal guarantees to control broader notions of predictive error. Instead of bounding the probability of exclusion, conformal risk control seeks to ensure

$$\mathbb{E}\left[\ell\left(\mathcal{C}(X_0), Y_0\right)\right] \leq \alpha, \tag{24}$$

for a user-specified loss function ℓ. This formulation allows practitioners to target task-dependent criteria, such as false negative rates or asymmetric penalties, rather than simple miscoverage. The approach typically constructs a family of prediction sets indexed by a tuning parameter, which is selected via calibration so that the empirical risk satisfies an adjusted constraint. Under exchangeability and mild regularity conditions, this yields finite-sample guarantees on the desired risk.

Covariate shift (Tibshirani et al., 2019). A line of work relaxes the assumption that calibration and test data are drawn from the same input distribution. Under covariate shift, the marginal distribution of X changes from P to P_{test}, while the conditional distribution $Y \mid X$ remains unchanged. In this setting, weighted conformal methods adjust the calibration procedure by assigning importance weights to each calibration point,

$$w(x) = \frac{dP_{\text{test}}(x)}{dP(x)}. \tag{25}$$

The conformal threshold is then computed as a weighted quantile of the calibration scores, so that samples more representative of the test distribution exert greater influence. When the weights are correctly specified, this approach restores validity under the shifted distribution.

Distribution drift (Hu et al., 2025a). CP has also been extended to settings where the data-generating process evolves over time. In such scenarios, calibration and test samples are no longer identically distributed, and the notion of a fixed test distribution may itself be ill-defined. A common strategy is to exploit temporal structure by assigning larger weights to more recent observations, reflecting their increased relevance for future predictions. The resulting weighted conformal procedures adapt to non-stationarity and produce prediction sets that track the evolving distribution. Under suitable assumptions on the rate of drift, these methods provide approximate coverage guarantees, with deviations that depend explicitly on the magnitude of the distributional change.

Taken together, these extensions illustrate the versatility of the CP framework in accommodating a wide range of deviations from the classical setting. A recurring theme across these developments is the need to move beyond marginal validity toward more refined control of prediction sets, whether through conditioning, alternative risk measures, or adaptation to distributional changes. This perspective

motivates approaches such as *Tunable Prediction Set Size (TPSS)*, which directly impose structural constraints on the prediction sets while retaining the core principles of conformal inference.

Evaluation. The performance of a conformal prediction method is typically assessed through two complementary criteria: the reliability of the prediction sets and their informativeness. These aspects are captured respectively by coverage and the size of the resulting sets, which together characterize the quality of uncertainty quantification.

Coverage. Coverage quantifies the probability that the constructed prediction set contains the true response. For a prescribed miscoverage level $\alpha \in (0, 1)$, the objective is to ensure

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0)\right) \geq 1 - \alpha, \tag{26}$$

for a new test point (X_0, Y_0) drawn from the same distribution as the training data. This guarantee is distribution-free and constitutes the defining property of conformal prediction.

In empirical evaluations, coverage is approximated using a held-out test sample $\{(X_i, Y_i)\}_{i=1}^{n_{\text{test}}}$ via

$$\widehat{\text{Cov}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\left\{Y_i \in \mathcal{C}(X_i)\right\}. \tag{27}$$

This quantity measures the fraction of test instances for which the prediction set successfully contains the true outcome. A method whose empirical coverage aligns with the nominal level $1 - \alpha$ is considered well-calibrated, whereas systematic under-coverage indicates overly aggressive predictions that may compromise reliability.

Prediction set size. While coverage captures validity, it does not reflect how informative the prediction sets are. This aspect is governed by their size, which determines the degree of uncertainty conveyed by the method. For a given coverage level, smaller sets are generally preferred, as they yield more precise and actionable predictions.

The notion of size depends on the structure of the output space. In classification problems, it corresponds to the number of labels included in the set, while in regression it is typically measured by interval length or volume. The empirical average size over a test sample $\{X_i\}_{i=1}^{n_{\text{test}}}$ is defined as

$$\widehat{\text{Size}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}(X_i)|. \tag{28}$$

A.2 Conformal e-value Prediction

Motivation. Split Conformal Prediction (sCP) is fundamentally built upon the notion of p-values, which provide a natural way to quantify how atypical a candidate output is relative to observed data. While p-values have long been a central tool in statistical inference, they have also attracted substantial scrutiny in recent years, motivating the development of alternative frameworks for measuring statistical evidence (Wasserstein et al., 2019). Among the most prominent alternatives are Bayes factors and their non-Bayesian counterparts, known as e-values, which have recently gained increasing attention in the literature (Grünwald et al., 2020; Shafer, 2019; Gauthier et al., 2025b).

Interestingly, ideas closely related to e-values appeared in early precursors of conformal methods, although under different terminology. In particular, what we now refer to as conformal e-value prediction (Vovk, 2025) was already implicitly present in early developments of conformal inference during the 1990s (Gammerman et al., 2013; Vovk et al., 1999). Despite these early developments, subsequent work has predominantly centered on p-value–based constructions such as sCP. One important reason is their direct and interpretable validity guarantee: prediction sets can be expressed through a simple probabilistic statement controlling the error rate at a prescribed level (Vovk et al., 2005; Nouretdinov et al., 2001). This formulation allows practitioners to immediately translate the parameter α into a bound on miscoverage, which is both intuitive and operationally convenient.

By contrast, guarantees derived from e-values are inherently less explicit. While e-values satisfy a moment condition, translating this into a bound on error probabilities requires the use of inequalities such as Markov's inequality, which can lead to looser and less interpretable guarantees (Vovk et al., 1999). Moreover, beyond interpretability, p-value–based conformal predictors exhibit a stronger form of validity in sequential settings: in the online framework, errors are shown to occur independently

(Vovk, 2002), implying that small nominal error probabilities translate, with high probability, into a low empirical frequency of mistakes over time (Vovk et al., 2005). This strengthened notion of validity is not directly available in the eCP framework.

Formulation. In Section 2, we briefly introduced conformal e-value prediction. We now provide a more detailed formulation and highlight its key components.

We consider the same data-splitting framework as in Split Conformal Prediction (sCP). The available data are partitioned into a training set and a calibration set

$$\mathcal{C} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}, \tag{29}$$

and a predictive model $\hat{f}$ is trained using only the training data. A nonconformity score function s is then employed to evaluate how well predictions align with observed outcomes, yielding calibration scores

$$s_i := s\big(\hat{f}(X_i), Y_i\big), \quad i = 1, \ldots, n, \tag{30}$$

and for a test pair (X_0, Y_0),

$$s_0 := s\big(\hat{f}(X_0), Y_0\big). \tag{31}$$

E-values and their properties. An e-value is a nonnegative random variable E satisfying the moment condition

$$\mathbb{E}[E] \leq 1. \tag{32}$$

This definition differs fundamentally from that of p-values, which are characterized by a uniform distribution under the null. Instead, e-values are defined through a first-moment constraint, and can be interpreted as measures of evidence against a null hypothesis: larger values of E correspond to stronger evidence.

A key advantage of this formulation is that it immediately yields a tail bound via Markov's inequality. Specifically, for any $t > 0$,

$$\mathbb{P}(E \geq t) \leq \frac{1}{t}. \tag{33}$$

This inequality provides a simple and general mechanism for controlling the probability that E exceeds a prescribed threshold, and forms the basis for constructing valid prediction sets in the conformal framework.

E-value construction. In the eCP framework, the calibration scores $\{s_i\}_{i=1}^n$ and the test score s_0 are combined into a single scalar quantity that measures how atypical the candidate pair (X_0, Y_0) is relative to the calibration data. Specifically, define

$$E(X_0, Y_0) := \frac{s_0}{\frac{1}{n+1}\left(s_0 + \sum_{i=1}^n s_i\right)}, \tag{34}$$

where $s_0 = s(\hat{f}(X_0), Y_0)$. This normalization compares the test score to the average score over the augmented sample $\{s_0, s_1, \ldots, s_n\}$. Consequently, larger values of $E(X_0, Y_0)$ indicate that the candidate label Y_0 is relatively more nonconforming than the calibration observations.

A crucial property of this construction is that, under the exchangeability of $(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n)$, the random variable $E(X_0, Y_0)$ is nonnegative and satisfies

$$\mathbb{E}\left[E(X_0, Y_0)\right] \leq 1. \tag{35}$$

Thus, $E(X_0, Y_0)$ is a valid e-value. Intuitively, this follows from the symmetry of the augmented sample: no single score, including s_0, plays a distinguished role in expectation, which prevents $E(X_0, Y_0)$ from being systematically large.

This observation enables a direct construction of prediction sets. Indeed, since $E(X_0, Y_0)$ satisfies a first-moment bound, Markov's inequality implies that the probability of $E(X_0, Y_0)$ exceeding a threshold t is at most $1/t$. This suggests selecting a threshold of order $1/\alpha$ in order to control the miscoverage probability at level α.

Prediction set. Motivated by the above, for a target level $\alpha \in (0, 1)$, we define

$$\mathcal{C}(X_0) = \left\{ y \in \mathcal{Y} : E(X_0, y) < \frac{1}{\alpha} \right\}. \tag{36}$$

In other words, a candidate label is retained if its associated e-value does not exceed the threshold $1/\alpha$, and is discarded otherwise.

The validity of this construction is formalized in the following result.

Theorem A.4 (Coverage guarantee of eCP (Vovk, 2025)). *Assume that* $(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n)$ *are exchangeable. Then*

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0)\right) \geq 1 - \alpha. \tag{37}$$

Proof sketch. By construction, $E(X_0, Y_0)$ is a nonnegative random variable satisfying $\mathbb{E}[E(X_0, Y_0)] \leq 1$. Applying Markov's inequality with $t = 1/\alpha$ yields

$$\mathbb{P}\left(E(X_0, Y_0) \geq \frac{1}{\alpha} \right) \leq \alpha. \tag{38}$$

Therefore,

$$\mathbb{P}\left(E(X_0, Y_0) < \frac{1}{\alpha} \right) \geq 1 - \alpha. \tag{39}$$

By definition of $\mathcal{C}(X_0)$ in equation 2, the event $\{Y_0 \in \mathcal{C}(X_0)\}$ is equivalent to $\{E(X_0, Y_0) < 1/\alpha\}$, which establishes the result. □

Comparison with sCP. Although eCP enjoys the same marginal coverage guarantee as sCP, the underlying mechanism is fundamentally different. In sCP, validity is derived from an exact finite-sample symmetry argument based on the rank of the test score, which yields a tight control of the error probability. In contrast, eCP relies solely on the moment condition $\mathbb{E}[E] \leq 1$ together with Markov's inequality. As a consequence, the bound

$$\mathbb{P}\left(E(X_0, Y_0) \geq \frac{1}{\alpha} \right) \leq \alpha$$

may be loose, since Markov's inequality does not exploit finer distributional structure.

This distinction implies that, while eCP allows greater flexibility—particularly in enabling data-dependent constructions, it generally provides weaker control compared to p-value-based methods. In particular, the resulting prediction sets may be more conservative, and the guarantee is not derived from an exact uniformity property but rather from a first-moment bound.

Recent extensions of Conformal e-value Prediction. Although conformal e-value prediction (eCP) is generally weaker than Split Conformal Prediction (sCP) in terms of the tightness and interpretability of its guarantees, its flexibility enables a range of constructions that are difficult or impossible to achieve within the classical p-value framework. In particular, the use of e-values allows for greater adaptability in how evidence is aggregated and thresholds are chosen, which can be leveraged to address more complex prediction settings. Recent work (Gauthier et al., 2025b) has demonstrated that this additional flexibility significantly broadens the scope of conformal prediction, leading to several novel extensions that go beyond the capabilities of standard sCP. We briefly highlight three representative examples.

Batch anytime-valid conformal prediction. In many practical scenarios, predictions are made sequentially or in batches over time, and one may wish to maintain validity uniformly over multiple prediction rounds without fixing the total number of predictions in advance. Classical conformal methods are not naturally suited for such *anytime-valid* settings, as their guarantees typically hold only for a fixed sample size. E-values provide a natural solution to this limitation. Due to their multiplicative and martingale-like structure, e-values can be aggregated across time while preserving validity. This enables the construction of prediction sets that remain valid at all time steps simultaneously, even under adaptive or sequential usage. As a result, eCP can be extended to batch or online settings where validity must hold uniformly over an evolving sequence of predictions.

Tunable prediction set size. Which is also referred to in the literature as *fixed-size conformal sets with data-dependent coverage*. As introduced in Section 2, the goal is to construct prediction sets with controlled size while allowing coverage to vary in a data-dependent manner. A key limitation of classical sCP is that set size is only indirectly controlled via the miscoverage level α, making explicit size constraints difficult to enforce. In many applications, it is more natural to fix the size and let coverage adapt, a formulation that is not readily handled within the p-value framework due to its reliance on inverting a fixed rank-based threshold. In contrast, as discussed earlier, the e-value formulation enables more flexible thresholding: since validity follows from a moment condition rather than an exact rank argument, it allows prediction sets with prescribed size and data-dependent coverage, at the cost of replacing a fixed coverage guarantee with an expected or adaptive one.

Conformal prediction under ambiguous ground truth. In some settings, the notion of a single ground-truth label is itself ill-defined, such as in problems involving annotator disagreement or probabilistic labeling. Classical conformal prediction assumes access to a well-defined target Y, and thus does not naturally extend to such ambiguous settings. E-values offer a flexible mechanism for handling this scenario by allowing conformity scores to incorporate distributions or multiple candidate labels. Rather than requiring a single deterministic label, the e-value construction can aggregate evidence across possible outcomes, leading to prediction sets that remain valid under uncertainty in the ground truth. This enables conformal prediction to be applied in settings with inherently ambiguous or noisy labels, which are increasingly common in modern machine learning applications.

A.3 The Tunable Prediction Set Size problem

Motivation. The tunable prediction set size (TPSS) problem was recently introduced in Gauthier et al. (2025b) to address a fundamental limitation of classical conformal prediction. Standard conformal methods operate by fixing a target coverage level α in advance, ensuring that the true outcome is contained in the prediction set with probability at least $1-\alpha$. While this provides strong and distribution-free guarantees, it leaves the size of the resulting prediction sets uncontrolled. In practice, however, the size of the prediction set is often of primary importance. For a fixed coverage level, the size of conformal sets can vary substantially across inputs and data distributions, leading to sets that may be either overly large and uninformative or too small to be reliable. This lack of direct control makes it difficult to enforce practical constraints on the number or volume of candidate outputs returned by the model. The TPSS problem reverses this perspective by treating the size of the prediction set as the primary quantity to be controlled. Instead of fixing the coverage level in advance, one specifies a size constraint and allows the resulting coverage to adapt accordingly. This shift enables the construction of prediction sets that are better aligned with practical requirements, where interpretability and usability often depend critically on the size of the output.

Formulation. In Section 2, we introduced the tunable prediction set size (TPSS) problem at a high level. We now provide a more detailed formulation.

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces. Unlike classical conformal prediction, which fixes a miscoverage level α and induces the size of the prediction set as a byproduct, the TPSS framework treats the size as a primary design constraint. To formalize this, we introduce a *budget function* $\mathcal{S} : \mathcal{X} \to \mathbb{R}_{\geq 0}$, which assigns to each input $x \in \mathcal{X}$ a maximum allowable size for the prediction set. The goal is to construct a set-valued predictor $\mathcal{C} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ such that, for a test point X_0,

$$|\mathcal{C}(X_0)| \leq \mathcal{S}(X_0). \tag{40}$$

The budget function $\mathcal{S}$ provides flexibility in specifying task-dependent constraints. For example, it may be constant, enforcing a uniform bound across all inputs, or depend on x, allowing larger sets for more uncertain inputs and smaller sets for more confident ones. Here, $|\cdot|$ denotes a user-specified notion of size defined on subsets of $\mathcal{Y}$. In discrete settings, where $\mathcal{Y}$ is finite or countable, $|\mathcal{C}(x)|$ naturally corresponds to the cardinality of the set. In continuous settings, where $\mathcal{Y} \subseteq \mathbb{R}^d$, $|\mathcal{C}(x)|$ can instead be defined via a suitable measure, typically the Lebesgue measure, so that it captures the volume of the prediction set. More generally, $|\cdot|$ may be interpreted as a size functional tailored to the structure of $\mathcal{Y}$, allowing the framework to accommodate a broad class of output spaces.

Beyond satisfying the size constraint, one also seeks to retain a meaningful coverage guarantee. Specifically, the aim is to construct $\mathcal{C}$ such that

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0)\right) \geq 1 - \alpha \tag{41}$$

for some $\alpha \in (0, 1)$, or a suitable relaxation thereof, while respecting the budget $\mathcal{S}(X_0)$. Thus, the objective is not to optimize coverage, but to preserve validity under an explicit size constraint.

In practice, enforcing the constraint exactly for all inputs can be difficult when using finite calibration data. We therefore consider a relaxed formulation in which the constraint holds with high probability,

$$\mathbb{P}\left(|\mathcal{C}(X_0)| \leq \mathcal{S}(X_0)\right) \to 1 \quad \text{as } n \to \infty, \tag{42}$$

where n is the calibration sample size. This captures the requirement that, with sufficient data, the prediction sets satisfy the prescribed budget with high probability.

Applications. To emphasize the importance of the Tunable Prediction Set Size problem in conformal prediction, we will now provide several real-world applications.

Medical diagnosis. In healthcare, predictive models are often used to assist clinicians in identifying potential diseases from patient data. Conformal prediction provides sets of plausible diagnoses with statistical guarantees, which can guide further testing or treatment. However, due to time and resource constraints, practitioners typically consider only a limited number of candidates (e.g., the top few diagnoses). The TPSS framework is particularly well-suited in this setting, as it allows one to directly constrain the number of returned diagnoses while maintaining a meaningful coverage guarantee, ensuring that the true condition is included with high probability.

Autonomous driving and perception. In safety-critical systems such as autonomous vehicles, models are used to detect and classify objects in the environment. Conformal prediction can provide sets of possible object labels or trajectories, capturing uncertainty in perception. However, downstream decision-making modules require concise outputs for real-time processing. TPSS enables the system to limit the number of hypotheses or candidate predictions, ensuring tractable and interpretable outputs while preserving reliability under uncertainty.

Recommendation systems. Modern recommendation systems often present users with a shortlist of items (e.g., products, movies, or songs). Conformal prediction can generate sets of relevant items with coverage guarantees, but unconstrained sets may be too large to be useful. TPSS directly addresses this by enforcing a fixed budget on the number of recommendations, allowing the system to provide a compact, high-quality set of suggestions while maintaining statistical validity.

Natural language generation and retrieval. In tasks such as machine translation, text generation, or document retrieval, models may produce multiple plausible outputs. Conformal methods can quantify uncertainty by returning sets of candidate sequences or documents. However, presenting large sets is impractical for users. TPSS allows one to restrict the number of generated candidates or retrieved documents, ensuring that outputs remain manageable while still covering the true or relevant responses with high probability.

Computer vision and image recognition. In computer vision tasks such as image classification and object detection, models operate on high-dimensional inputs and often require substantial computational and memory resources. Conformal prediction can provide sets of plausible labels to capture uncertainty, but large prediction sets can significantly exacerbate computational costs. In computer vision, images are high-dimensional objects that already require substantial time and memory to process. When multiple candidate labels are considered, each may require additional passes over the image or further feature extraction and scoring, effectively multiplying the computational cost. As a result, large prediction sets can quickly become prohibitively expensive, making real-time or large-scale deployment impractical. The TPSS framework addresses this issue by enforcing a strict budget on the number of candidate labels. By limiting the size of the prediction set, it reduces computational overhead and prevents the accumulation of redundant hypotheses, while still maintaining a principled coverage guarantee.

Anomaly detection and monitoring. In monitoring systems for cybersecurity, finance, or industrial processes, models are used to identify anomalous or high-risk events. Conformal prediction can produce sets of candidate anomalies or alert categories. In practice, however, only a limited number of alerts can be investigated due to resource constraints. TPSS allows the system to restrict the number of alerts while maintaining coverage guarantees, improving both efficiency and reliability in decision-making.

B THEORETICAL PROOFS

B.1 PROOF OF THE COVERAGE GUARANTEE - THEOREM 4.2

Proof. Fix $\delta \in (0, \zeta]$. As in the standard split conformal analysis, we condition throughout on the training sample used to construct $\widehat{f}$, so that $\widehat{f}$ and the nonconformity score function are fixed. All probabilities and expectations below are therefore taken with respect to $\mathcal{C}_a$, $\mathcal{C}_c$, and the independent test pair (X_0, Y_0).

Recall that the calibration scores associated with $\mathcal{C}_c$ are $s_i^{(c)} = s(\widehat{f}(X_i^c), Y_i^c)$, for $i = 1, \ldots, n$. Let $s_{(1)}^{(c)} < \cdots < s_{(n)}^{(c)}$ denote their order statistics, where the strict inequalities hold almost surely under Assumption 4.1. By the definition of the empirical conformal quantile,

$$\widehat{q}_c\left(\alpha_\delta(X_0)\right) = s_{(\lceil (1-\alpha_\delta(X_0))(n+1) \rceil)}^{(c)}. \tag{43}$$

When $\lceil (1-\alpha_\delta(X_0))(n+1) \rceil = n+1$, we use the standard conformal convention $s_{(n+1)}^{(c)} = +\infty$, which is consistent with $\mathrm{BinTail}_n(n+1, p) = 0$.

Write $s_0 = s(\widehat{f}(X_0), Y_0)$. By the definition of $C_\delta(X_0)$,

$$\{Y_0 \notin C_\delta(X_0)\} = \{s_0 > \widehat{q}_c\left(\alpha_\delta(X_0)\right)\} \tag{44}$$

$$= \left\{ s_0 > s_{(\lceil (1-\alpha_\delta(X_0))(n+1) \rceil)}^{(c)} \right\}. \tag{45}$$

Since ties occur with probability zero, the event that s_0 is strictly larger than the $\lceil (1-\alpha_\delta(X_0))(n+1) \rceil$-th order statistic is equivalent to the event that at least $\lceil (1-\alpha_\delta(X_0))(n+1) \rceil$ calibration scores are strictly smaller than s_0. Hence,

$$\{Y_0 \notin C_\delta(X_0)\} = \left\{ \sum_{i=1}^{n} \mathbf{1}\left\{ s_i^{(c)} < s_0 \right\} \geq \lceil (1-\alpha_\delta(X_0))(n+1) \rceil \right\}. \tag{46}$$

We now condition on $(X_0, Y_0, \mathcal{C}_a)$. Under this conditioning, s_0 and $\alpha_\delta(X_0)$ are fixed. Moreover, because $\mathcal{C}_c$ is independent of $(X_0, Y_0, \mathcal{C}_a)$, the scores $s_1^{(c)}, \ldots, s_n^{(c)}$ remain independent draws from the marginal score distribution with CDF F. Therefore, almost surely,

$$\mathcal{L}\left(\sum_{i=1}^{n} \mathbf{1}\left\{ s_i^{(c)} < s_0 \right\} \middle| X_0, Y_0, \mathcal{C}_a \right) = \mathrm{Binomial}\left(n, F(s_0)\right). \tag{47}$$

It follows from the definition of $\mathrm{BinTail}_n$ that

$$\mathbb{P}\left(Y_0 \notin C_\delta(X_0) | X_0, Y_0, \mathcal{C}_a\right) = \mathrm{BinTail}_n\left(\lceil (1-\alpha_\delta(X_0))(n+1) \rceil, F\left(s\left(\widehat{f}(X_0), Y_0 \right) \right) \right). \tag{48}$$

Applying the tower property conditional on $\mathcal{C}_a$ gives

$$\mathbb{P}\left(Y_0 \notin C_\delta(X_0) | \mathcal{C}_a\right) = \mathbb{E}\left[\mathrm{BinTail}_n\left(\lceil (1-\alpha_\delta(X_0))(n+1) \rceil, F\left(s\left(\widehat{f}(X_0), Y_0 \right) \right) \right) \middle| \mathcal{C}_a \right]. \tag{49}$$

By the definition of $\Delta_{\delta,n}(\mathcal{C}_a)$,

$$\Delta_{\delta,n}(\mathcal{C}_a) = \mathbb{E}\left[\mathrm{BinTail}_n\left(\lceil (1-\alpha_\delta(X_0))(n+1) \rceil, F\left(s\left(\widehat{f}(X_0), Y_0 \right) \right) \right) - \alpha_\delta(X_0) \middle| \mathcal{C}_a \right] \tag{50}$$

$$= \mathbb{P}\left(Y_0 \notin C_\delta(X_0) | \mathcal{C}_a\right) - \mathbb{E}\left[\alpha_\delta(X_0) | \mathcal{C}_a\right]. \tag{51}$$

Rearranging yields the following exact conditional coverage identity:

$$\mathbb{P}\left(Y_0 \in C_\delta(X_0) | \mathcal{C}_a\right) = 1 - \mathbb{E}\left[\alpha_\delta(X_0) | \mathcal{C}_a\right] - \Delta_{\delta,n}(\mathcal{C}_a). \tag{52}$$

Finally, taking expectation over $\mathcal{C}_a$ and applying the tower property once more, we obtain

$$\mathbb{P}\left(Y_0 \in C_\delta(X_0)\right) = \mathbb{E}\left[\mathbb{P}\left(Y_0 \in C_\delta(X_0) | \mathcal{C}_a\right)\right] \tag{53}$$

$$= 1 - \mathbb{E}\left[\mathbb{E}\left[\alpha_\delta(X_0) | \mathcal{C}_a\right]\right] - \mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right] \tag{54}$$

$$= 1 - \mathbb{E}\left[\alpha_\delta(X_0)\right] - \mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right]. \tag{55}$$

which proves the stated result. □

B.2 Proof of Soft Size Control- Theorem 4.4

Proof. We first derive the following formula

Lemma B.1. *Under Assumption 4.1, suppose additionally that the quantile function $Q(u) := F^{-1}(u)$ is continuous on $(\varepsilon/2, 1-\varepsilon/2)$. Then, for every $\eta > 0$ and every fixed $\varepsilon \in (0,1)$, we have*

$$\mathbb{P}\left(\sup_{\alpha\in[\varepsilon,1-\varepsilon]} |\hat{q}_a(\alpha) - \hat{q}_c(\alpha)| > \eta\right) \to 0 \qquad \text{as } n \to \infty. \tag{56}$$

Proof. Under Assumption 4.1, the samples in $\mathcal{C}_a$ and $\mathcal{C}_c$ are i.i.d. draws from the same distribution, and the score function s is deterministic. Consequently, the nonconformity scores $s_i^{(a)} = s\big(\hat{f}(X_i^a), Y_i^a\big)$ and $s_i^{(c)} = s\big(\hat{f}(X_i^c), Y_i^c\big)$ are i.i.d. draws from a common distribution with cumulative distribution function F. In particular, the score sets $\left\{s_i^{(a)}\right\}_{i=1}^n$ and $\left\{s_i^{(c)}\right\}_{i=1}^n$ are independent samples from F. Let $\hat{F}_{\mathcal{C}_a}(t)$ and $\hat{F}_{\mathcal{C}_c}(t)$ denote the corresponding empirical distribution functions. By the Glivenko–Cantelli theorem,

$$\sup_{t\in\mathbb{R}} |\hat{F}_{\mathcal{C}_a}(t) - F(t)| \xrightarrow{a.s.} 0, \qquad \sup_{t\in\mathbb{R}} |\hat{F}_{\mathcal{C}_c}(t) - F(t)| \xrightarrow{a.s.} 0. \tag{57}$$

Since $\varepsilon > 0$ is fixed, for all sufficiently large n, the empirical quantile ranks corresponding to $\alpha \in [\varepsilon, 1-\varepsilon]$ lie in $\{1, \dots, n\}$.

Since the quantile function $Q = F^{-1}$ is continuous on the interval $(\varepsilon/2, 1-\varepsilon/2)$ by assumption, uniform convergence of the empirical distribution functions implies uniform convergence of the corresponding empirical quantile functions on $[\varepsilon, 1-\varepsilon]$. Hence,

$$\sup_{\alpha\in[\varepsilon,1-\varepsilon]} |\hat{q}_a(\alpha) - q(\alpha)| \xrightarrow{a.s.} 0, \qquad \sup_{\alpha\in[\varepsilon,1-\varepsilon]} |\hat{q}_c(\alpha) - q(\alpha)| \xrightarrow{a.s.} 0. \tag{58}$$

Therefore,

$$\sup_{\alpha\in[\varepsilon,1-\varepsilon]} |\hat{q}_a(\alpha) - \hat{q}_c(\alpha)| \xrightarrow{a.s.} 0. \tag{59}$$

Equivalently, for every $\eta > 0$,

$$\mathbb{P}\left(\sup_{\alpha\in[\varepsilon,1-\varepsilon]} |\hat{q}_a(\alpha) - \hat{q}_c(\alpha)| > \eta\right) \to 0. \tag{60}$$

This completes the proof. □

For all sufficiently large n, the set defining $\alpha_\delta(X_0)$ is nonempty almost surely by assumption. Hence, by construction of $\alpha_\delta(X_0)$,

$$N_{X_0}\left(\hat{q}_a(\alpha_\delta(X_0))\right) \le \mathcal{S}(X_0) - \delta. \tag{61}$$

On the other hand, $|C_\delta(X_0)| = N_{X_0}\left(\hat{q}_c(\alpha_\delta(X_0))\right)$. Hence $|C_\delta(X_0)| \le \mathcal{S}(X_0)$ whenever

$$N_{X_0}\left(\hat{q}_c(\alpha_\delta(X_0))\right) - N_{X_0}\left(\hat{q}_a(\alpha_\delta(X_0))\right) \le \delta. \tag{62}$$

Therefore,

$$\mathbb{P}\left(|C_\delta(X_0)| \le \mathcal{S}(X_0)\right) \ge \mathbb{P}\left(N_{X_0}\left(\hat{q}_c(\alpha_\delta(X_0))\right) - N_{X_0}\left(\hat{q}_a(\alpha_\delta(X_0))\right) \le \delta\right). \tag{63}$$

By definition, $\alpha_\delta(X_0) \in [\varepsilon, 1-\varepsilon]$, where $\varepsilon \in (0, 1/2)$ is the truncation parameter. Therefore,

$$|\hat{q}_c(\alpha_\delta(X_0)) - \hat{q}_a(\alpha_\delta(X_0))| \le \sup_{\alpha\in[\varepsilon,1-\varepsilon]} |\hat{q}_c(\alpha) - \hat{q}_a(\alpha)|. \tag{64}$$

By Lemma B.1, the right-hand side converges to 0 in probability. Hence,

$$|\hat{q}_c(\alpha_\delta(X_0)) - \hat{q}_a(\alpha_\delta(X_0))| \to 0 \quad \text{in probability}, \tag{65}$$

and consequently,

$$\mathbb{P}\left(|\hat{q}_c(\alpha_\delta(X_0)) - \hat{q}_a(\alpha_\delta(X_0))| < \eta_0\right) \to 1. \tag{66}$$

On this event, applying Assumption 4.3 and the fact that $L > 0$, we have

$$N_{X_0}\left(\hat{q}_c(\alpha_\delta(X_0))\right) - N_{X_0}\left(\hat{q}_a(\alpha_\delta(X_0))\right) \le L. \tag{67}$$

Since $\delta \ge L$, it follows that $N_{X_0}\left(\hat{q}_c(\alpha_\delta(X_0))\right) - N_{X_0}\left(\hat{q}_a(\alpha_\delta(X_0))\right) \le \delta$ with probability tending to 1. Therefore,

$$\mathbb{P}\left(|C_\delta(X_0)| \le \mathcal{S}(X_0)\right) \to 1. \tag{68}$$

This completes the proof. □

B.3 Proof of LOO Estimator Convergence Rate - Theorem 5.2

Proof. We prove the two convergence statements jointly. The main argument relies on the same stability property of the adaptive level under the deletion or replacement of a single approximation observation. We therefore begin by establishing this property and then use it for both $\hat{\alpha}_\delta$ and $\widehat{\Delta}_{\delta,n}$.

Recall that

$$\hat{\alpha}_\delta = \frac{1}{n}\sum_{k=1}^{n} \alpha_\delta^k(X_k^a), \tag{69}$$

where, for each $k \in [n]$,

$$\alpha_\delta^k(x) := \inf\left\{\alpha \in [\varepsilon, 1-\varepsilon] : \left|\left\{y \in \mathcal{Y} : s(\hat{f}(x), y) \le \hat{q}_a^k(\alpha)\right\}\right| \le \mathcal{S}(x) - \delta\right\}, \tag{70}$$

and

$$\mathcal{C}_a^k = \mathcal{C}_a \setminus \{(X_k^a, Y_k^a)\}. \tag{71}$$

For every $x \in \mathcal{X}$, define

$$N_x(t) := \left|\left\{y \in \mathcal{Y} : s(\hat{f}(x), y) \le t\right\}\right|. \tag{72}$$

By construction, $N_x(t)$ is nondecreasing in t.

Since the observations in $\mathcal{C}_a$ are i.i.d. under Assumption 4.1, the distribution of any quantity induced by an approximation sample depends on the sample only through its size. For each $m \ge 1$, define

$$\alpha_{\delta,m}(x) := \inf\left\{\alpha \in [\varepsilon, 1-\varepsilon] : \left|\left\{y \in \mathcal{Y} : s(\hat{f}(x), y) \le \hat{q}_m(\alpha)\right\}\right| \le \mathcal{S}(x) - \delta\right\}, \tag{73}$$

where $\hat{q}_m(\alpha)$ denotes the empirical $(1-\alpha)$-quantile constructed from an independent approximation sample of size m. In particular,

$$\alpha_\delta(X_0) = \alpha_{\delta,n}(X_0). \tag{74}$$

Step 1: A common rank-stability property.

Fix $x \in \mathcal{X}$. Let $R_1, \dots, R_n$ denote the conformity scores computed from an approximation sample of size n, and write

$$R_{(1)} \le \cdots \le R_{(n)} \tag{75}$$

for their order statistics. If one observation is deleted, write

$$R_{(1)}^{(-)} \le \cdots \le R_{(n-1)}^{(-)} \tag{76}$$

for the resulting order statistics.

For a sample of size m, let $j_m(\alpha)$ denote the rank associated with the empirical $(1-\alpha)$-quantile. The map $j_m(\alpha)$ takes values in $[m]$, is nonincreasing in α, and adjacent values of α on the empirical quantile grid differ by at most C_0/m for a universal constant $C_0 > 0$. Hence,

$$\alpha_{\delta,m}(x) = \inf\left\{\alpha \in [\varepsilon, 1-\varepsilon] : N_x\left(R_{(j_m(\alpha))}^{(m)}\right) \le \mathcal{S}(x) - \delta\right\}, \tag{77}$$

where $R_{(r)}^{(n)} := R_{(r)}$ and $R_{(r)}^{(n-1)} := R_{(r)}^{(-)}$.

Deleting one score changes the rank location of any threshold by at most one. In particular, for every admissible rank r,

$$R_{(r-1)} \le R_{(r)}^{(-)} \le R_{(r+1)}, \tag{78}$$

with the usual modifications at the boundaries.

Since the empirical quantile map is piecewise constant in α, the infimum in equation 77 depends only on the induced empirical rank. Thus, the optimization over $\alpha \in [\varepsilon, 1-\varepsilon]$ may equivalently be restricted to the empirical quantile grid. Moreover, because $N_x(t)$ is nondecreasing in t, the feasible-rank set

$$\left\{ r : N_x\left(R_{(r)}^{(m)}\right) \le \mathcal{S}(x) - \delta \right\} \tag{79}$$

is an initial segment of $\{1, \dots, m\}$.

Let $\alpha_r^{(m)}$ denote the smallest value on the empirical grid for which

$$j_m\left(\alpha_r^{(m)}\right) = r. \tag{80}$$

Adjacent ranks correspond to adjacent, or possibly identical, grid values, so

$$\left|\alpha_{r+1}^{(m)} - \alpha_r^{(m)}\right| \le \frac{C_0}{m}. \tag{81}$$

By the interlacing property in equation 78, deleting one observation can move the boundary rank of the feasible set in equation 79 by at most one. Combining this fact with the grid bound equation 81, there exists a universal constant $C_\alpha > 0$ such that

$$|\alpha_{\delta,n-1}(x) - \alpha_{\delta,n}(x)| \le \frac{C_\alpha}{n} \qquad \text{for every } x \in \mathcal{X}. \tag{82}$$

Consequently,

$$|\mathbb{E}\left[\alpha_{\delta,n-1}(X_0)\right] - \mathbb{E}\left[\alpha_{\delta,n}(X_0)\right]| = \mathcal{O}\left(n^{-1}\right). \tag{83}$$

The same argument applies when one observation in a sample of fixed size m is replaced rather than deleted. Indeed, replacing one score changes the empirical counting function at every threshold by at most one and, equivalently, changes the rank location of every threshold by at most one. Thus the boundary rank of the feasible-rank set also moves by at most one. Consequently, if $\alpha_{\delta,m}$ and $\alpha'_{\delta,m}$ are constructed from two approximation samples of size m differing in exactly one observation, then

$$\left|\alpha_{\delta,m}(x) - \alpha'_{\delta,m}(x)\right| \le \frac{C_\alpha}{m}, \qquad x \in \mathcal{X}. \tag{84}$$

We will repeatedly use equation 82 and equation 84 below.

Step 2: Convergence of $\hat{\alpha}_\delta$.

For every $k \in [n]$, the i.i.d. assumption gives

$$\left(\mathcal{C}_a^k, (X_k^a, Y_k^a)\right) \overset{d}{=} \left(\mathcal{C}_{n-1}, (X_0, Y_0)\right), \tag{85}$$

where $\mathcal{C}_{n-1}$ is an independent approximation sample of size $n-1$. Therefore,

$$\mathbb{E}[\hat{\alpha}_\delta] = \mathbb{E}\left[\alpha_{\delta,n-1}(X_0)\right]. \tag{86}$$

Combining equation 86 with equation 83 gives

$$\begin{aligned} \mathbb{E}\left[\alpha_{\delta,n}(X_0)\right] &= \mathbb{E}[\hat{\alpha}_\delta] + \left(\mathbb{E}\left[\alpha_{\delta,n}(X_0)\right] - \mathbb{E}\left[\alpha_{\delta,n-1}(X_0)\right]\right) \\ &= \mathbb{E}[\hat{\alpha}_\delta] + \mathcal{O}\left(n^{-1}\right). \end{aligned} \tag{87}$$

It remains to control the random fluctuation of $\hat{\alpha}_\delta$ around its expectation. Define the symmetric functional

$$\Phi(Z_1, \ldots, Z_n) := \hat{\alpha}_\delta, \qquad Z_i := (X_i^a, Y_i^a). \tag{88}$$

Let Z_i' be an independent replacement of the i-th observation, and define

$$\Phi_i' := \Phi\left(Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n\right). \tag{89}$$

Write

$$\Phi = \frac{1}{n}\sum_{k=1}^{n} A_k, \qquad \Phi_i' = \frac{1}{n}\sum_{k=1}^{n} A_k', \tag{90}$$

where

$$A_k := \alpha_\delta^k(X_k^a) \tag{91}$$

and A_k' denotes the corresponding quantity after replacing Z_i by Z_i'.

We distinguish two cases. First, when $k = i$, both A_i and A_i' belong to $[\varepsilon, 1-\varepsilon]$, and hence

$$|A_i - A_i'| \le 1. \tag{92}$$

Second, for $k \neq i$, the pseudo-test point X_k^a is unchanged, and the two pseudo-calibration samples differ in exactly one observation. Therefore, by equation 84,

$$|A_k - A_k'| \le \frac{C_\alpha}{n-1}, \qquad k \neq i. \tag{93}$$

Equivalently, this second bound can be seen directly from the empirical rank representation. If

$$R_1^{(k)}, \ldots, R_{n-1}^{(k)} \tag{94}$$

are the conformity scores in $\mathcal{C}_a^k$, with order statistics

$$R_{(1)}^{(k)} < \cdots < R_{(n-1)}^{(k)} \tag{95}$$

almost surely under Assumption 4.1, then the feasible-rank set

$$\left\{ r \in \{1, \ldots, n-1\} : N_x\left(R_{(r)}^{(k)}\right) \le \mathcal{S}(x) - \delta \right\} \tag{96}$$

is an initial segment of $\{1, \ldots, n-1\}$. Replacing one observation changes the empirical counting function

$$C_k(t) := \sum_{\ell=1}^{n-1} \mathbf{1}\left\{ R_\ell^{(k)} \le t \right\} \tag{97}$$

by at most one at every threshold t. Therefore, the boundary feasible rank changes by at most one, which together with the empirical-grid mesh bound again yields equation 93.

Combining equation 92 and equation 93,

$$\begin{aligned} |\Phi - \Phi_i'| &\le \frac{1}{n}\left(1 + \sum_{k \neq i} \frac{C_\alpha}{n-1}\right) \\ &= \frac{1 + C_\alpha}{n}. \end{aligned} \tag{98}$$

Thus, Φ satisfies the bounded-difference condition with coordinate constants $(1 + C_\alpha)/n$. McDiarmid's inequality therefore gives, for every $t > 0$,

$$\begin{aligned} \mathbb{P}\left(|\hat{\alpha}_\delta - \mathbb{E}[\hat{\alpha}_\delta]| \ge t\right) &\le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}((1 + C_\alpha)/n)^2}\right) \\ &= 2\exp\left(-\frac{2nt^2}{(1 + C_\alpha)^2}\right). \end{aligned} \tag{99}$$

Hence,

$$\sqrt{n}\,(\hat{\alpha}_\delta - \mathbb{E}[\hat{\alpha}_\delta]) = \mathcal{O}_P(1), \tag{100}$$

or equivalently,

$$\hat{\alpha}_\delta - \mathbb{E}[\hat{\alpha}_\delta] = \mathcal{O}_P(n^{-1/2}). \tag{101}$$

Finally, by equation 74 and equation 87,

$$\mathbb{E}[\alpha_\delta(X_0)] = \mathbb{E}[\hat{\alpha}_\delta] + \mathcal{O}(n^{-1}), \tag{102}$$

and therefore

$$\begin{aligned} |\hat{\alpha}_\delta - \mathbb{E}[\alpha_\delta(X_0)]| &\le |\hat{\alpha}_\delta - \mathbb{E}[\hat{\alpha}_\delta]| + |\mathbb{E}[\hat{\alpha}_\delta] - \mathbb{E}[\alpha_\delta(X_0)]| \\ &= \mathcal{O}_P(n^{-1/2}) + \mathcal{O}(n^{-1}) \\ &= \mathcal{O}_P(n^{-1/2}). \end{aligned} \tag{103}$$

Step 3: A conditional rank-interval bound.

To analyze $\widehat{\Delta}_{\delta,n}$, we additionally require a bound on the probability that a test score falls between calibration order statistics whose ranks differ by only a fixed number.

Let $T_1, \dots, T_n$ be i.i.d. draws from the marginal score distribution F, with order statistics

$$T_{(1)} < \cdots < T_{(n)} \tag{104}$$

almost surely. Set

$$F(T_{(0)}) := 0, \qquad F(T_{(n+1)}) := 1. \tag{105}$$

Fix an integer $h \ge 1$. For any possibly random rank $R \in [n]$ that is independent of $T_1, \dots, T_n$, define

$$I_{R,h} := \left[T_{((R-h)\vee 0)}, T_{((R+h)\wedge(n+1))}\right]. \tag{106}$$

Then

$$\mathbb{E}\left[F(I_{R,h})\right] \le \frac{2h}{n+1}, \tag{107}$$

and

$$\mathbb{E}\left[F(I_{R,h})^2\right] \le \frac{2h(2h+1)}{(n+1)(n+2)}. \tag{108}$$

Both bounds hold uniformly over the distribution of R.

To verify these bounds, Assumption 4.1 implies that the marginal score distribution has no atoms, so its CDF F is continuous. Hence,

$$U_j := F(T_j), \qquad j = 1, \dots, n, \tag{109}$$

are i.i.d. $\mathrm{Unif}(0,1)$ random variables. Let

$$U_{(1)} < \cdots < U_{(n)}, \qquad U_{(0)} := 0, \qquad U_{(n+1)} := 1. \tag{110}$$

Conditioning on $R = r$,

$$F(I_{R,h}) = U_{((r+h)\wedge(n+1))} - U_{((r-h)\vee 0)}. \tag{111}$$

This quantity is the sum of at most $2h$ consecutive uniform spacings. The vector of the $n+1$ uniform spacings has a $\mathrm{Dirichlet}(1, \dots, 1)$ distribution. Consequently, if W_d denotes the total length of $d \le 2h$ consecutive spacings, then

$$\mathbb{E}[W_d] = \frac{d}{n+1}, \tag{112}$$

and

$$\mathbb{E}[W_d^2] = \frac{d(d+1)}{(n+1)(n+2)}. \tag{113}$$

Therefore,

$$\mathbb{E}\left[F(I_{R,h}) \mid R\right] \le \frac{2h}{n+1}, \tag{114}$$

and

$$\mathbb{E}\left[F(I_{R,h})^2 \mid R\right] \le \frac{2h(2h+1)}{}(n+1)(n+2). \tag{115}$$

Taking expectation over R proves equation 107 and equation 108.

We now incorporate Assumption 5.1. Let

$$S := s(\widehat{f}(X), Y), \tag{116}$$

and let μ_x denote the conditional distribution of S given $X = x$, with μ denoting the marginal distribution of S. By Assumption 5.1,

$$\frac{\mathrm{d}\mu_x}{\mathrm{d}\mu} \le K \quad \mu\text{-a.e.}, \tag{117}$$

and therefore, for every Borel set $B \subseteq \mathbb{R}$,

$$\mu_x(B) \le K\mu(B) \tag{118}$$

for P_X-almost every x.

If R is conditionally independent of Y given X, then $I_{R,h}$ is conditionally independent of Y given X, and

$$\begin{aligned}\mathbb{P}\left(S \in I_{R,h} \mid X, R, T_1, \ldots, T_n\right) &= \mu_X(I_{R,h}) \\ &\le K\mu(I_{R,h}).\end{aligned} \tag{119}$$

Taking expectation and applying equation 107,

$$\mathbb{P}(S \in I_{R,h}) \le \frac{2Kh}{n+1} = \mathcal{O}(n^{-1}). \tag{120}$$

Similarly, let (X_1, Y_1) and (X_2, Y_2) be two independent observations and let R_1, R_2 be two ranks that are conditionally independent of (Y_1, Y_2) given (X_1, X_2) and jointly independent of the calibration sample. Writing

$$S_i := s(\widehat{f}(X_i), Y_i), \qquad i \in \{1, 2\}, \tag{121}$$

conditional independence gives

$$\begin{aligned}&\mathbb{P}\left(S_1 \in I_{R_1,h}, S_2 \in I_{R_2,h} \mid X_1, X_2, R_1, R_2, T_1, \ldots, T_n\right) \\ &\quad \le K^2 F(I_{R_1,h}) F(I_{R_2,h}).\end{aligned} \tag{122}$$

By Cauchy–Schwarz and equation 115,

$$\begin{aligned}&\mathbb{E}\left[F(I_{R_1,h})F(I_{R_2,h}) \mid R_1, R_2\right] \\ &\quad \le \sqrt{\mathbb{E}[F(I_{R_1,h})^2 \mid R_1]\mathbb{E}[F(I_{R_2,h})^2 \mid R_2]} \\ &\quad \le \frac{2h(2h+1)}{}(n+1)(n+2) = \mathcal{O}(n^{-2}).\end{aligned} \tag{123}$$

Consequently,

$$\mathbb{P}\left(S_1 \in I_{R_1,h}, S_2 \in I_{R_2,h}\right) = \mathcal{O}(n^{-2}). \tag{124}$$

Step 4: Bias of $\widehat{\Delta}_{\delta,n}$.

Recall that

$$\widehat{\Delta}_{\delta,n} = \frac{1}{n}\sum_{k=1}^{n} \Delta_{\delta,n}^{k}, \tag{125}$$

where

$$\Delta_{\delta,n}^{k} = I_k - A_k, \tag{126}$$

with

$$A_k := \alpha_\delta^k(X_k^a), \tag{127}$$

and

$$I_k := \mathbf{1}\left\{ s_k^{(a)} > \widehat{q}_c(A_k) \right\}, \qquad s_k^{(a)} := s(\widehat{f}(X_k^a), Y_k^a). \tag{128}$$

For $m \geq 1$, let $\alpha_{\delta,m}(X_0)$ be the adaptive level in equation 73, constructed using an independent approximation sample of size m. Let $\widehat{q}_c$ denote the conformal quantile constructed from the independent calibration sample $\mathcal{C}_c$ of size n, and define

$$D_{m,n} := \mathbf{1}\left\{ s_0 > \widehat{q}_c\left(\alpha_{\delta,m}(X_0)\right) \right\} - \alpha_{\delta,m}(X_0), \tag{129}$$

where

$$s_0 := s(\widehat{f}(X_0), Y_0). \tag{130}$$

By the i.i.d. assumption, for every $k \in [n]$, $\mathcal{C}_a^k$ has the same distribution as an independent approximation sample of size $n-1$, while (X_k^a, Y_k^a) has the distribution of the independent test pair. Therefore,

$$\mathbb{E}\left[\Delta_{\delta,n}^{k}\right] = \mathbb{E}[D_{n-1,n}], \tag{131}$$

and consequently,

$$\mathbb{E}\left[\widehat{\Delta}_{\delta,n}\right] = \mathbb{E}[D_{n-1,n}]. \tag{132}$$

On the other hand, the exact identity established earlier for $\Delta_{\delta,n}(\mathcal{C}_a)$ gives

$$\begin{aligned}\Delta_{\delta,n}(\mathcal{C}_a) = \mathbb{P}\left(s_0 > \widehat{q}_c\left(\alpha_{\delta,n}(X_0)\right) | \mathcal{C}_a\right) \\ - \mathbb{E}\left[\alpha_{\delta,n}(X_0) \mid \mathcal{C}_a\right].\end{aligned} \tag{133}$$

Taking expectation over $\mathcal{C}_a$ yields

$$\mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right] = \mathbb{E}[D_{n,n}]. \tag{134}$$

Thus, it remains to compare $\mathbb{E}[D_{n-1,n}]$ and $\mathbb{E}[D_{n,n}]$.

Couple the approximation samples of sizes n and $n-1$ by first drawing an i.i.d. sample of size n and then deleting one observation. By equation 82,

$$|\alpha_{\delta,n}(X_0) - \alpha_{\delta,n-1}(X_0)| \leq \frac{C_\alpha}{n} \qquad \text{a.s.} \tag{135}$$

Consequently, there exists a fixed integer $J < \infty$ such that

$$|r_n\left(\alpha_{\delta,n}(X_0)\right) - r_n\left(\alpha_{\delta,n-1}(X_0)\right)| \leq J \qquad \text{a.s.} \tag{136}$$

Let

$$s_{(1)}^{(c)} < \cdots < s_{(n)}^{(c)} \tag{137}$$

denote the calibration-score order statistics. The two indicators appearing in $D_{n,n}$ and $D_{n-1,n}$ can differ only when s_0 lies between two calibration order statistics whose ranks differ by at most J.

Consequently, for a suitable random rank R that is independent of $\mathcal{C}_c$ and conditionally independent of Y_0 given X_0,

$$\begin{aligned}&|\mathbf{1}\{s_0 > \widehat{q}_c(\alpha_{\delta,n}(X_0))\} - \mathbf{1}\{s_0 > \widehat{q}_c(\alpha_{\delta,n-1}(X_0))\}| \\ &\quad \le \mathbf{1}\{s_0 \in I_{R,J}\}.\end{aligned} \tag{138}$$

By equation 120,

$$\mathbb{P}(\text{the two indicators in equation 138 differ}) = \mathcal{O}(n^{-1}). \tag{139}$$

Combining equation 135 and equation 139,

$$|\mathbb{E}[D_{n,n}] - \mathbb{E}[D_{n-1,n}]| = \mathcal{O}(n^{-1}). \tag{140}$$

Finally, using equation 132 and equation 134,

$$\left|\mathbb{E}\left[\widehat{\Delta}_{\delta,n}\right] - \mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right]\right| = \mathcal{O}(n^{-1}). \tag{141}$$

Step 5: Mean-square stability of $\widehat{\Delta}_{\delta,n}$.

We next show that replacing any single observation in either $\mathcal{C}_a$ or $\mathcal{C}_c$ changes $\widehat{\Delta}_{\delta,n}$ by $\mathcal{O}(n^{-1})$ in mean square.

Write

$$Z_i^a := (X_i^a, Y_i^a), \qquad Z_j^c := (X_j^c, Y_j^c), \tag{142}$$

and regard the estimator as the functional

$$\widehat{\Delta}_{\delta,n} = \Psi(Z_1^a, \dots, Z_n^a, Z_1^c, \dots, Z_n^c). \tag{143}$$

Replacement of one approximation observation.

Fix $i \in [n]$, let $Z_i^{a\prime}$ be an independent copy of Z_i^a, and let Ψ_i^a denote the functional after replacing Z_i^a by $Z_i^{a\prime}$. Write A_k', I_k' for the corresponding quantities after the replacement.

For $k = i$, both the pseudo-test point and its associated adaptive level may change. Since $I_i, I_i' \in \{0, 1\}$ and $A_i, A_i' \in [\varepsilon, 1-\varepsilon]$,

$$|(I_i - A_i) - (I_i' - A_i')| \le 2. \tag{144}$$

For $k \neq i$, the pseudo-test observation (X_k^a, Y_k^a) is unchanged, and only one observation in the corresponding pseudo-calibration sample is replaced. Hence, by equation 84,

$$|A_k - A_k'| \le \frac{C_\alpha}{n-1}. \tag{145}$$

Since the corresponding ranks in $\mathcal{C}_c$ differ by at most a fixed constant J. Define

$$E_{ik}^a := \{I_k \neq I_k'\}, \qquad H_i^a := \sum_{k \neq i} \mathbf{1}\{E_{ik}^a\}. \tag{146}$$

Then

$$\begin{aligned}|\Psi - \Psi_i^a| &\le \frac{1}{n}\left(2 + \sum_{k \neq i} |A_k - A_k'| + H_i^a\right) \\ &\le \frac{1}{n}(2 + C_\alpha + H_i^a).\end{aligned} \tag{147}$$

We claim that

$$\mathbb{E}\left[(H_i^a)^2\right] = \mathcal{O}(1) \tag{148}$$

uniformly in i.

For any fixed $k \neq i$, both A_k and A'_k are constructed without using Y_k^a. Conditional on X_k^a and all remaining approximation data, the corresponding calibration ranks are therefore independent of Y_k^a. Since those ranks differ by at most J, equation 120 gives

$$\mathbb{P}(E_{ik}^a) = \mathcal{O}(n^{-1}). \tag{149}$$

To control the cross terms, take $k, \ell \neq i$ with $k \neq \ell$. The levels A_k and A_ℓ cannot directly be treated as mutually independent of Y_ℓ^a and Y_k^a, respectively, because each pseudo-calibration sample contains the other point. We therefore introduce leave-two-out reference levels.

Let $\bar{A}_k$ denote the adaptive level evaluated at X_k^a using the approximation sample with both Z_k^a and Z_ℓ^a removed, and define $\bar{A}_\ell$ analogously. By one deletion, one replacement, and another application of the stability bounds equation 82–equation 84, there exists a constant $C_* < \infty$ such that

$$|A_k - \bar{A}_k| \leq \frac{C_*}{n}, \qquad |A'_k - \bar{A}_k| \leq \frac{C_*}{n}, \tag{150}$$

and similarly,

$$|A_\ell - \bar{A}_\ell| \leq \frac{C_*}{n}, \qquad |A'_\ell - \bar{A}_\ell| \leq \frac{C_*}{n}. \tag{151}$$

Consequently, there exists a fixed integer $J_* < \infty$ such that

$$\begin{aligned} E_{ik}^a &\subseteq \left\{ s_k^{(a)} \in I_{\bar{R}_k, J_*} \right\}, \\ E_{i\ell}^a &\subseteq \left\{ s_\ell^{(a)} \in I_{\bar{R}_\ell, J_*} \right\}, \end{aligned} \tag{152}$$

where

$$\bar{R}_k := r_n(\bar{A}_k), \qquad \bar{R}_\ell := r_n(\bar{A}_\ell). \tag{153}$$

Crucially, conditional on X_k^a, X_ℓ^a and the approximation observations other than Z_k^a, Z_ℓ^a, the ranks $\bar{R}_k, \bar{R}_\ell$ do not depend on either Y_k^a or Y_ℓ^a. Since the two held-out responses are conditionally independent, the two-point bound equation 124 yields

$$\mathbb{P}\left(E_{ik}^a \cap E_{i\ell}^a\right) = \mathcal{O}(n^{-2}). \tag{154}$$

Expanding the square of H_i^a and using equation 149 and equation 154,

$$\begin{aligned} \mathbb{E}\left[(H_i^a)^2\right] &= \sum_{k \neq i} \mathbb{P}(E_{ik}^a) + 2 \sum_{\substack{k<\ell \\ k,\ell \neq i}} \mathbb{P}\left(E_{ik}^a \cap E_{i\ell}^a\right) \\ &= (n-1)\mathcal{O}(n^{-1}) + \mathcal{O}(n^2)\mathcal{O}(n^{-2}) \\ &= \mathcal{O}(1), \end{aligned} \tag{155}$$

which proves equation 148.

Using equation 147 and equation 148,

$$\mathbb{E}\left[(\Psi - \Psi_i^a)^2\right] = \mathcal{O}(n^{-2}) \qquad \text{uniformly in } i. \tag{156}$$

Replacement of one calibration observation.

Fix $j \in [n]$, let $Z_j^{c\prime}$ be an independent copy of Z_j^c, and let Ψ_j^c denote the resulting functional. The adaptive levels $A_1, \ldots, A_n$ do not change, since they depend only on $\mathcal{C}_a$.

For each k, let I'_k denote the indicator after replacing Z_j^c. Two calibration samples differing in one observation have empirical counting functions differing by at most one at every threshold. Their order statistics therefore interlace:

$$s_{(r-1)}^{(c)} \leq s_{(r)}^{(c)\prime} \leq s_{(r+1)}^{(c)} \tag{157}$$

for every admissible rank r, with the usual boundary modifications.

Define

$$E_{jk}^c := \{I_k \neq I_k'\}. \tag{158}$$

By equation 157, the event E_{jk}^c implies that $s_k^{(a)}$ lies within one calibration rank of the threshold associated with A_k. Let

$$H_j^c := \sum_{k=1}^{n} \mathbf{1}\left\{E_{jk}^c\right\}. \tag{159}$$

Since the adaptive levels themselves are unchanged,

$$|\Psi - \Psi_j^c| \leq \frac{H_j^c}{n}. \tag{160}$$

For each fixed k, A_k is constructed without using Y_k^a. Therefore, the one-point rank-interval bound gives

$$\mathbb{P}(E_{jk}^c) = \mathcal{O}(n^{-1}). \tag{161}$$

For $k \neq \ell$, introduce the same leave-two-out reference levels $\bar{A}_k, \bar{A}_\ell$. Deleting the ℓ-th approximation observation changes A_k by only $\mathcal{O}(n^{-1})$, and deleting the k-th approximation observation changes A_ℓ by only $\mathcal{O}(n^{-1})$. Hence, there exists a fixed $J_{**} < \infty$ such that

$$\begin{aligned} E_{jk}^c &\subseteq \left\{s_k^{(a)} \in I_{\bar{R}_k, J_{**}}\right\}, \\ E_{j\ell}^c &\subseteq \left\{s_\ell^{(a)} \in I_{\bar{R}_\ell, J_{**}}\right\}. \end{aligned} \tag{162}$$

The leave-two-out ranks are independent of (Y_k^a, Y_ℓ^a) conditional on (X_k^a, X_ℓ^a) and the remaining approximation data. Therefore, by equation 124,

$$\mathbb{P}\left(E_{jk}^c \cap E_{j\ell}^c\right) = \mathcal{O}(n^{-2}). \tag{163}$$

It follows that

$$\begin{aligned} \mathbb{E}\left[(H_j^c)^2\right] &= \sum_{k=1}^{n} \mathbb{P}(E_{jk}^c) + 2\sum_{k<\ell} \mathbb{P}\left(E_{jk}^c \cap E_{j\ell}^c\right) \\ &= n\mathcal{O}(n^{-1}) + \mathcal{O}(n^2)\mathcal{O}(n^{-2}) \\ &= \mathcal{O}(1). \end{aligned} \tag{164}$$

Combining equation 160 and equation 164,

$$\mathbb{E}\left[(\Psi - \Psi_j^c)^2\right] = \mathcal{O}(n^{-2}) \qquad \text{uniformly in } j. \tag{165}$$

Step 6: Concentration of $\widehat{\Delta}_{\delta,n}$ around its expectation.

The functional $\Psi = \widehat{\Delta}_{\delta,n}$ depends on the $2n$ independent observations

$$Z_1^a, \ldots, Z_n^a, Z_1^c, \ldots, Z_n^c. \tag{166}$$

Applying the Efron–Stein inequality and using equation 156 and equation 165,

$$\begin{aligned} \mathrm{Var}\left(\widehat{\Delta}_{\delta,n}\right) &\leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[(\Psi - \Psi_i^a)^2\right] + \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}\left[(\Psi - \Psi_j^c)^2\right] \\ &= \mathcal{O}(n^{-1}). \end{aligned} \tag{167}$$

Chebyshev's inequality therefore gives

$$\widehat{\Delta}_{\delta,n} - \mathbb{E}\left[\widehat{\Delta}_{\delta,n}\right] = \mathcal{O}_P(n^{-1/2}). \tag{168}$$

Combining the centered fluctuation equation 168 with the bias bound equation 141,

$$\begin{aligned}
&\widehat{\Delta}_{\delta,n} - \mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right] \\
&\quad = \left(\widehat{\Delta}_{\delta,n} - \mathbb{E}\left[\widehat{\Delta}_{\delta,n}\right]\right) + \left(\mathbb{E}\left[\widehat{\Delta}_{\delta,n}\right] - \mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right]\right) \\
&\quad = \mathcal{O}_P(n^{-1/2}) + \mathcal{O}(n^{-1}) \\
&\quad = \mathcal{O}_P(n^{-1/2}).
\end{aligned} \tag{169}$$

Thus,

$$\left|\widehat{\Delta}_{\delta,n} - \mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right]\right| = \mathcal{O}_P(n^{-1/2}). \tag{170}$$

Step 7: Convergence of the combined estimator.

Finally, combining equation 103 and equation 170, we obtain

$$\begin{aligned}
&\left|\hat{\alpha}_\delta + \widehat{\Delta}_{\delta,n} - \mathbb{E}\left[\alpha_\delta(X_0) + \Delta_{\delta,n}(\mathcal{C}_a)\right]\right| \\
&\quad \le \left|\hat{\alpha}_\delta - \mathbb{E}[\alpha_\delta(X_0)]\right| + \left|\widehat{\Delta}_{\delta,n} - \mathbb{E}\left[\Delta_{\delta,n}(\mathcal{C}_a)\right]\right| \\
&\quad = \mathcal{O}_P(n^{-1/2}) + \mathcal{O}_P(n^{-1/2}) \\
&\quad = \mathcal{O}_P(n^{-1/2}).
\end{aligned} \tag{171}$$

This proves all three claims. □

B.4 Proof of LOO Estimator Stability - Theorem 5.3

Proof. Both variance bounds follow from the Efron–Stein inequality together with the stability properties established in the proof of Theorem 5.2. We first recall that, for any measurable function f of independent random variables $Z_1, \dots, Z_m$, if Z_i' is an independent copy of Z_i and

$$f_i' := f(Z_1, \dots, Z_{i-1}, Z_i', Z_{i+1}, \dots, Z_m), \tag{172}$$

then the Efron–Stein inequality gives

$$\mathrm{Var}\left(f(Z_1, \dots, Z_m)\right) \le \frac{1}{2}\sum_{i=1}^{m} \mathbb{E}\left[\left(f(Z_1, \dots, Z_m) - f_i'\right)^2\right]. \tag{173}$$

Variance of $\hat{\alpha}_\delta$.

Recall from the proof of Theorem 5.2 that

$$\hat{\alpha}_\delta = \Phi(Z_1, \dots, Z_n), \qquad Z_i := (X_i^a, Y_i^a), \tag{174}$$

where $Z_1, \dots, Z_n$ are independent under Assumption 4.1. For each $i \in [n]$, let Z_i' be an independent copy of Z_i, and define

$$\Phi_i' := \Phi\left(Z_1, \dots, Z_{i-1}, Z_i', Z_{i+1}, \dots, Z_n\right). \tag{175}$$

The bounded-difference argument established in the proof of Theorem 5.2 shows that, for every $i \in [n]$,

$$|\Phi - \Phi_i'| \le \frac{1 + C_\alpha}{n} \qquad \text{almost surely.} \tag{176}$$

In particular,

$$\mathbb{E}\left[(\Phi - \Phi_i')^2\right] \le \left(\frac{1 + C_\alpha}{n}\right)^2. \tag{177}$$

Applying equation 173 to the functional Φ and using equation 177, we obtain

$$\begin{aligned}\mathrm{Var}\left(\hat{\alpha}_\delta\right) &= \mathrm{Var}(\Phi) \\ &\leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[(\Phi-\Phi_i')^2\right] \\ &\leq \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1+C_\alpha}{n}\right)^2 \\ &= \frac{(1+C_\alpha)^2}{2n} \\ &= \mathcal{O}(n^{-1}).\end{aligned} \tag{178}$$

Therefore,

$$\mathrm{Var}\left(\hat{\alpha}_\delta\right) = \mathcal{O}(n^{-1}). \tag{179}$$

Variance of $\widehat{\Delta}_{\delta,n}$.

Recall also from the proof of Theorem 5.2 that $\widehat{\Delta}_{\delta,n}$ can be written as

$$\widehat{\Delta}_{\delta,n} = \Psi\left(Z_1^a,\ldots,Z_n^a,Z_1^c,\ldots,Z_n^c\right), \tag{180}$$

where

$$Z_i^a := (X_i^a, Y_i^a), \qquad Z_j^c := (X_j^c, Y_j^c), \tag{181}$$

and the $2n$ observations in equation 180 are independent under Assumptions 4.1 and 5.1.

For each $i \in [n]$, let Ψ_i^a denote the value of the functional after replacing Z_i^a by an independent copy. Similarly, for each $j \in [n]$, let Ψ_j^c denote the value obtained after replacing Z_j^c by an independent copy.

The mean-square stability argument established in the proof of Theorem 5.2 gives, uniformly over $i, j \in [n]$,

$$\mathbb{E}\left[(\Psi-\Psi_i^a)^2\right] = \mathcal{O}(n^{-2}), \qquad \mathbb{E}\left[(\Psi-\Psi_j^c)^2\right] = \mathcal{O}(n^{-2}). \tag{182}$$

Applying the Efron–Stein inequality equation 173 to all $2n$ independent coordinates of Ψ, and then using equation 182, gives

$$\begin{aligned}\mathrm{Var}\left(\widehat{\Delta}_{\delta,n}\right) &= \mathrm{Var}(\Psi) \\ &\leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[(\Psi-\Psi_i^a)^2\right] + \frac{1}{2}\sum_{j=1}^{n}\mathbb{E}\left[(\Psi-\Psi_j^c)^2\right] \\ &= \frac{1}{2}\left(n\,\mathcal{O}(n^{-2}) + n\,\mathcal{O}(n^{-2})\right) \\ &= \mathcal{O}(n^{-1}).\end{aligned} \tag{183}$$

Hence,

$$\mathrm{Var}\left(\widehat{\Delta}_{\delta,n}\right) = \mathcal{O}(n^{-1}). \tag{184}$$

Variance of the combined estimator.

Finally, although $\hat{\alpha}_\delta$ and $\widehat{\Delta}_{\delta,n}$ are not independent, the two variance bounds above are sufficient to control the variance of their sum. Indeed,

$$\begin{aligned}\mathrm{Var}\left(\hat{\alpha}_\delta + \widehat{\Delta}_{\delta,n}\right) = \mathrm{Var}\left(\hat{\alpha}_\delta\right) + \mathrm{Var}\left(\widehat{\Delta}_{\delta,n}\right) \\ + 2\,\mathrm{Cov}\left(\hat{\alpha}_\delta, \widehat{\Delta}_{\delta,n}\right).\end{aligned} \tag{185}$$

By the Cauchy–Schwarz inequality,

$$\left|\mathrm{Cov}\left(\hat{\alpha}_\delta, \widehat{\Delta}_{\delta,n}\right)\right| \le \sqrt{\mathrm{Var}\left(\hat{\alpha}_\delta\right)\mathrm{Var}\left(\widehat{\Delta}_{\delta,n}\right)} = \mathcal{O}(n^{-1}), \tag{186}$$

where the last equality follows from equation 179 and equation 184. Substituting equation 179, equation 184, and equation 186 into equation 185, we conclude that

$$\mathrm{Var}\left(\hat{\alpha}_\delta + \widehat{\Delta}_{\delta,n}\right) = \mathcal{O}(n^{-1}). \tag{187}$$

This proves all three claims. □

C Further Discussion and Analysis

C.1 Comparison between eCP and sCP for the Tunable Prediction Set Size problem

In this section, we provide a more detailed discussion of conformal e-value prediction (eCP) and Split Conformal Prediction (sCP) in the context of the TPSS problem.

C.1.1 Why eCP admits a more direct extension to TPSS than sCP

A key advantage of the eCP framework is its ability to naturally handle data-dependent miscoverage levels. This flexibility arises from the post-hoc validity property of eCP, which ensures that the coverage level can be chosen adaptively based on the observed data while still retaining a valid guarantee.

More precisely, if $\tilde{\alpha}$ is any (possibly data-dependent) coverage level, then eCP yields the control (Gauthier et al., 2025b).

$$\mathbb{E}\left[\frac{\mathbb{P}\left(Y_0 \notin \hat{\mathcal{C}}_{\tilde{\alpha}}(X_0)\middle|\tilde{\alpha}\right)}{\tilde{\alpha}}\right] \le 1. \tag{188}$$

Importantly, this guarantee holds without requiring $\tilde{\alpha}$ to be independent of the data. As a result, one can directly define a data-dependent level $\tilde{\alpha}(X_0)$ to enforce a size constraint, for example via

$$\tilde{\alpha}(X_0) := \inf\left\{\alpha \in (0,1) : \left|\left\{y : E(X_0, y) < \frac{1}{\alpha}\right\}\right| \le \mathcal{S}(X_0)\right\}, \tag{189}$$

without breaking the validity of the procedure. In this sense, the extension of eCP to the TPSS setting is immediate at the level of construction.

To obtain a more interpretable coverage statement, one may further pass from the post-hoc guarantee to a marginal form. Starting from

$$\mathbb{E}\left[\frac{\mathbb{P}\left(Y_0 \notin \hat{\mathcal{C}}_{\tilde{\alpha}}(X_0)\middle|\tilde{\alpha}\right)}{\tilde{\alpha}}\right] \le 1, \tag{190}$$

one can heuristically rewrite

$$\mathbb{P}\big(Y_0 \notin \hat{\mathcal{C}}_{\tilde{\alpha}}(X_0)\big) = \mathbb{E}\left[\tilde{\alpha} \cdot \frac{\mathbb{P}\left(Y_0 \notin \hat{\mathcal{C}}_{\tilde{\alpha}}(X_0)\middle|\tilde{\alpha}\right)}{\tilde{\alpha}}\right]. \tag{191}$$

Since the ratio inside the expectation is controlled and typically close to 1, a first-order Taylor expansion around this value yields

$$\mathbb{P}\big(Y_0 \notin \hat{\mathcal{C}}_{\tilde{\alpha}}(X_0)\big) \approx \mathbb{E}[\tilde{\alpha}], \tag{192}$$

and therefore

$$\mathbb{P}\big(Y_0 \in \hat{\mathcal{C}}_{\tilde{\alpha}}(X_0)\big) \approx 1 - \mathbb{E}[\tilde{\alpha}(X_0)]. \tag{193}$$

Thus, handling data-dependent miscoverage in eCP only requires this additional Taylor approximation step, while the underlying construction remains unchanged.

In contrast, such a direct extension is not available in the classical sCP framework. In sCP, the miscoverage level α determines the empirical quantile threshold and hence directly controls the size of the prediction set. This naturally suggests defining

$$\mathcal{C}_\alpha(X_0) = \{y \in \mathcal{Y} : s(\hat{f}(X_0), y) \le \hat{q}_\alpha\}, \tag{194}$$

and then selecting a data-dependent level

$$\tilde{\alpha}(X_0) := \inf\{\alpha \in (0,1) : |\mathcal{C}_\alpha(X_0)| \le \mathcal{S}(X_0)\} \tag{195}$$

to enforce the desired size constraint.

However, this approach breaks the fundamental rank argument underlying sCP. Let $s_1, \dots, s_n$ denote the calibration scores and let s_0 denote the test score. Under exchangeability, the rank

$$R := 1 + \sum_{i=1}^{n} \mathbf{1}\{s_i \le s_0\} \tag{196}$$

is uniformly distributed over $\{1, \dots, n+1\}$. Consequently, for any fixed $\alpha \in (0,1)$,

$$\mathbb{P}\big(R > \lceil (n+1)(1-\alpha) \rceil\big) = \frac{n+1-\lceil (n+1)(1-\alpha) \rceil}{n+1} \le \alpha, \tag{197}$$

which gives the standard finite-sample coverage guarantee

$$\mathbb{P}\big(Y_0 \notin \mathcal{C}_\alpha(X_0)\big) \le \alpha. \tag{198}$$

The difficulty appears when α is replaced by the adaptive choice $\tilde{\alpha}(X_0)$. In that case,

$$Y_0 \notin \mathcal{C}_{\tilde{\alpha}(X_0)}(X_0) \quad \Longleftrightarrow \quad R > \lceil (n+1)(1-\tilde{\alpha}(X_0)) \rceil. \tag{199}$$

Unlike the fixed-α setting, the cutoff $\lceil (n+1)(1-\tilde{\alpha}(X_0)) \rceil$ is now itself a random quantity depending on X_0. As a result, the event above involves a data-dependent threshold, and the uniformity of R alone is no longer sufficient to evaluate its probability. Intuitively, the cutoff is now chosen adaptively based on the test input, which introduces dependence between the rank and the threshold and destroys the symmetry exploited by the classical exchangeability argument. Therefore, while eCP can be adapted to TPSS rather directly through its post-hoc validity framework, the analogous extension in sCP is fundamentally non-trivial. This is precisely the difficulty addressed by our proposed construction.

C.1.2 Advantages of TsCP over eCP-Based TPSS Methods

Although eCP extends more directly to the TPSS setting, its resulting coverage statement is less direct than the guarantee obtained from our sCP-based construction. In particular, eCP naturally accommodates data-dependent levels, but the guarantee obtained in prior TPSS extensions is not immediately expressed in the standard marginal coverage form. To obtain an interpretable statement of the form $1 - \mathbb{E}[\tilde{\alpha}(X_0)]$, one must further apply a first-order Taylor approximation. Thus, although eCP provides a flexible mechanism for enforcing size constraints, the final coverage interpretation is indirect and approximation-based.

In contrast, our sCP-based construction yields a direct marginal guarantee. By using one calibration split to select the data-dependent level and an independent split for conformal calibration, we obtain

$$\mathbb{P}\big(Y_0 \in \mathcal{C}_\delta(X_0)\big) \ge 1 - \mathbb{E}\left[\alpha_\delta(X_0) + \Delta_{\delta,n}(\mathcal{C}_a)\right], \tag{200}$$

without relying on a Taylor approximation. Moreover, the soft size-control result in Theorem 4.4 shows that the resulting prediction set asymptotically satisfies the desired size constraint with high probability. Hence, in the TPSS setting, the sCP construction provides both a more direct coverage interpretation and explicit asymptotic control of the prediction set size.

A second advantage appears in the analysis of the leave-one-out estimator. In TsCP, the LOO functional enjoys a strong locality property: replacing a single observation perturbs the estimator by at most order $(1 + C_0)/n$ almost surely. This is precisely the content of the bounded-difference argument established in Step 2 of Section B.3, where one shows that the corresponding symmetric functional changes by at most

$$\frac{1 + C_0}{n} \tag{201}$$

when one data point is replaced. More fundamentally, this follows from the rank-based structure of sCP: deleting or replacing a single calibration observation changes the relevant empirical quantile rank by at most one, and hence changes the corresponding adaptive level by only $\mathcal{O}(n^{-1})$. As a consequence, the expectation of the LOO estimator can be related directly to the target quantity. In particular, by exchangeability,

$$\mathbb{E}[\hat{\alpha}_\delta] = \mathbb{E}[\alpha_{\delta,n-1}(X_0)], \tag{202}$$

while the one-point stability of the adaptive level yields

$$|\mathbb{E}[\alpha_{\delta,n-1}(X_0)] - \mathbb{E}[\alpha_{\delta,n}(X_0)]| = \mathcal{O}(n^{-1}). \tag{203}$$

Thus, the bias incurred by the leave-one-out construction is controlled directly through the local rank stability of sCP. The same structural property also gives the bounded-difference condition needed for concentration and, through the Efron–Stein inequality, the $\mathcal{O}(n^{-1})$ variance bound. Intuitively, because the sCP construction depends only on empirical quantiles and rank comparisons, the influence of any individual observation remains localized and can be controlled explicitly.

By contrast, obtaining analogous LOO results in the eCP literature requires substantially stronger assumptions. In addition to i.i.d. sampling and the no-ties condition, their analysis relies on a structural constraint on the label space (P1) and a global stability assumption (P3), both of which can be restrictive in practice. These additional assumptions are needed because the corresponding eCP quantities do not exhibit the same direct rank-locality property as sCP.

The underlying difficulty stems from the non-local nature of the eCP construction. In eCP, the prediction rule is built from quantities of the form

$$E = \frac{S(X_{n+1}, Y_{n+1})}{\frac{1}{n+1}\sum_{i=0}^{n} S(X_i, Y_i)}, \tag{204}$$

where the denominator aggregates contributions from all observations. When a single data point is replaced, the denominator changes globally, affecting the value of E in a non-local manner. In particular, even a small perturbation of the empirical average propagates through the ratio and can affect the resulting prediction rule globally. This makes it substantially less direct to obtain the one-point $\mathcal{O}(n^{-1})$ stability that arises naturally from the rank structure of sCP, and consequently complicates the control of the corresponding leave-one-out expectation.

C.2 Effect of the Truncation Parameter ε

In this section, we analyze the role of the truncation parameter ε introduced in equation 6 in Section 4. To motivate its necessity, recall that when adapting the eCP framework to the TPSS problem (see Section 3), one may define a data-dependent miscoverage level as

$$\tilde{\alpha}(X_0) := \inf\left\{\alpha \in (0,1) : \left|\left\{y : E(X_0, y) < \frac{1}{\alpha}\right\}\right| \leq \mathcal{S}(X_0)\right\}. \tag{205}$$

Both equation 205 and equation 6 pursue the same objective: selecting a data-dependent level α to enforce a size constraint on the resulting prediction set. The key distinction, however, lies in the admissible range of α. In equation 205, the optimization is performed over the full interval $(0, 1)$, whereas in equation 6 it is restricted to the truncated interval $[\varepsilon, 1 - \varepsilon]$. At first glance, this restriction may appear to limit the flexibility of our proposed procedure. In what follows, we show that this truncation is in fact essential for the validity of the sCP-based construction, while having a negligible impact on its practical expressiveness.

C.2.1 WHY ε IS NECESSARY

The introduction of the truncation parameter ε is primarily required for the proof of the soft size control guarantee in Theorem 4.4. More precisely, as shown in the proof in Section B.2, the argument relies on comparing the two empirical quantile functions $\hat{q}_a(\alpha)$ and $\hat{q}_c(\alpha)$ uniformly over all admissible values of α. This is exactly the content of Lemma B.1, which establishes that

$$\sup_{\alpha\in[\varepsilon,1-\varepsilon]} |\hat{q}_a(\alpha) - \hat{q}_c(\alpha)| \to 0 \qquad \text{in probability.} \tag{206}$$

This uniform control is the key step in the proof. Indeed, by construction of $\alpha_\delta(X_0)$, the selected level is random and depends on the test input X_0. Therefore, in order to control the difference

$$|\hat{q}_c(\alpha_\delta(X_0)) - \hat{q}_a(\alpha_\delta(X_0))|\,, \tag{207}$$

it is not sufficient to know pointwise convergence for each fixed α. Instead, one must control the quantile difference uniformly over all possible values that $\alpha_\delta(X_0)$ may take. This is why the proof passes through the bound

$$|\hat{q}_c(\alpha_\delta(X_0)) - \hat{q}_a(\alpha_\delta(X_0))| \le \sup_{\alpha\in[\varepsilon,1-\varepsilon]} |\hat{q}_c(\alpha) - \hat{q}_a(\alpha)|\,, \tag{208}$$

which is valid only because $\alpha_\delta(X_0)$ is constrained to lie in the interval $[\varepsilon, 1-\varepsilon]$.

The role of ε is therefore to restrict the optimization to a fixed compact subinterval strictly contained in $(0,1)$. On such an interval, the quantile levels are bounded away from the boundary points 0 and 1, and the empirical quantile functions admit uniform consistency under the continuity of the population quantile function on $[\varepsilon, 1-\varepsilon]$.

If the truncation were removed and one allowed $\alpha_\delta(X_0)$ to range over the full interval $(0,1)$, then the above argument would break down. The reason is that one would need to control

$$\sup_{\alpha\in(0,1)} |\hat{q}_a(\alpha) - \hat{q}_c(\alpha)|\,, \tag{209}$$

that is, uniform convergence of the empirical quantile functions all the way up to the boundary points 0 and 1. Near these endpoints, empirical quantiles become unstable, as they correspond to extreme order statistics. Small perturbations in the empirical distribution can produce disproportionately large changes in the associated quantiles, and the continuity argument used in Lemma B.1 is no longer available in the same uniform form. As a result, the step that transfers the size constraint from the approximation set to the final prediction set can no longer be justified.

C.2.2 WHY ε DOES NOT WEAKEN THE PROPOSED METHOD

The introduction of ε serves only to ensure that $\alpha_\delta(X_0)$ lies in a compact subset of $(0,1)$, and has essentially no impact in practical settings. In particular, ε can be chosen arbitrarily small (e.g., 10^{-10}, 10^{-100}, or even smaller), far below any resolution that is meaningful at finite sample sizes. From an empirical standpoint, such extremely small values are indistinguishable from 0. Indeed, in finite-sample settings, the effective resolution of α is inherently limited by the sample size n, since quantile levels are realized through ranks of the form $\lceil (n+1)(1-\alpha) \rceil$. Consequently, any variation of α below the scale of $1/n$ has no observable effect on the resulting prediction set. Choosing $\varepsilon \ll 1/n$ therefore does not alter the behavior of the method in practice. This phenomenon is common in statistical procedures: theoretical constraints are often imposed to exclude pathological boundary cases, while having no measurable impact on empirical performance. In our setting, the truncation removes only extreme regimes of α that would correspond to degenerate prediction sets, while leaving the practically relevant range entirely unaffected. In summary, although ε is required for theoretical stability, it can be taken sufficiently small so that its presence is empirically negligible.

C.3 DISCUSSION OF ASSUMPTION 5.1

In Section 5, we introduced a leave-one-out procedure to obtain a practical approximation of the expected-value terms appearing in the coverage guarantee of Theorem 4.2. The theoretical guarantees established for this estimator, including its consistency and variance decay, rely on Assumption 5.1. In

this section, we discuss the role and interpretation of this assumption and compare it with assumptions imposed in the existing literature. We begin by recalling the assumption.

Assumption 5.1. Let μ denote the marginal distribution of $s(\hat{f}(X), Y)$, and let μ_x denote its conditional distribution given $X = x$. Assume that there exists a constant $K < \infty$ such that, for P_X-almost every $x \in \mathcal{X}$,

$$\mu_x \ll \mu, \qquad \frac{d\mu_x}{d\mu} \leq K \quad \mu\text{-a.e.}$$

Interpretation of Assumption 5.1. Assumption 5.1 controls how different the conditional distribution of the nonconformity score can be from its marginal distribution. Recall that μ is the marginal distribution of the score $S = s(\hat{f}(X), Y)$, whereas μ_x is the distribution of the same score conditional on $X = x$. The condition

$$\mu_x \ll \mu, \qquad \frac{d\mu_x}{d\mu} \leq K, \tag{210}$$

implies, equivalently, that for every measurable set B in the score space,

$$\mu_x(B) \leq K\mu(B) \tag{211}$$

for P_X-almost every x. Thus, a region of the score space that has small probability under the marginal score distribution must also have small conditional probability for almost every covariate value x, up to the multiplicative factor K. Intuitively, the assumption prevents the conditional score distribution at a particular x from concentrating a large amount of probability on a region that is extremely unlikely under the overall score distribution.

When the corresponding densities exist, Assumption 5.1 can be viewed more directly as a bounded density-ratio condition. If f and g_x denote the marginal and conditional densities of the score, respectively, then the assumption requires

$$\frac{g_x(t)}{f(t)} \leq K \tag{212}$$

for μ-almost every t and P_X-almost every x. Importantly, this condition is imposed only on the one-dimensional distribution of the nonconformity score rather than directly on the full distribution of (X, Y). It therefore controls the amount of heterogeneity that remains in the score distribution after conditioning on the covariates.

The role of Assumption 5.1 in our analysis is closely connected to the rank-based structure of sCP. In the proof of Theorem 5.2, replacing or deleting a single observation changes the relevant empirical quantile by only a bounded number of ranks. Consequently, whether the associated coverage indicator changes is determined by whether the score of the held-out observation falls inside an interval bounded by a fixed number of neighboring calibration order statistics. Under the marginal score distribution, such an interval has probability of order $\mathcal{O}(n^{-1})$; moreover, the second-moment argument used in the proof gives an order $\mathcal{O}(n^{-2})$ bound for the corresponding two-point event. Assumption 5.1 allows these marginal probability bounds to be transferred to the conditional score distribution through equation 211. In particular, if I denotes such a bounded-rank interval, then

$$\mathbb{P}(S \in I \mid X = x) = \mu_x(I) \leq K\mu(I). \tag{213}$$

This is the key step that ensures that a small perturbation of the empirical quantile changes the relevant coverage indicator only with sufficiently small probability. We emphasize that this additional assumption is not needed for the convergence analysis of $\hat{\alpha}_\delta$ itself, for which the rank stability and i.i.d. structure are sufficient; it enters when controlling the coverage-related component of the LOO estimator.

Assumption 5.1 is satisfied in several natural settings. For example, if the nonconformity score is independent of X, then

$$\mu_x = \mu \tag{214}$$

for P_X-almost every x, and the assumption holds with $K = 1$. More generally, it holds whenever the conditional score densities are uniformly dominated by the marginal score density with a bounded

likelihood ratio. A simple example arises when X belongs to finitely many groups. Suppose that $X \in \{1, \dots, M\}$ and that

$$p_j := \mathbb{P}(X = j) > 0, \qquad j \in [M]. \tag{215}$$

Since the marginal score distribution satisfies

$$\mu = \sum_{j=1}^{M} p_j \mu_j, \tag{216}$$

we have, for every measurable set B,

$$\mu(B) \geq p_j \mu_j(B), \tag{217}$$

and hence

$$\mu_j(B) \leq \frac{1}{p_j} \mu(B). \tag{218}$$

Thus, if $\min_j p_j \geq p_{\min} > 0$, Assumption 5.1 holds with $K \leq p_{\min}^{-1}$. More broadly, the condition is reasonable whenever the conditional score distributions vary with x but do not become arbitrarily more concentrated than their marginal mixture.

Nevertheless, Assumption 5.1 should not be regarded as a mild assumption in complete generality. It requires a *uniform* bound on the conditional-to-marginal score distribution ratio, and this requirement can fail when certain covariate regions induce highly concentrated score distributions in portions of the score space that have very small marginal probability. For instance, if the score becomes nearly deterministic for some values of X, or if rare covariate regions induce substantially different score distributions, a finite uniform constant K may not exist. We therefore do not claim that Assumption 5.1 is innocuous or universally satisfied.

Comparison with prior methods. While Assumption 5.1 is not mild, we *emphasize* that nontrivial regularity assumptions are also required by existing approaches to establish the consistency of a leave-one-out approximation. The closest related analysis is the eCP-based LOO estimator developed in Backward Conformal Prediction and its subsequent extensions (Gauthier et al., 2025a). Their convergence result relies on several additional regularity conditions on the transformation, the induced leave-one-out quantities, and the size-control rule. To make the comparison explicit, we briefly recall the main assumptions required in their analysis.

First, the transformation $h(s; D, X)$ is assumed to be uniformly bounded as

$$0 < h_{\min} \leq h(s; D, X) \leq h_{\max} < \infty, \tag{219}$$

and the calibration size is required to satisfy $n > h_{\max}/h_{\min}$. In addition, the transformation h and the size-constraint rule must satisfy the following three properties almost surely.

(P1) Concentration. For every $y \in \mathcal{Y}$, the sequences

$$\{h_j\}_{j=1}^n, \qquad \{h_i^y\}_{i=1}^n, \qquad \{h_j(y)\}_{j=1}^n \tag{220}$$

are assumed to satisfy Chernoff–Hoeffding concentration bounds. More precisely, for every $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} h_i - \mathbb{E}[h_i]\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(h_{\max} - h_{\min})^2}\right), \tag{221}$$

and, uniformly over $y \in \mathcal{Y}$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} h_i^y - \mathbb{E}[h_i^y]\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(h_{\max} - h_{\min})^2}\right), \tag{222}$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} h_i(y) - \mathbb{E}[h_i(y)]\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(h_{\max} - h_{\min})^2}\right). \tag{223}$$

(P2) Expectation stability. For every $y \in \mathcal{Y}$, they require

$$|\mathbb{E}[h_i] - \mathbb{E}[h_i^y]| \leq \frac{h_{\max}}{\sqrt{n}}, \tag{224}$$

$$|\mathbb{E}[h_i(y)] - \mathbb{E}[h_{n+1}(y)]| \leq \frac{h_{\max}}{\sqrt{n}}. \tag{225}$$

(P3) Global stability. There must exist a constant $L > 0$ such that

$$\left| \sum_{i=1}^{n} (\widetilde{\alpha}_i(h) - \mathbb{E}[\widetilde{\alpha}_{n+1}(h)]) \right| \leq L \max_{y \in \mathcal{Y}} \left| \sum_{i=1}^{n} \left(E^i(y,h) - E^{n+1}(y,h) \right) \right|. \tag{226}$$

These assumptions play different roles in controlling the dependence induced by the leave-one-out construction. We next explain why, taken together, they constitute substantial regularity requirements and can be restrictive in practice.

Property (P1) requires, uniformly over $y \in \mathcal{Y}$, several sequences arising from the LOO construction to satisfy Chernoff–Hoeffding concentration bounds. This is a strong requirement because these quantities are themselves constructed from highly overlapping leave-one-out datasets and are therefore generally dependent; such concentration does not follow directly from exchangeability or i.i.d. sampling. Moreover, the requirement is imposed for every candidate label y, making it particularly restrictive when the label space is large or continuous. Property (P2) further requires the expectations of several leave-one-out, label-perturbed, and test-point quantities to differ by at most order $\mathcal{O}(n^{-1/2})$. This amounts to an additional stability requirement connecting the finite-sample leave-one-out construction to its test-point counterpart and, again, is not implied by the basic sampling assumption. Finally, Property (P3) imposes a global stability condition linking perturbations of the data-dependent miscoverage levels to perturbations of the corresponding e-value vectors. Informally, it requires the mapping from the collection of e-values to the selected miscoverage level to behave in a uniformly Lipschitz manner. Such a condition can be difficult to verify for general size-constraint rules, particularly when the rule involves discrete thresholding or optimization over the prediction-set size.

We note that the comparison between the set of assumptions imposed by Gauthier et al. (2025a) and Assumption 5.1 is not entirely one-sided. The two frameworks impose regularity on fundamentally different objects, so it is not straightforward to state that one set of assumptions is strictly stronger than the other. The Backward CP assumptions are primarily *procedural*: they require concentration of several leave-one-out quantities, stability of their expectations, and a global relationship between perturbations of the e-value representation and the resulting data-dependent miscoverage level. By contrast, Assumption 5.1 is purely *distributional*: it places no direct restriction on the LOO mapping or the size-control rule, but instead requires the conditional score distribution μ_x to be uniformly dominated by the marginal score distribution μ. Thus, the two analyses obtain LOO consistency through rather different forms of regularity.

From a practical perspective, we view Assumption 5.1 as somewhat easier to interpret and potentially more amenable to control. The assumptions in Backward CP depend on the global behavior of the complete leave-one-out procedure and on quantities induced jointly by the data, the transformation, and the size-control rule. Verifying such properties can therefore become difficult as the underlying procedure becomes more complicated. In contrast, Assumption 5.1 concerns only the one-dimensional nonconformity score distribution. Moreover, unlike the underlying data-generating distribution, the choice of nonconformity score is part of the design of the conformal procedure. Hence, at least in principle, one may select or regularize the score so that its conditional distributions are less heterogeneous across x, making Assumption 5.1 more plausible in practice.

We emphasize, however, that this does not make Assumption 5.1 mild. The required uniform density-ratio bound can still fail in important settings, just as the concentration and stability conditions required by Backward CP can fail. Rather, the two approaches trade different forms of regularity, and our framework expresses this regularity through a more localized condition on the score distribution. Relaxing this requirement, for example to a local or average domination condition around the relevant conformal quantiles, is therefore an important and interesting direction for future work.

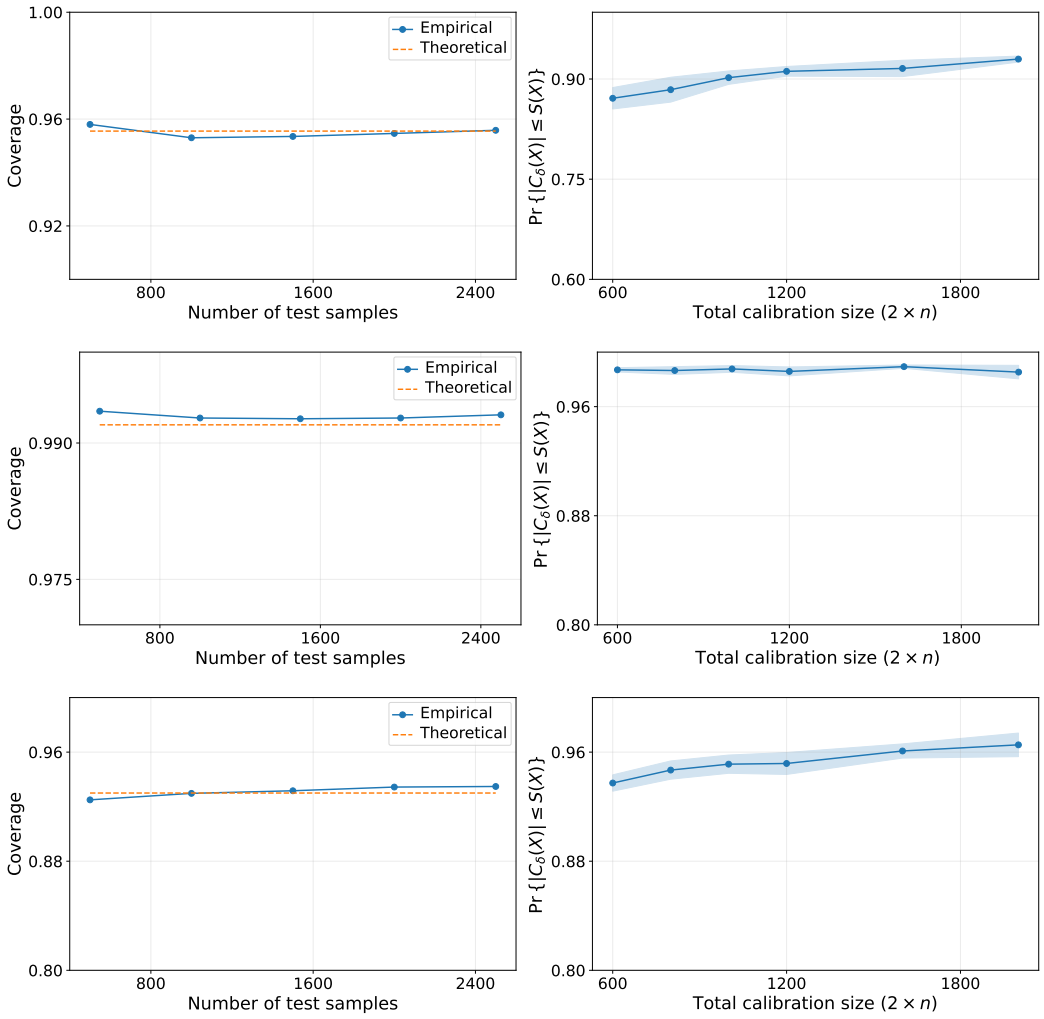


Figure 6: Empirical validation of Theorem 4.2 (Left column) and Theorem 4.4 (Right column) on Classification Synthetic Data (Top row), Covertype data (Middle row), and MNIST data (Bottom row). See Subsection D.2.1 for a detailed description of the experimental setup.

D Experimental Details

D.1 Experimental Setup for Regression Setting

Datasets. We evaluate TsCP on both synthetic and real-world regression datasets, covering a range of data distributions and noise structures.

- **Synthetic data.** We generate a SyntheticRegression dataset with $n = 20{,}000$ samples and input dimension $d = 20$, where features are drawn as $X \sim \mathcal{N}(0, I_d)$. The response is generated according to

$$Y = 0.05 \cdot (X^\top \beta + \varepsilon), \quad \varepsilon \sim \mathcal{N}(0, 3.0^2), \tag{227}$$

  where $\beta \in \mathbb{R}^d$ is fixed.

- **California Housing data.** The California Housing dataset (Pace & Barry, 1997) consists of 20,640 samples derived from the 1990 U.S. Census. Each sample corresponds to a census block group.

  The dataset contains 8 input features, including income, housing, population, and geographic variables (latitude and longitude), with the target being the median house value.

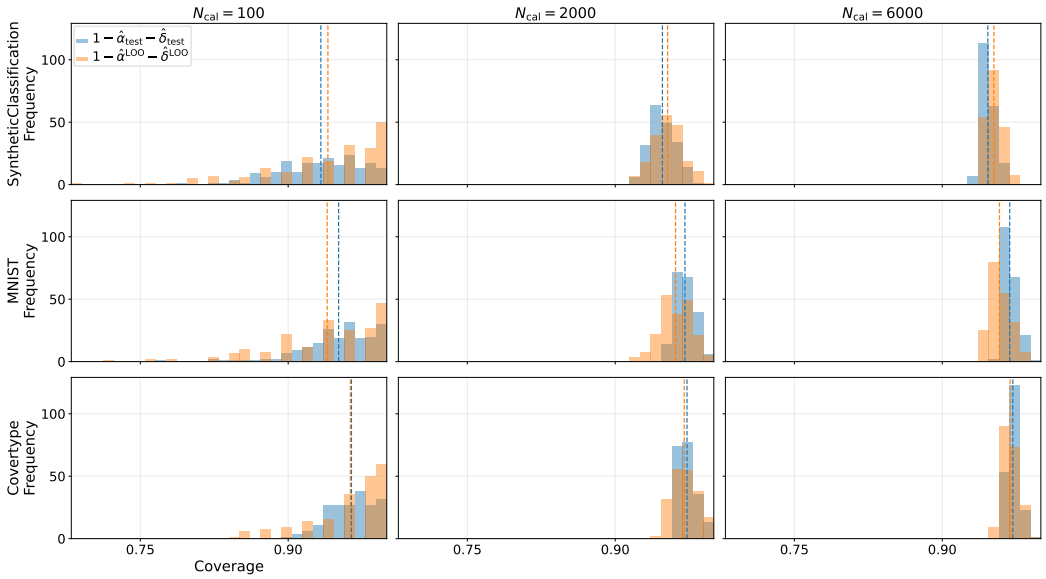


Figure 7: Empirical behavior of the LOO estimator corresponding to the convergence and variance properties in Theorems 5.2 and 5.3 on the Synthetic Classification dataset (top row), MNIST (middle row), and Covertype (bottom row). See Subsection D.2.1 for details of the experimental setup.

- **Superconductivity data.** The Superconductivity dataset (Hamidieh, 2018) is obtained from OpenML and consists of 21,263 samples with 81 input features describing physical and chemical properties of superconducting materials. The prediction task is to estimate the critical temperature of each material.

D.1.1 Validation of Theoretical Guarantees

Common setup. For all experiments, we use the normalized nonconformity score

$$s(x, y) = \frac{|y - \hat{f}(x)|}{\hat{\sigma}(x)}, \tag{228}$$

which induces prediction intervals of the form

$$C(x) = \left[\hat{f}(x) - \hat{\sigma}(x)q,\ \hat{f}(x) + \hat{\sigma}(x)q\right], \tag{229}$$

where q is a calibrated quantile. The size of the prediction set is given by $|C(x)| = 2\hat{\sigma}(x)q$.
Validation of Theorem 4.2. To validate Theorem 4.2, we follow a Monte Carlo procedure that mirrors the theoretical construction. For each run, we randomly sample a calibration subset of size $2n$ from the calibration pool and split it into two equal subsets $\mathcal{C}_a$ and $\mathcal{C}_c$, and randomly select a test point (X_0, Y_0).

For this test point, we compute $\alpha_\delta(X_0)$ using $\mathcal{C}_a$ by selecting the smallest exact empirical breakpoint $\alpha \in [\varepsilon, 1 - \varepsilon]$ such that

$$2\hat{\sigma}(X_0)\hat{q}_a(\alpha) \le c - \delta, \tag{230}$$

where $\hat{q}_a(\alpha)$ is the finite-sample conformal $(1 - \alpha)$-quantile of the scores computed on $\mathcal{C}_a$, and $\varepsilon = 10^{-10}$.

We then construct the prediction interval as

$$C(X_0) = \left[\hat{f}(X_0) - \hat{\sigma}(X_0)\hat{q}_c(\alpha_\delta(X_0)),\ \hat{f}(X_0) + \hat{\sigma}(X_0)\hat{q}_c(\alpha_\delta(X_0))\right], \tag{231}$$

where $\hat{q}_c(\cdot)$ is computed from $\mathcal{C}_c$.

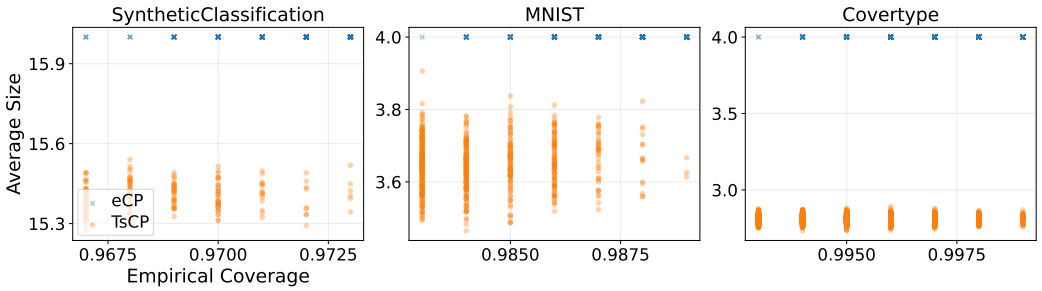


Figure 8: Comparison of TsCP with the eCP baseline on task **(a)**. Columns correspond to the Synthetic Classification, MNIST, and Covertype datasets, respectively. See Subsection D.2.2 for details of the experimental setup.

We evaluate the empirical coverage at different numbers of Monte Carlo test trials $m \in \{500, 1000, 1500, 2000, 2500\}$. For each value of m, we repeat the above procedure m times, resampling both the calibration subset and the test point in every trial, and compute

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{Y_i \in \mathcal{C}_i(X_i)\}\,. \tag{232}$$

Independently, using $R = 2000$ reference trials, we estimate the theoretical coverage quantity as

$$1 - \frac{1}{R}\sum_{r=1}^{R}\left[\alpha_\delta(X_r) + \Delta_{\delta,n}(\mathcal{C}_{a,r})\right]. \tag{233}$$

The empirical and reference streams are generated independently. Consequently, the theoretical curve does not depend on m, whereas the empirical curve converges to it as the number of Monte Carlo trials increases.

We use a total calibration size of $2n = 2000$. The dataset-specific budget and slack values are $(c, \delta) = (15, 1)$ for Synthetic Regression, $(c, \delta) = (2, 0.2)$ for California Housing, and $(c, \delta) = (60, 8)$ for Superconductivity.

Validation of Theorem 4.4. To validate Theorem 4.4, we fix the dataset-specific budget c and slack parameter δ, and vary the total calibration size $2n$. We use $(c, \delta) = (15, 1)$ for Synthetic Regression, $(c, \delta) = (2, 0.2)$ for California Housing, and $(c, \delta) = (60, 8)$ for Superconductivity. For each dataset, we consider

$$2n \in \{500, 1000, 2000, 4000, 6000\}. \tag{234}$$

For each calibration size, we repeatedly sample a calibration subset of size $2n$, split it into two equal subsets $\mathcal{C}_a$ and $\mathcal{C}_c$, and independently sample a test point X_0. The adaptive level $\alpha_\delta(X_0)$ is selected from the exact empirical quantile breakpoints, and the resulting prediction interval has length

$$|\mathcal{C}(X_0)| = 2\hat{\sigma}(X_0)\hat{q}_c(\alpha_\delta(X_0))\,. \tag{235}$$

We estimate the probability that the constructed interval satisfies the prescribed size constraint using

$$\widehat{\Pr}\{|\mathcal{C}(X)| \le c\} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{|\mathcal{C}_i(X_i)| \le c\}\,, \tag{236}$$

where both the calibration subset and test point are independently resampled in every Monte Carlo trial. We use $N = 1500$ trials for each calibration size and repeat the experiment across five independent outer seeds. We plot the mean size-control probability as a function of the total calibration size, with shaded regions indicating one standard deviation across seeds. Theorem 4.4 predicts that this probability approaches one as the calibration size increases.

Validation of Theorems 5.2 and 5.3. To validate Theorems 5.2 and 5.3, we evaluate the behavior of the leave-one-out (LOO) coverage estimator under repeated calibration sampling. For each total calibration size

$$2n \in \{100, 2000, 6000\}, \tag{237}$$

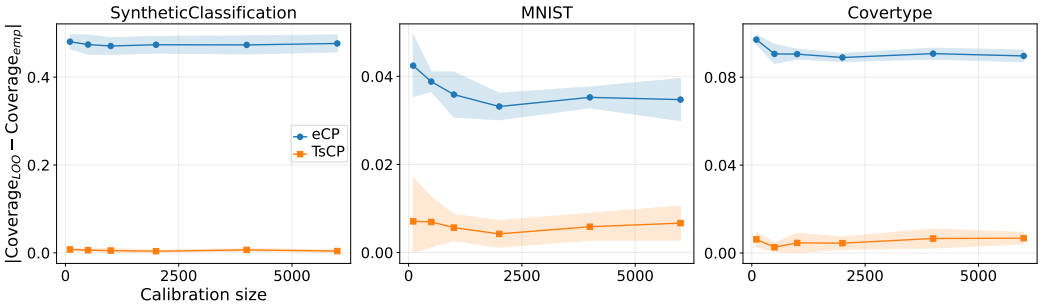


Figure 9: Comparison of TsCP with the eCP baseline on task **(b)**. Columns correspond to the Synthetic Classification, MNIST and Covertype datasets, respectively. See Subsection D.2.2 for details of the experimental setup.

we perform 200 independent Monte Carlo trials. In each trial, we randomly sample a calibration subset of size $2n$ and split it into two equal subsets $\mathcal{C}_a$ and $\mathcal{C}_c$. The adaptive levels are selected using the exact empirical quantile breakpoints in $[\varepsilon, 1-\varepsilon]$, with $\varepsilon = 10^{-10}$.

For each sampled calibration split, we compute the corrected LOO estimator

$$1 - \widehat{\alpha}_{\delta}^{\mathrm{LOO}} - \widehat{\Delta}_{\delta,n}^{\mathrm{LOO}}. \tag{238}$$

We compare this estimator against an independent-test estimate computed on a held-out batch of $m = 2500$ test observations:

$$1 - \widehat{\alpha}_{\mathrm{test}} - \widehat{\Delta}_{\mathrm{test}}, \tag{239}$$

where

$$\widehat{\alpha}_{\mathrm{test}} = \frac{1}{m}\sum_{j=1}^{m} \alpha_{\delta}(X_j), \tag{240}$$

$$\widehat{\Delta}_{\mathrm{test}} = \frac{1}{m}\sum_{j=1}^{m} \left[\mathbf{1}\{Y_j \notin \mathcal{C}(X_j)\} - \alpha_{\delta}(X_j)\right]. \tag{241}$$

By construction,

$$1 - \widehat{\alpha}_{\mathrm{test}} - \widehat{\Delta}_{\mathrm{test}} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{Y_j \in \mathcal{C}(X_j)\}, \tag{242}$$

so this quantity is also the empirical coverage on the independent test batch.

For each calibration size, we plot histograms of the independent-test coverage estimate and the corrected LOO estimate across the 200 Monte Carlo trials. We additionally report their empirical means using vertical reference lines. To assess the convergence result in Theorem 5.2, we examine whether the two distributions and their means become increasingly aligned as the calibration size grows. To assess the stability result in Theorem 5.3, we examine whether the distribution of the corrected LOO estimator becomes more concentrated as n increases.

We use the dataset-specific budgets $c = 15$ for Synthetic Regression, $c = 2$ for California Housing, and $c = 60$ for Superconductivity, with $\delta = 0.1$ for all datasets. Regression conformity scores are the normalized residual scores defined in the common experimental setup.

D.1.2 Comparison with Existing Methods

Task (a). We compare TsCP with the eCP baseline (Gauthier et al., 2025b;a) under the same experimental setup described above.

We evaluate the coverage-size trade-off using the same dataset-specific budget c for both methods: $c = 15$ for Synthetic Regression, $c = 3$ for California Housing, and $c = 60$ for Superconductivity.

For each dataset and seed, TsCP and eCP are evaluated using the same randomly sampled calibration subsets and test points. Empirical coverage and average prediction-set size are computed over 1000 shared test points, yielding a paired comparison under identical data realizations. We use a total calibration size of $2n = 2000$ and set $\delta = 0.001$ for TsCP.

To enable a fair comparison, we compare the prediction-set sizes of TsCP and eCP in regions where the two methods achieve similar empirical coverage.

Task (b). We compare the LOO-based coverage estimates of TsCP and truncated eCP (Gauthier et al., 2025a;b) under the same regression setup. We vary the total calibration size over $N_{\text{cal}} = 2n \in \{500, 1000, 2000, 4000, 6000\}$. At each size, both methods use the same randomly sampled calibration subsets and test points across 50 evaluation trials. Their LOO-based coverage estimates are $1 - \widehat{\alpha}^{\text{LOO}}$ for eCP and $1 - \widehat{\alpha}^{\text{LOO}} - \widehat{\Delta}^{\text{LOO}}_{\delta,n}$ for TsCP. For each method, we report the absolute difference between its mean LOO-based coverage estimate and its empirical coverage, estimated using 1,000 independent, paired calibration-test trials.

We repeat the experiment over five outer seeds, re-splitting the data and refitting the base and scale models for each seed. The curves show the mean absolute coverage difference across seeds, with shaded regions indicating one standard deviation. Both methods use the same dataset-specific constant budget: $c = 15$ for Synthetic Regression, $c = 2$ for California Housing, and $c = 60$ for Superconductivity. We set $\delta = 0.1$ for TsCP. For both methods, the adaptive levels are selected using the exact empirical breakpoints in $[\varepsilon, 1 - \varepsilon]$, with $\varepsilon = 10^{-10}$, rather than a fixed grid of miscoverage levels. When no feasible eCP level satisfies the prescribed budget, truncated eCP uses the fallback value $\alpha = 1 - \varepsilon$.

D.2 Experiments on the Classification Setting

Datasets. We evaluate TsCP on both synthetic and real-world classification datasets.

- **Synthetic data.** We generate $n = 20{,}000$ samples with $d = 20$ features using `make_classification` from `sklearn.datasets`, with 20 classes, 16 informative features, and 2 redundant features.
- **Covertype.** The Covertype dataset (Blackard & Dean, 1999) consists of 581,012 samples with 54 features and 7 classes. We use a subsampled version with 30,000 samples.
- **MNIST.** We use a subset of the MNIST dataset (LeCun et al., 2002) with 20,000 samples, where each input is a 28×28 grayscale image (flattened into 784 features). The task is digit classification with 10 classes.

Common setup. For all experiments, both TsCP and eCP use the same label-wise nonconformity score derived from the predicted class probabilities:

$$s(x, y) = 1 - \hat{p}_y(x), \tag{243}$$

where $\hat{p}_y(x)$ denotes the predicted probability of label y. Prediction sets are constructed by thresholding these scores.

D.2.1 Validation of Theoretical Guarantees

Validation of Theorem 4.2. To validate Theorem 4.2 in classification settings, we follow the same Monte Carlo procedure as in the regression setting in Section D.1.1. For each run, we randomly sample a calibration subset of size $2n$ from the calibration pool and split it into two equal subsets $\mathcal{C}_a$ and $\mathcal{C}_c$, and randomly select a test point (X_0, Y_0).

For this test point, we compute $\alpha_\delta(X_0)$ using $\mathcal{C}_a$ by selecting the smallest exact empirical breakpoint $\alpha \in [\varepsilon, 1 - \varepsilon]$ such that

$$|\mathcal{C}_a(X_0; \alpha)| \leq c - \delta, \tag{244}$$

where $\mathcal{C}_a$ is the prediction set constructed using the quantile from $\mathcal{C}_a$. We use $\varepsilon = 10^{-10}$. We then construct the final prediction set $\mathcal{C}(X_0)$ using $\mathcal{C}_c$.

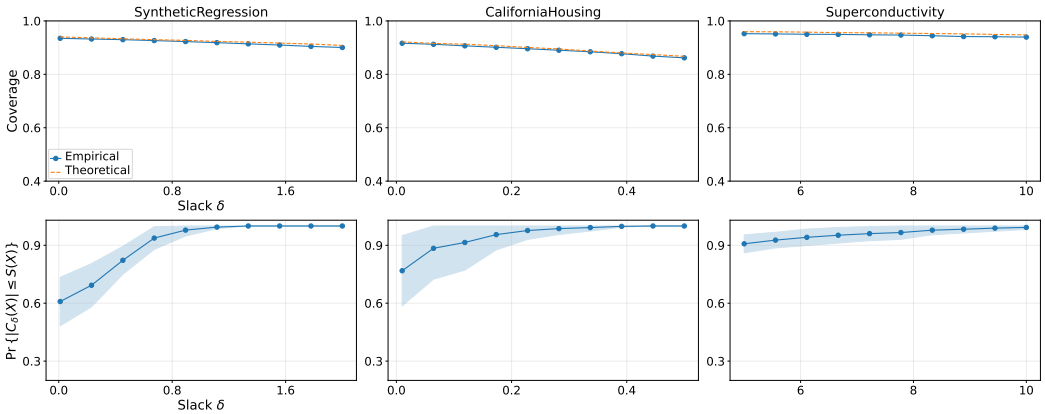


Figure 10: Coverage and size-control probability across different values of the slack parameter δ for the SyntheticRegression, California Housing, and Superconductivity datasets. See Subsection D.3 for details of the experimental setup.

We evaluate the empirical coverage at different numbers of test samples $m \in \{500, 1000, 1500, 2000, 2500\}$. For each value of m, we repeat the above procedure m times, each time resampling both the calibration subset and the test point, and compute

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{Y_i \in \mathcal{C}(X_i)\}\,. \tag{245}$$

Using an independent stream of $R = 2000$ reference trials, we compute its theoretical counterpart as

$$1 - \frac{1}{R}\sum_{r=1}^{R}\left[\alpha_\delta(X_r) + \Delta_{\delta,n}(\mathcal{C}_{a,r})\right]. \tag{246}$$

We use a total calibration size of $2n = 1000$. The dataset-specific budget and slack values are $(c, \delta) = (16, 2)$ for Synthetic Classification, $(c, \delta) = (4, 2)$ for MNIST, and $(c, \delta) = (4, 1)$ for Covertype. Results are averaged over five outer seeds.

Validation of Theorem 4.4. We adopt the same protocol as in the regression experiments (Section D.1.1) and focus on the classification setting. The budget c is fixed ($c = 16$ for Synthetic, $c = 5$ for MNIST, and $c = 4$ for Covertype), while the total calibration size is varied over $2n \in \{600, 800, 1000, 1200, 1600, 2000\}$.

For each configuration, prediction sets are constructed as described above, and we measure the probability that their cardinality satisfies the prescribed budget:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{|C(X_i)| \le c\}\,. \tag{247}$$

For each value of $2n$, we repeat the experiment across 5 independent runs and visualize the resulting mean size-control probability with one standard deviation. The results confirm that the probability of satisfying the prescribed budget c approaches one as $2n$ increases.

We use $\delta = 2$ for Synthetic and MNIST, $\delta = 1$ for Covertype, select α using the exact empirical breakpoints in $[\varepsilon, 1-\varepsilon]$ with $\varepsilon = 10^{-10}$, and use $N = 1500$ test samples.

Empirical behavior of the LOO estimator in classification We adopt the same LOO evaluation protocol as in the regression setting (Section D.1.1), and analyze the behavior of the leave-one-out estimator $\widehat{\alpha}_\delta$ in the classification setting. For each calibration size $2n \in \{100, 2000, 6000\}$, we perform 200 independent trials. In each trial, we resample the calibration subset, evaluate the resulting predictor on an independent batch of 2500 test points, and compute both the independent-test quantities $\widehat{\alpha}_{\text{test}}$ and $\widehat{\Delta}_{\text{test}}$, and their LOO estimates $\widehat{\alpha}_\delta^{\text{LOO}}$ and $\widehat{\Delta}_{\delta,n}^{\text{LOO}}$.

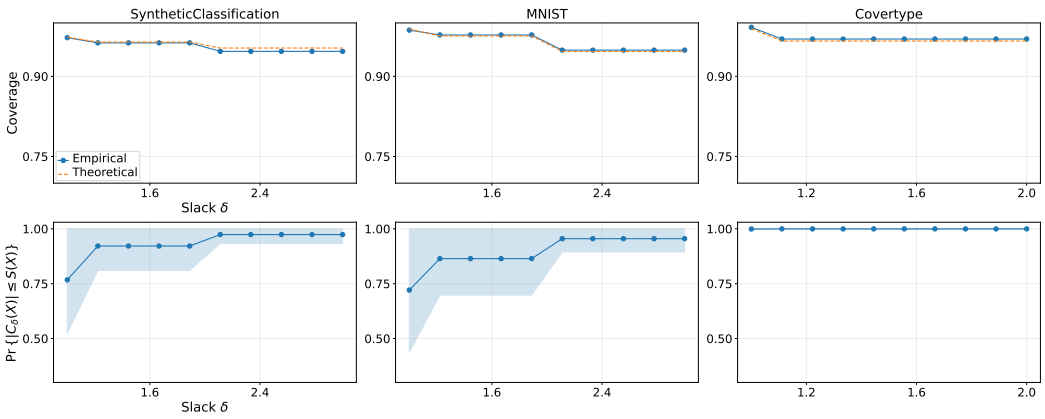


Figure 11: Coverage and size-control probability across different values of the slack parameter δ for the Classification Synthetic, MNIST, and Covertype datasets. See Subsection D.3 for details of the experimental setup.

We visualize the distributions of $1-\widehat{\alpha}_{\delta}^{\mathrm{LOO}}-\widehat{\Delta}_{\delta,n}^{\mathrm{LOO}}$ and $1-\widehat{\alpha}_{\mathrm{test}}-\widehat{\Delta}_{\mathrm{test}}$ using histograms, together with their empirical means indicated by vertical reference lines.

We use $c=16$ for Synthetic Classification and $c=4$ for Covertype and MNIST, $\delta=1$, and select α using the exact empirical breakpoints in $[\varepsilon, 1-\varepsilon]$, with $\varepsilon=10^{-10}$.

Results. Empirical validations of Theorems 4.2 and 4.4 on the Synthetic, Covertype, and MNIST datasets are presented in Figure 6. Figure 7 illustrates the empirical behavior of the LOO estimator corresponding to the convergence and variance properties characterized in Theorems 5.2–5.3. Overall, the empirical results demonstrate that the empirical coverage closely agrees with the theoretical coverage quantity in Theorem 4.2, while the probability that the prediction set satisfies the prescribed budget c approaches one as the calibration size increases, supporting Theorem 4.4 in practice. Furthermore, the empirical mean of $1-\widehat{\alpha}_{\delta}^{\mathrm{LOO}}-\widehat{\Delta}_{\delta,n}^{\mathrm{LOO}}$ closely matches that of $1-\widehat{\alpha}_{\mathrm{test}}-\widehat{\Delta}_{\mathrm{test}}$, showing analogous empirical convergence behavior in the classification setting. In addition, the dispersion of the histogram of $1-\widehat{\alpha}_{\delta}^{\mathrm{LOO}}-\widehat{\Delta}_{\delta,n}^{\mathrm{LOO}}$ decreases as n increases, indicating a corresponding reduction in estimator variability.

D.2.2 Comparison with Existing Methods

Task (a). We compare TsCP with the eCP baseline (Gauthier et al., 2025b;a) under the same setup as above in the classification setting.

For each dataset, both methods use the same fixed budget: $c=16$ for Synthetic Classification and $c=4$ for MNIST and Covertype. For each dataset and seed, TsCP and eCP are evaluated using the same randomly sampled calibration subsets and test points. Empirical coverage and average prediction-set size are computed over 1000 shared test points, yielding a paired comparison under identical data realizations. We use a total calibration size of $2n=2000$ and set $\delta=1$ for TsCP.

To enable a fair comparison, we compare the prediction-set sizes of TsCP and eCP in regions where the two methods achieve similar empirical coverage.

Task (b). We compare the LOO-based coverage estimates of TsCP and the truncated eCP baseline (Gauthier et al., 2025a;b) under the same classification protocol. We vary the total calibration size over

$$2n \in \{100, 500, 1000, 2000, 4000, 6000\}.$$

For each size, both methods use the same randomly sampled calibration subsets and test points across 50 evaluation trials. Their LOO-based coverage estimates are $1-\widehat{\alpha}^{\mathrm{LOO}}$ for eCP and $1-\widehat{\alpha}^{\mathrm{LOO}}-\widehat{\Delta}_{\delta,n}^{\mathrm{LOO}}$ for TsCP. For each method, we report the absolute difference between its mean LOO-based coverage estimate and its empirical coverage, estimated from 1,000 independent, paired calibration-test trials. We repeat the experiment over five outer seeds, re-splitting the data and refitting the logistic-regression

Table 1: Hard-constraint satisfaction under the proposed soft size-control procedure for both regression and classification datasets across different calibration set sizes. Each entry reports the hard-constraint satisfaction rate, defined as the percentage of test points satisfying $|C_\delta(X)| \leq S(X)$. See Subsection D.4 for details of the experimental setup.

Dataset	**Calibration Set Size**					
	1000	**2000**	**3000**	**4000**	**5000**	**6000**
Synthetic Regression	99.1	99.7	99.2	99.1	100	100
California Housing	95.7	100	95.4	96.1	97.8	100
Superconductivity	100	99.5	100	100	100	100
Synthetic Classification	100	96.6	100	100	99.6	99.6
MNIST	94.9	99.9	100	98.8	95.9	100
Covertype	99.9	100	100	99.9	100	99.9

classifier for each seed. The curves show the mean absolute difference across seeds, with one-standard-deviation bands. Both methods use the same constant budget within each dataset: $c = 16$ for Synthetic Classification and $c = 4$ for MNIST and Covertype. We set $\delta = 1$ for TsCP. For both methods, the adaptive level is selected from the exact empirical quantile levels within $[\varepsilon, 1 - \varepsilon]$. Both TsCP and eCP use the $1 - \widehat{p}_y(x)$ nonconformity score.

Results. The results for Task **(a)** are reported in Figure 8, while those for Task **(b)** are presented in Figure 9. Across all three datasets, TsCP consistently outperforms the eCP baseline on Task **(a)**, TsCP achieves smaller average prediction-set sizes while maintaining comparable empirical coverage to eCP. For Task **(b)**, we compare the LOO-based coverage bounds against the empirical coverage over a range of calibration sizes. Across all datasets, TsCP tracks the empirical coverage much more closely than eCP, whereas the eCP bound remains noticeably more conservative. This indicates that preserving the sCP structure within the TPSS setting leads to a substantially tighter practical coverage guarantee.

D.3 Ablation on the slack parameter δ

As described in Section 4, one of the key hyperparameters in TsCP is the slack parameter δ. This parameter plays a central role, as it influences both the theoretical coverage bound in Theorem 4.2 and the convergence behavior of the soft size constraint in Theorem 4.4. In this section, we examine the effect of δ in greater detail.

Why decreasing δ favors Coverage. From Theorem 4.2, the coverage guarantee is influenced by the quantity $\mathbb{E}[\alpha_\delta(X_0)]$. Recall the definition

$$\alpha_\delta(X) := \inf\left\{\alpha \in [\varepsilon, 1-\varepsilon] : \left|\left\{y \in \mathcal{Y} : s(\hat{f}(X), y) \leq \hat{q}_a(\alpha)\right\}\right| \leq \mathcal{S}(X) - \delta\right\}. \tag{248}$$

As δ decreases, the constraint $\mathcal{S}(X) - \delta$ becomes less restrictive. More precisely, for $\delta_1 < \delta_2$, we have

$$\mathcal{S}(X) - \delta_1 \geq \mathcal{S}(X) - \delta_2, \tag{249}$$

which implies

$$\{\alpha : N_X\left(\hat{q}_a(\alpha)\right) \leq \mathcal{S}(X) - \delta_2\} \subseteq \{\alpha : N_X\left(\hat{q}_a(\alpha)\right) \leq \mathcal{S}(X) - \delta_1\}, \tag{250}$$

where recall that $N_X(t) := \left|y \in \mathcal{Y} : s(\hat{f}(X), y) \leq t\right|$. Taking the infimum over these sets yields

$$\alpha_{\delta_1}(X) \leq \alpha_{\delta_2}(X), \tag{251}$$

so that $\alpha_\delta(X)$ is non-decreasing in δ. Consequently, $\mathbb{E}[\alpha_{\delta_1}(X_0)] \leq \mathbb{E}[\alpha_{\delta_2}(X_0)]$. Thus, decreasing δ decreases the $\mathbb{E}[\alpha_\delta(X_0)]$ component of the coverage bound, suggesting a more favorable coverage–size tradeoff. We note, however, that the full bound in Theorem 4.2 also contains the correction term $\Delta_{\delta,n}$, which itself depends on δ; therefore, the monotonicity of the complete theoretical coverage guarantee does not follow from the above argument alone and is examined empirically.

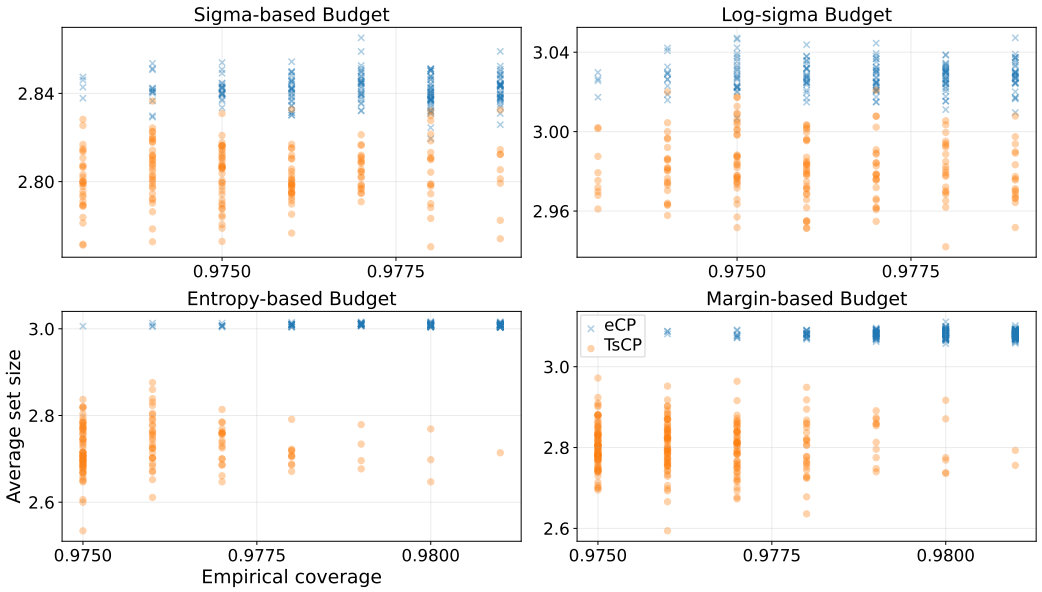


Figure 12: Comparison between TsCP and eCP under different budget functions on the California Housing regression dataset (top row) and the MNIST classification dataset (bottom row). See Subsection D.5 for details of the experimental setup.

‘‘‘

Why increasing δ favors Size Control. Recall that, by construction of $\alpha_\delta(X_0)$, we have

$$N_{X_0}\left(\hat{q}_a(\alpha_\delta(X_0))\right) \leq \mathcal{S}(X_0) - \delta. \tag{252}$$

On the other hand, since

$$|\mathcal{C}_\delta(X_0)| = N_{X_0}\left(\hat{q}_c(\alpha_\delta(X_0))\right), \tag{253}$$

we have

$$|\mathcal{C}_\delta(X_0)| \leq \mathcal{S}(X_0) \quad \text{whenever} \quad N_{X_0}\left(\hat{q}_c(\alpha_\delta(X_0))\right) N_{X_0}\left(\hat{q}_a(\alpha_\delta(X_0))\right) \leq \delta. \tag{254}$$

Thus, the slack δ provides a margin for absorbing the discrepancy between the two independently constructed quantiles $\hat{q}_a$ and $\hat{q}_c$. A larger δ provides a larger admissible margin against this discrepancy, which suggests improved size control and motivates the empirical trend observed in our experiments. However, because $\alpha_\delta(X_0)$, and hence the corresponding quantile level, also changes with δ, the above argument alone does not establish monotonicity of the probability that $|\mathcal{C}_\delta(X_0)| \leq \mathcal{S}(X_0)$.

Therefore, the slack parameter δ serves as a tunable hyperparameter that allows the user to balance coverage and size control, with smaller values favoring higher coverage and larger values enforcing stricter adherence to the size constraint. We now empirically examine how coverage and prediction set size vary with respect to δ in the following experiment.

Empirical Validations We will now empirically validate the role of the slack parameter δ discussed in the previous subsections, across both regression and classification settings.

Setup. We study the effect of the slack parameter δ while fixing the budget c and calibration size. For each dataset, we use dataset-specific budgets ($c = 15$ for SyntheticRegression, $c = 2$ for California Housing, and $c = 60$ for Superconductivity in regression; $c = 16$ for Synthetic Classification and $c = 4$ for MNIST and Covertype in classification), and sweep δ over dataset-specific ranges: $\delta \in [0.01, 2.0]$ for SyntheticRegression, $\delta \in [0.01, 0.5]$ for California Housing, and $\delta \in [5.0, 10.0]$ for Superconductivity in regression; $\delta \in [1, 3]$ for Synthetic Classification and $\delta \in [1, 2]$ for MNIST and Covertype in classification. Each range is evaluated at 10 equally spaced values. For each value of δ, we construct prediction sets using the same procedure as above with total calibration size $2n = 2000$ and evaluate performance on 1500 test samples. We report the empirical coverage, the

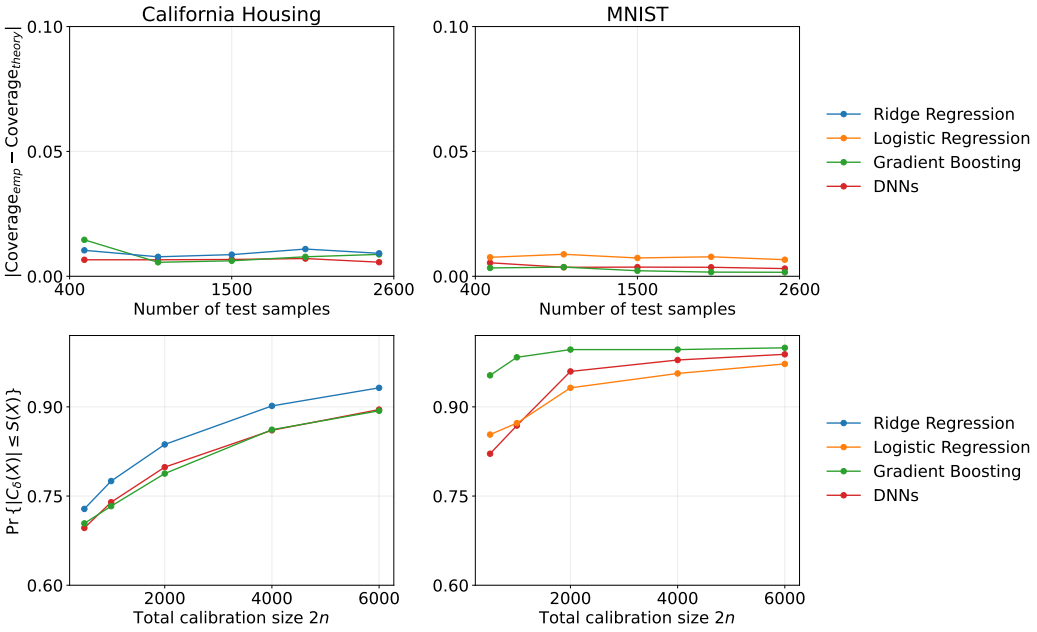


Figure 13: Absolute discrepancy between empirical and theoretical coverage (top) and size-control probability of TsCP across different numbers of test points and underlying models $\hat{f}$ on the California Housing (Left column) and the MNIST classification dataset (Right column). See Subsection D.6 for details of the experimental setup.

corresponding theoretical quantity $1 - \mathbb{E}\left[\alpha_\delta(X_0) + \Delta_{\delta,n}(\mathcal{C}_a)\right]$, and the size-control probability

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{|C_\delta(X_i)| \leq S(X_i)\right\}. \tag{255}$$

The adaptive level is selected from the exact empirical quantile levels within $[\varepsilon, 1-\varepsilon]$. We use the same models as in the main experiments.

Results. The results for the regression datasets are presented in Figure 10, while those for the classification datasets are reported in Figure 11. Across all datasets, we observe a consistent pattern: as the slack parameter δ increases, both empirical coverage and its theoretical counterpart decrease, while the size-control probability increases. These results are consistent with the discussion above and demonstrate that δ serves as a tunable hyperparameter governing the trade-off between stronger coverage guarantees and stricter control of prediction set size.

D.4 Ablation on Size Constraint Variants and Calibration Set Size

In the literature on the Tunable Prediction Set Size problem, the size constraint is typically imposed as a hard constraint of the form

$$|\mathcal{C}(X_0)| \leq \mathcal{S}(X_0). \tag{256}$$

In contrast, TsCP enforces a softer, probabilistic version of the constraint, given by

$$\mathbb{P}\big(|\mathcal{C}(X_0)| \leq \mathcal{S}(X_0)\big) \to 1 \quad \text{as } n \to \infty. \tag{257}$$

Although this soft constraint is theoretically weaker, as it only guarantees asymptotic satisfaction in probability, we demonstrate empirically that the proposed method effectively satisfies the hard constraint in practice. In particular, even for moderate to small calibration set sizes n, the prediction set size $|\mathcal{C}(X_0)|$ consistently remains below the prescribed budget $\mathcal{S}(X_0)$. This indicates that, despite only providing a soft guarantee theoretically, the method achieves behavior comparable to the hard constraint in finite-sample settings.

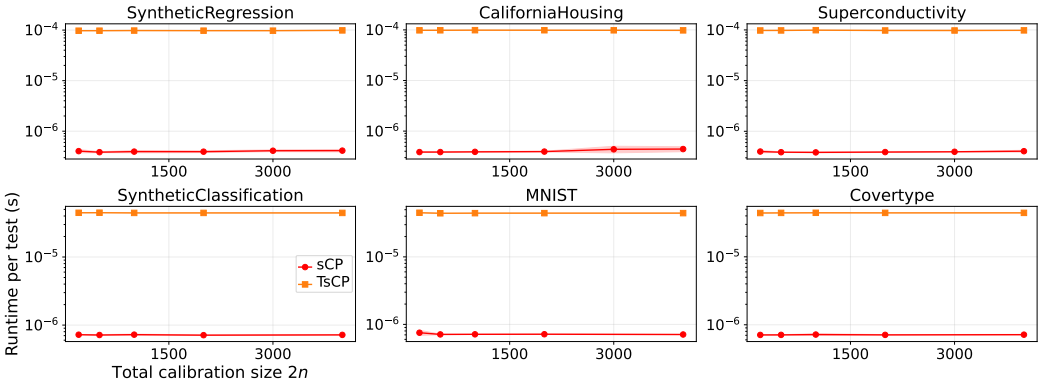


Figure 14: Comparison of the runtime of TsCP and standard sCP across different calibration set sizes. Results for the regression and classification datasets are shown in the top and bottom rows, respectively. See Subsection D.7 for details of the experimental setup.

Setup. We empirically evaluate hard-constraint satisfaction by measuring the fraction of test points for which the constructed prediction set satisfies the prescribed budget.

We vary the total calibration size $2n$ over a fixed grid $2n \in \{1000, 2000, 3000, 4000, 5000, 6000\}$. For each value of $2n$, we apply TsCP with a constant budget $S(x) = c$. We evaluate the resulting prediction sets on at most 1000 held-out test samples.

For each calibration size, we report the hard-constraint accuracy

$$\mathrm{HCA} := \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \mathbf{1}\{|C_\delta(X_i)| \le c\}. \tag{258}$$

This quantity measures the empirical probability that the hard constraint $|C_\delta(X)| \le c$ is satisfied. Thus, values close to one indicate that the soft guarantee in Theorem 4.4 yields near-hard size control in finite samples. We use $c = 15, \delta = 2$ for SyntheticRegression, $c = 2, \delta = 0.2$ for California Housing, $c = 60, \delta = 8$ for Superconductivity, $c = 16, \delta = 2$ for Synthetic Classification, $c = 4, \delta = 2$ for MNIST, $c = 4, \delta = 1$ for Covertype data.

Results. The results are reported in Table 1. Even at the smallest total calibration size considered ($2n = 1000$), the proposed soft size-control procedure satisfies the hard constraint for at least 94.9% of test points across all datasets, with four of the six datasets exceeding 99%. At the largest calibration size considered ($2n = 6000$), four datasets achieve 100% hard-constraint satisfaction, while the remaining two achieve 99.6% and 99.9%, respectively. Overall, these results show that, although TsCP is theoretically guaranteed only to satisfy the probabilistic size constraint in equation 4 asymptotically, it achieves near-hard size control at finite calibration sizes across all evaluated datasets.

Discussion on Calibration Set Size of TsCP. As discussed in Section 4, extending standard Split Conformal Prediction to the TPSS setting requires an additional approximation step. In particular, besides the usual calibration set $\mathcal{C}_c$, which is used to compute the final conformal threshold, TsCP uses an auxiliary approximation set $\mathcal{C}_a$ to estimate the size-aware quantile level. This naturally increases the total amount of calibration data required by the procedure and may appear, at first sight, to be a potential limitation.

However, this increase is moderate in practice. As shown empirically by the experiment at the beginning of this section, with results reported in Table 1, the size of $\mathcal{C}_a$ needed for TsCP to satisfy the hard size constraint in Equation 3 with high probability is typically small to moderate. Therefore, the total calibration budget used by TsCP is usually no more than about twice that of standard Split Conformal Prediction. Thus, TsCP should be viewed as requiring only a constant-factor increase in calibration data, rather than a qualitatively larger amount of data.

This constant-factor increase is also mild from the perspective of conformal calibration. Indeed, the behavior of the conditional coverage induced by conformal calibration is well understood. The following classical result, gives the exact distribution of the conditional coverage.

Theorem D.1 (Distribution of conformal conditional coverage (Angelopoulos & Bates, 2021)). *Assume that the calibration scores and test score are i.i.d. and that there are no ties. Let $\hat{q}$ be the conformal threshold given by the l-th largest calibration-based order statistic, equivalently corresponding to the conformal level determined by l. Then the conditional coverage satisfies*

$$\mathbb{P}\left(Y_0 \in C(X_0) | \{(X_i^c, Y_i^c)\}_{i=1}^n\right) \sim \mathrm{Beta}(n+1-l, l). \tag{259}$$

In particular, if $l = \lfloor (n+1)\alpha \rfloor$, then the conditional coverage concentrates around $1-\alpha$ as n increases.

Theorem D.1 shows that the randomness of the realized conditional coverage is controlled directly by the number of calibration samples. For the beta distribution above, the variance is of order $\mathcal{O}(1/n)$. Moreover, because the beta parameters are both of order n when α is fixed away from 0 and 1, standard order-statistic concentration bounds imply that the probability of any fixed deviation from the target coverage decays exponentially fast in n. Hence, increasing the total calibration requirement by only a constant factor does not fundamentally change the sample-efficiency profile of conformal calibration. In TsCP, the set $\mathcal{C}_c$ still plays the same role as the usual sCP calibration set, while $\mathcal{C}_a$ is used only to learn the size-aware quantile level. Therefore, the additional cost of $\mathcal{C}_a$ is empirically modest and theoretically consistent with the known concentration behavior of conformal calibration.

D.5 Ablation on Budget Function $\mathcal{S}$

In the main experiments in Section 6, we consider the case where the budget function $\mathcal{S}(X)$ is constant across all inputs. However, in practice, the desired prediction set size may depend on the input X. For instance, in classification tasks, more ambiguous inputs may require larger prediction sets, while in regression settings, regions with higher uncertainty may benefit from more flexible budgets. In this section, we evaluate the performance of the proposed method under different choices of the budget function $\mathcal{S}(X)$ and compare it with baseline methods for TPSS, demonstrating the robustness of our approach to varying budget specifications.

Setup. We consider the same task as Task **(a)** in Subsection 6.2, but instead of using a constant budget function $S(X) \equiv c$, we consider several input-dependent budget functions. To construct adaptive budgets, we define $S(X)$ as a convex combination of two user-specified endpoints, $S_{\min}$ and $S_{\max}$, where the interpolation weight is determined by an input-dependent uncertainty score. This design ensures that $S(X)$ remains within a prescribed feasible range while allowing the allowable prediction set size to adapt to the difficulty of each input: easier inputs receive tighter budgets closer to $S_{\min}$, whereas more ambiguous or uncertain inputs are assigned more flexible budgets closer to $S_{\max}$. In this way, the budget function preserves explicit user control over the admissible size range while introducing instance-dependent adaptivity.

- **Classification Setting.** We define our score function similarly to that of Subsection D.2. Furthermore, given predicted class probabilities $\hat{p}(x) = (\hat{p}_1(x), \ldots, \hat{p}_K(x))$, let $\hat{p}_{(1)}(x)$ and $\hat{p}_{(2)}(x)$ denote the largest and second-largest probabilities, respectively. We consider the following budget functions:
  - **Entropy-based:**

$$S(x) = \left\lceil S_{\min} \cdot \left(1 - \frac{H(\hat{p}(x))}{\log K}\right) + S_{\max} \cdot \frac{H(\hat{p}(x))}{\log K} \right\rceil, \tag{260}$$

$$\text{where } H(\hat{p}(x)) = -\sum_{k=1}^{K} \hat{p}_k(x) \log \hat{p}_k(x). \tag{261}$$

This budget assigns larger allowable prediction set sizes to more uncertain inputs, using the entropy of the predictive distribution as a measure of uncertainty. When the model is confident, the entropy is low, and the budget remains close to $S_{\min}$; when the prediction is more ambiguous, the entropy increases and the budget moves to $S_{\max}$.

– **Margin-based:**

$$S(x) = \left\lceil S_{\min} \cdot (\hat{p}_{(1)}(x) - \hat{p}_{(2)}(x)) + S_{\max} \cdot \left(1 - (\hat{p}_{(1)}(x) - \hat{p}_{(2)}(x))\right) \right\rceil. \quad (262)$$

This budget function uses the gap between the two most likely classes as a measure of predictive confidence. When the top class is clearly separated from the runner-up, the margin $\hat{p}_{(1)}(x) - \hat{p}_{(2)}(x)$ is large and the budget remains close to $S_{\min}$. In contrast, when the top two classes are close, indicating greater ambiguity, the margin is small and the budget increases toward $S_{\max}$.

- **Regression Setting.** We define our score to be the normalized score (Angelopoulos & Bates, 2021):

$$s(\hat{f}(X), Y) = \frac{|y - \hat{f}(x)|}{\hat{\sigma}(x)}, \quad (263)$$

where $\hat{\sigma}(x)$ is the fitted scale model. Furthermore, given the calibration-scale values $\hat{\sigma}_i$, we let $q_{0.05}$ and $q_{0.95}$ denote the 5% and 95% quantiles of $\hat{\sigma}_i$, and let $q^{\log}_{0.05}$ and $q^{\log}_{0.95}$ denote the corresponding quantiles of $\log \hat{\sigma}_i$. With these notations, we then define the following adaptive budget functions.

– **Sigma-based budget.**

$$S(x) = S_{\min} + (S_{\max} - S_{\min}) \cdot \text{clip}\left(\frac{\hat{\sigma}(x) - q_{0.05}}{q_{0.95} - q_{0.05}}, 0, 1\right). \quad (264)$$

This budget uses the fitted scale estimate $\hat{\sigma}(x)$ as a proxy for local predictive uncertainty. Inputs with smaller estimated scale are assigned tighter budgets closer to $S_{\min}$, while inputs with larger estimated scale receive more flexible budgets closer to $S_{\max}$. Normalizing $\hat{\sigma}(x)$ between the empirical 5% and 95% quantiles of the calibration scales yields an input-dependent budget that adapts to heteroscedastic uncertainty while remaining within a prescribed range.

– **Log-sigma-based budget.**

$$S(x) = S_{\min} + (S_{\max} - S_{\min}) \cdot \text{clip}\left(\frac{\log \hat{\sigma}(x) - q^{\log}_{0.05}}{q^{\log}_{0.95} - q^{\log}_{0.05}}, 0, 1\right). \quad (265)$$

This budget applies the same construction as the first one but on the log scale, replacing $\hat{\sigma}(x)$ with $\log \hat{\sigma}(x)$. This transformation reduces sensitivity to extreme scale values and provides a more stable interpolation when the fitted scales are highly skewed. As a result, inputs with larger uncertainty still receive larger budgets, but the growth is moderated on the log scale, yielding a more conservative adaptive budget under heavy-tailed or highly variable uncertainty.

Additionally, for classification we use the MNIST dataset with $S_{\min} = 2$ and $S_{\max} = 4$, while for regression we use the California Housing dataset with $S_{\min} = 2.5$ and $S_{\max} = 3.5$.

Results. The results are presented in Figure 12. Overall, TsCP consistently achieves substantially smaller prediction set sizes while maintaining similar coverage to that of eCP across all budget functions. These results demonstrate that the proposed method remains effective under a range of input-dependent budget specifications.

D.6 Ablation on the model $\hat{f}$

In the main experiments in Section 6, for simplicity, we use Random Forest and Logistic Regression as the base models for regression and classification tasks, respectively. However, in practice, a wide range of models are employed depending on the application, including more complex models such as gradient boosting methods and neural networks, which can better capture nonlinear patterns and high-dimensional structures. In this section, we evaluate the performance of the proposed method across different choices of base models $\hat{f}$ for both regression and classification tasks to demonstrate that our approach remains robust and effective regardless of the underlying predictive model.

Setup. To assess model dependence, we evaluate the proposed method using different choices of the base predictor $\hat{f}$ in both classification and regression settings. The main experiments use logistic regression for classification and random forests for regression. For classification, we consider logistic regression, histogram-based gradient boosting, and a neural network implemented as a multilayer perceptron (MLP) with two hidden layers. For regression, we consider ridge regression, gradient boosting, and an MLP regressor with two hidden layers. In each case, we follow the same experimental setup as in Subsection 6.1 for regression and Subsection D.2.1 for classification.

Results. The results are presented in Figure 13. Across all choices of the underlying model, the empirical coverage achieved by TsCP closely matches the theoretical coverage quantity in Theorem 4.2, as indicated by the consistently small absolute differences. Moreover, the probability that the prediction-set size satisfies the prescribed budget c generally increases with the calibration size and reaches high values across both regression and classification settings, in agreement with Theorem 4.4. These results demonstrate that the proposed method remains robust across a diverse range of model classes.

D.7 Ablation on Inference Time

D.7.1 Computational Overhead of TsCP compared to sCP

While TsCP extends standard Split Conformal Prediction to the TPSS setting and provides explicit user control over the prediction set size, it also incurs additional inference-time overhead. The main reason is that, unlike classical Split Conformal Prediction, TsCP requires an additional step to determine the data-dependent level $\alpha_\delta(X)$ for each test input X, before constructing the final prediction set using the independent calibration set.

To make this precise, we quantify the additional overhead per test input X_0. In standard Split Conformal Prediction, once the calibration quantile $\hat{q}(\alpha)$ is fixed, constructing $\mathcal{C}(X_0)$ requires a single evaluation of the score function over the candidate labels, resulting in an inference cost of order $O(M)$, where M denotes the cost of forming the prediction set.

In contrast, TsCP additionally computes $\alpha_\delta(X_0)$ using the monotonicity of the prediction-set size with respect to the conformal quantile. The approximation and calibration scores are sorted once during preprocessing. For each test input, the required empirical rank is then found by binary search over the sorted approximation scores in $O(\log n)$ time, after which the corresponding calibration quantile is retrieved directly. Therefore, the inference cost per test input is

$$\mathrm{Time}_{\mathrm{TsCP}}(X_0) = O(M + \log n), \tag{266}$$

compared to

$$\mathrm{Time}_{\mathrm{sCP}}(X_0) = O(M) \tag{267}$$

for standard Split Conformal Prediction. The one-time sorting cost of $O(n \log n)$ is treated as calibration preprocessing and is excluded from the per-input inference time.

D.7.2 Empirical Comparison with standard Split Conformal Prediction

As conformal prediction is often used as an inference-time uncertainty quantification procedure, this additional cost should remain moderate in practice. In this section, we therefore empirically investigate the runtime of the proposed method and compare it with standard Split Conformal Prediction in order to assess the computational overhead induced by the TPSS extension.

Setup. For both TsCP and standard Split Conformal Prediction, we measure the runtime required to construct prediction sets over a fixed collection of test inputs while varying the calibration set size. Specifically, we evaluate 1000 test points and report the average inference time per test input. To isolate the computational cost of the conformal procedure itself, model predictions (i.e., evaluations of $\hat{f}(x)$ and the scale function, when applicable) are precomputed and excluded from the runtime measurement. We vary the calibration set size to assess its effect on computational overhead: for classification, we consider calibration sizes in $\{200, 500, 1000, 2000\}$, while for regression we use $\{200, 500, 1000, 2000, 3000, 4000\}$. All experiments are repeated over five independent random seeds, and we report the average runtime across runs.

Results. The results are presented in Figure 14. Consistent with the complexity analysis above, TsCP incurs a higher inference-time cost than standard Split Conformal Prediction. In relative terms, TsCP is more than two orders of magnitude slower than sCP across the considered datasets. Nevertheless, its absolute runtime remains small, at approximately 10^{-4} seconds per test input. Moreover, the runtime changes only slightly as the calibration set size increases, consistent with the logarithmic dependence on n after the one-time preprocessing of the calibration scores. Thus, although the relative overhead is substantial, the absolute inference latency remains computationally manageable in the settings considered.

D.8 Supplementary Statements

Computational Infrastructure. All reported experiments were conducted on CPU-only machines with 4 CPU cores and 4GB of RAM; no GPU acceleration was used.