CONFORMAL PREDICTION UNDER DISTRIBUTION SHIFT: FALSE COVERAGE PROPORTION GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal Prediction (CP) is a distribution-free and model-agnostic framework for uncertainty quantification that constructs prediction sets with rigorous finite-sample coverage guarantees. Despite these advantages, the classical CP guarantee is marginal and therefore does not directly control the realized proportion of miscovered observations across a test sample. This has motivated the study of the False Coverage Proportion (FCP) in the literature, which provides a stronger form of coverage assessment over multiple predictions. Nevertheless, both classical CP and existing FCP theory typically rely on exchangeability between calibration and test data, an assumption that may fail under calibration-to-test distribution shift. Recent works have proposed Weighted CP to address this issue for classical CP by reweighting calibration observations to account for such distribution shifts. However, existing FCP theory does not directly extend to this weighted setting and has largely remained restricted to exchangeable data, leaving FCP control under distribution shift substantially unexplored. Therefore, in this work, we establish finite-sample FCP guarantees for Weighted CP under calibration-to-test distribution shift. We further study the inverse FCP problem of selecting a miscoverage level to achieve a prescribed target FCP. For practical implementation, we also develop data-driven procedures for estimating the quantities appearing in our bounds and establish their convergence. Empirically, we validate our theoretical guarantees across multiple classification and regression datasets under both classical Covariate Shift and more challenging one-phase shifts. We further investigate the large-sample behavior of the proposed bounds through asymptotic studies.

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; Park et al., 2021; Pournaderi & Xiang, 2024) is a general framework for uncertainty quantification that provides distribution-free, model-agnostic predictive guarantees under a mild exchangeability assumption. At a high level, CP augments the output of an arbitrary predictive model by using previously observed data to calibrate its prediction errors and construct a set of plausible outcomes. Among the various formulations of CP, *Split Conformal Prediction* (sCP) is particularly widely used in practice due to its simplicity and low computational overhead (Papadopoulos, 2008; Angelopoulos & Bates, 2021). In sCP, the predictive model is fitted on a training set, while a held-out calibration set is used to assess prediction errors and construct the resulting prediction sets. These prediction sets enjoy rigorous finite-sample guarantees, ensuring that the true response is covered with probability at least a user-specified target level over repeated sampling. Owing to this flexibility and practical simplicity, CP and sCP have been adopted in a wide range of applications, 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., 2025), 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; Yapicioglu et al., 2025), and federated learning (Lu & Kalpathy-Cramer, 2021; Lemercier et al., 2021; Lu et al., 2023; Plassier et al., 2023; Humbert et al., 2023).

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.

Despite these advantages, the classical CP guarantee is marginal, characterizing coverage for an individual test observation rather than the realized coverage behavior over an entire test sample. This distinction becomes particularly important when prediction sets are constructed for multiple test observations simultaneously. Although marginal coverage controls the expected proportion of miscovered observations, it does not directly control the fraction of miscoverages realized on a particular test sample. This has motivated the study of the stronger notion of *False Coverage Proportion* (FCP) (Vovk, 2013; Gazin et al., 2024; 2025; Blain et al., 2025; Song et al., 2026; Fermanian et al., 2026), which measures the proportion of test observations whose true responses are not contained in their corresponding prediction sets. FCP guarantees aim to control this realized fraction of miscovered test points with high probability, thereby providing a stronger form of simultaneous coverage assessment across multiple predictions.

While providing stronger coverage control, existing FCP theory inherits the same exchangeability assumption as classical CP (see Definition A.1) between the calibration and test observations. This mild assumption, however, can still be violated in many real-world deployment settings, where the test distribution differs systematically from the distribution used for calibration, a phenomenon commonly referred to as *Distribution Shift*. For example, a model calibrated using data from one patient population may later be deployed on a population with a different distribution of demographic or clinical characteristics. To accommodate such settings for classical CP, recent work has developed *Weighted Conformal Prediction* (Tibshirani et al., 2019; Barber et al., 2023; Gibbs & Candes, 2021; Aolaritei et al., 2026; Xu et al., 2025). The key idea is to assign nonuniform weights to the calibration observations so that their weighted empirical distribution better reflects the test distribution, thereby adapting the conformal calibration procedure to the underlying Distribution Shift. Nonetheless, existing FCP theory does not directly extend to this weighted setting and has largely remained restricted to the standard exchangeable setting. To our knowledge, the closest related work that moves beyond exchangeability is Gazin (2024), which studies the behavior of the FCP under Distribution Shift, but only from an asymptotic perspective; we provide a more detailed comparison in Appendix A.2. This leaves open the problem of obtaining finite-sample FCP guarantees under Distribution Shift, where one must quantify the deviation of the realized FCP at the available calibration and test sample sizes.

Contribution. Motivated by this gap, we develop finite-sample theoretical guarantees for the FCP of CP under Distribution Shift. Our analysis covers both fixed miscoverage levels and guarantees that hold uniformly over the entire range of miscoverage levels, and we further study the corresponding inverse FCP problem of selecting a conformal level to achieve a prescribed target FCP. To bridge these theoretical results with practical implementation, we additionally propose data-driven procedures for estimating the quantities appearing in our bounds and establish their uniform consistency, together with convergence rates, under suitable assumptions. The paper is organized as follows:

1. In Section 3, we discuss the limitations of existing theoretical FCP guarantees under Distribution Shift and provide empirical evidence showing that these bounds can fail substantially when the calibration and test distributions differ.
2. In Section 4, we present our main theoretical finite-sample FCP guarantees of CP under Distribution Shift for both fixed and uniform miscoverage levels.
3. In Section 5, we develop practical algorithms for estimating the theoretical FCP bounds from observed data and establish their convergence properties. We also study the inverse FCP problem, deriving theoretical guarantees and practical estimation procedures.
4. In Section 6, we empirically validate our method across multiple classification and regression datasets under both classical Covariate Shift and the more challenging stratified score-transport shift under different weight functions. We further investigate the large-sample behavior of our proposed bounds through asymptotic studies.

Section 2 reviews background on CP, Split CP, Weighted CP, and FCP. A table of notation, additional background on CP, theoretical details, and further experimental results are deferred to the Appendix.

2 PRELIMINARIES

In this section, we review Conformal Prediction (CP), Weighted CP, One-Phase Distribution Shift (OPDS), and False Coverage Proportion (FCP). Further details are given in Appendix A.

Conformal Prediction (CP). 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 CP methods is to construct a random set-valued predictor $\mathcal{C}^\alpha : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, where $\mathcal{P}(\mathcal{Y})$ denotes the collection of all subsets of $\mathcal{Y}$, such that

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

Here, the probability is taken jointly over the randomness of the set $\mathcal{C}^\alpha$ and the test pair (X_0, Y_0).

In particular, in *split Conformal Prediction* (Angelopoulos & Bates, 2023), the data are partitioned into three disjoint subsets: a training set $\mathcal{D}_{\text{train}}$, a calibration set $\mathcal{D}_{\text{cal}} = \{Z_i = (X_i, Y_i)\}_{i=1}^n$, and a test set $\mathcal{D}_{\text{test}} = \{Z_{n+j} = (X_{n+j}, Y_{n+j})\}_{j=1}^m$. The prediction model $\hat{f}$ is trained on $\mathcal{D}_{\text{train}}$, independently of the calibration and test data. Then, a fixed nonconformity score function

$$s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

is specified by the user, and the calibration and test scores are defined, respectively, by $S_i := s(X_i, Y_i)$, $i \in [n]$, and $T_j := s(X_{n+j}, Y_{n+j})$, $j \in [m]$. For each test input X_{n+j}, $j \in [m]$, the prediction set at a user-specified miscoverage level α is defined as

$$\mathcal{C}^\alpha(X_{n+j}) := \left\{ y \in \mathcal{Y} : s(X_{n+j}, y) \leq Q_{1-\alpha}\left(\frac{1}{n+1} \sum_{i=1}^n \delta(S_i) + \frac{1}{n+1}\delta(\infty) \right) \right\}, \tag{2}$$

where $\delta(a)$ denotes a point mass at a, and $Q_{1-\alpha}(\cdot)$ denotes the $(1-\alpha)$-quantile. Under this construction, split CP admits the following finite-sample coverage guarantee.

Theorem 2.1. *[Coverage guarantee of sCP (Angelopoulos & Bates, 2023)] Assume that $\mathcal{D}_{cal} \cup \{(X_{n+1}, Y_{n+1})\}$ are exchangeable (See Definition A.1). Then the following inequality holds:*

$$1 - \alpha \;\leq\; \mathbb{P}(Y_{n+1} \in \mathcal{C}^\alpha(X_{n+1}))$$

Weighted Conformal Prediction. In real-world applications, the exchangeability assumption in Theorem 2.1 may fail to apply under Distribution Shift (see Appendix A.3). To accommodate such settings, previous works (Tibshirani et al., 2019; Barber et al., 2023; Gazin, 2024) introduce a non-negative, nontrivial weight function $w : \mathcal{X} \to \mathbb{R}_{\geq 0}$ for the calibration observations, together with an additional positive weight $w_\infty \in \mathbb{R}_{>0}$ for the test point[1]. These weights induce a weighted empirical distribution over the calibration scores, allowing the conformal procedure to adjust to differences between the calibration and test distributions. For a test input X_{n+1}, define the normalizing constant $W := w_\infty + \sum_{k\in[n]} w(X_k)$. The weighted CP set is then given by

$$\mathcal{C}^{w,\alpha}(X_{n+1}) := \left\{ y \in \mathcal{Y} : s(X_{n+1}, y) \leq Q_{1-\alpha}\left(\frac{w_\infty}{W}\delta(\infty) + \sum_{k\in[n]} \frac{w(X_k)}{W}\delta(S_k) \right) \right\}. \tag{3}$$

Equivalently, for $i \in [m]$, if we define the p-value for the test point X_{n+i} by

$$p^{w,(n)}(X_{n+i}, y) := \frac{w_\infty + \sum_{k=1}^n w(X_k)\mathbf{1}\{S_k \geq s(X_{n+i}, y)\}}{w_\infty + \sum_{k=1}^n w(X_k)}, p_i^{w,(n)} := p^{w,(n)}(X_{n+i}, Y_{n+i}).$$

Then, the prediction set $\mathcal{C}^{w,\alpha}$ in equation 3 can be written as

$$\mathcal{C}^{w,\alpha}(X_{n+i}) = \left\{ y \in \mathcal{Y} : p^{w,(n)}(X_{n+i}, y) > \alpha \right\}, \tag{4}$$

One-Phase Distribution Shift (OPDS). In this work, we will focus on the OPDS setting, which we formalize in the following assumption. Possible extensions to more general distribution shifts are discussed in Appendix C.3.

Assumption 2.2 (One-Phase Distribution-Shift)**.** *Conditionally on the independent training data, the score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and the weight function w are fixed measurable functions. The calibration observations $\{Z_k^{\text{cal}} = (X_k^{\text{cal}}, Y_k^{\text{cal}})\}_{k\in[n]}$, are i.i.d. with distribution P_{cal}, while the test observations $\{Z_i^{\text{test}} = (X_i^{\text{test}}, Y_i^{\text{test}})\}_{i\in[n+1,n+m]}$, are i.i.d. with distribution P_{test}. The two families are independent. Let $F_{\text{cal}}(t) := P_{\text{cal}}(s(X, Y) \leq t)$, and $F_{\text{test}}(t) := P_{\text{test}}(s(X, Y) \leq t)$ denote the score CDF. We assume further that F_{cal} and F_{test} are continuous.*

[1] Some prior work, such as Tibshirani et al. (2019), allows w_∞ to depend on the test data. In this work, however, we restrict attention to fixed w_∞, following Gazin (2024). Further discussion is provided in Appendix C.1.

Additionally, we impose the following mild assumption on the weight function w.

Assumption 2.3. *The weight function w is uniformly bounded. That is, there exists a constant $0 < \varepsilon_w < \infty$ such that $w(x) \le \varepsilon_w$ for all $x \in \mathcal{X}$.*

Discussion of Assumptions 2.2 and 2.3. The OPDS setting specified by Assumption 2.2 is standard and has been widely considered in the literature (Gazin, 2024; Tibshirani et al., 2019). Notably, when $P_{\text{cal}} = P_{\text{test}}$, this assumption reduces the setting to the usual i.i.d. assumption. Assumption 2.3 is satisfied, for example, when the covariate space is compact and w is continuous and finite-valued on that space. For the oracle weight (See equation 5), it rules out cases in which the test distribution places substantially more mass than the calibration distribution in regions where the latter has arbitrarily small density. In empirical settings with a finite dataset, we observe that it is sufficient that $w(X_i) < \infty$ for every observed data point. Accordingly, it is typically a mild assumption in practice.

Covariate Shift. A well-known special case of OPDS is *Covariate Shift* (Tibshirani et al., 2019). In this setting, one would impose the following additional assumption on the shift.

Assumption 2.4. *The calibration and test distributions satisfy $P_{\text{cal}}(Y \mid X) = P_{\text{test}}(Y \mid X)$. Moreover, the test covariate distribution is absolutely continuous with respect to the calibration covariate distribution: $P^X_{\text{test}} \ll P^X_{\text{cal}}$.*

Additionally, for Covariate Shift, the weight function is chosen as the *oracle weight function*

$$w(x) = w_0(x) := \frac{dP^X_{\text{test}}}{dP^X_{\text{cal}}}(x), \qquad x \in \mathcal{X}. \tag{5}$$

The resulting weighted CP procedure satisfies the following marginal coverage guarantee.

Theorem 2.5 (Coverage guarantee under Covariate Shift (Tibshirani et al., 2019))**.** *For $w = w_0$ as defined in equation 5, and with the test-point weight chosen adaptively as $w_\infty = w_0(X_{n+1})$, under Assumptions 2.2 and 2.4, the resulting weighted Conformal Prediction set satisfies*

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}^{w_0,\alpha}(X_{n+1})) \ge 1 - \alpha.$$

We emphasize that, as stated above, our analysis takes w_∞ to be fixed, rather than setting $w_\infty = w_0(X_{\text{test}})$ as in Tibshirani et al. (2019). Consequently, Theorem 2.5 need not hold under our convention. This does not affect our objective, however, since our focus is on controlling the FCP over multiple test points, which is a stronger form of error control than marginal coverage.

False Coverage Proportion (FCP). While Theorem 2.5 provides a marginal coverage guarantee for weighted CP, our focus is on the realized coverage behavior across multiple test points. Following Vovk (2013), we consider the False Coverage Proportion over m test observations, defined as

$$\text{FCP}^{w,(n)}_m(\alpha) := \frac{1}{m}\sum_{i\in[m]} \mathbf{1}\{Y_{n+i} \notin \mathcal{C}^{w,\alpha}(X_{n+i})\} = \frac{1}{m}\sum_{i\in[m]} \mathbf{1}\left\{p_i^{w,(n)} \le \alpha\right\}. \tag{6}$$

Following prior works, the main theoretical objective for FCP is to obtain a high-probability upper bound on the realized FCP, either at a fixed miscoverage level α or, more strongly, uniformly over all $\alpha \in (0,1)$. In particular, for $\delta \in (0,1)$, we seek bounds $\overline{\text{FCP}}^{w,\delta,\alpha}_{\text{fix}}$ and $\overline{\text{FCP}}^{w,\delta,\alpha}_{\text{uniform}}$ such that

$$\mathbb{P}\left(\text{FCP}^{w,(n)}_m(\alpha) \le \overline{\text{FCP}}^{w,\delta,\alpha}_{\text{fix}}\right) \ge 1-\delta, \text{for fixed } \alpha \in (0,1), \tag{7}$$

$$\mathbb{P}\left(\forall \alpha \in (0,1), \text{FCP}^{w,(n)}_m(\alpha) \le \overline{\text{FCP}}^{w,\delta,\alpha}_{\text{uniform}}\right) \ge 1-\delta. \tag{8}$$

In addition to this forward (direct) problem, one may consider the *inverse FCP problem*, where a target FCP level β is specified and the goal is to choose a conformal level that achieves this target with high probability. Specifically, we seek $\widehat{\text{FCP}}^{w,\delta,\beta}_{\text{fix}}$ and $\widehat{\text{FCP}}^{w,\delta,\beta}_{\text{uniform}}$ satisfying

$$\mathbb{P}\left(\text{FCP}^{w,(n)}_m\left(\widehat{\text{FCP}}^{w,\delta,\beta}_{\text{fix}}\right) \le \beta\right) \ge 1-\delta, \text{for fixed } \beta \in (0,1), \tag{9}$$

$$\mathbb{P}\left(\forall \beta \in (0,1), \text{FCP}^{w,(n)}_m\left(\widehat{\text{FCP}}^{w,\delta,\beta}_{\text{uniform}}\right) \le \beta\right) \ge 1-\delta. \tag{10}$$

We note that the marginal guarantee given in Theorem 2.5 is equivalent to $\mathbb{E}\left[\text{FCP}^{w_0,(n)}_m(\alpha)\right] \le \alpha$, which constitutes a weaker control objective than the high-probability guarantees in equation 7 and 8.

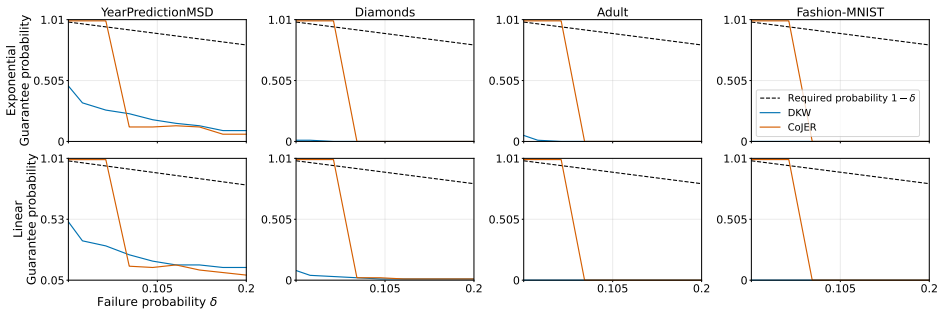

Figure 1: Empirical forward pass rates of the uniform DKW (Gazin et al., 2024) and CoJER (Blain et al., 2025) bounds under Covariate Shift.

3 Limitations of Existing FCP Bounds under Distribution Shift

Prior work on controlling the False Coverage Proportion (FCP) in Conformal Prediction (CP) has established bounds analogous to those in equation 8 and equation 10. For example, Gazin et al. (2024) apply the Dvoretzky-Kiefer-Wolfowitz inequality to control the empirical distribution of conformal errors uniformly over the miscoverage level, thereby obtaining an FCP bound under exchangeability. Separately, Blain et al. (2025) propose CoJER, which builds on the joint error rate framework of Blanchard et al. (2020) to derive an upper bound on the FCP. They further combine CoJER with a template-based procedure to invert this guarantee, yielding a bound analogous to equation 10.

However, neither approach accounts for Distribution Shift between the calibration and test data. Both are developed for standard, unweighted CP and rely on exchangeability between the calibration and test samples. In our setting, this assumption is generally violated because the two samples may follow different distributions. As a result, the corresponding guarantees need not remain valid under Distribution Shift. We illustrate this limitation empirically across multiple datasets, miscoverage levels, and Covariate Shift settings, showing that these bounds can substantially underestimate the realized FCP. Further details of the experimental setup are provided in Section D.2.

Results. The results are presented in Figure 1. Across both classification and regression datasets, the bounds of Gazin et al. (2024) and Blain et al. (2025) have empirical pass rates below target level. In several settings, the pass rates deteriorate sharply as δ increases and even reach zero, indicating systematic violations of the nominal guarantee under Distribution Shift. This demonstrates that their guarantees need not remain valid under Distribution Shift, motivating our development of FCP bounds that explicitly account for such shifts.

4 Main Theoretical Guarantees

In this section, we derive our main theoretical guarantees for the False Coverage Proportion (FCP) of Weighted Conformal Prediction (CP) under Distribution-Shift.

Notation. Let $(X_1, Y_1) \sim P_{\text{cal}}$ be a generic calibration point, and let $S = s(X_1, Y_1)$ denote its corresponding score. We denote the mean of the weight function w by

$$\mu_w := \mathbb{E}_{\text{cal}}[w(X_1)].$$

Under Assumption 2.3, μ_w is positive and finite. We additionally denote $a_n := \dfrac{w_\infty}{n}$. For $t \in \mathbb{R}$, define the weighted calibration distribution function by

$$A_w(t) := \mathbb{E}_{\text{cal}}\left[w(X_1)\mathbf{1}\{S \le t\}\right].$$

We further define the population-weighted CDF, incorporating the finite-n correction induced by the additional weight w_∞, as $F_{w,n}(t) := A_w(t)(\mu_w + a_n)^{-1}$. Finally, let $(X_{n+1}, Y_{n+1}) \sim P_{\text{test}}$ be a generic test point and let $T = s(X_{n+1}, Y_{n+1})$. We define the population CDF of the ideal weighted p-value under the test distribution by

$$G_{w,n}(\beta) := P_{\text{test}}\left(F_{w,n}(T) \ge 1 - \beta\right), \qquad \beta \in \mathbb{R}.$$

Main Results. We begin by establishing a finite-sample probability bound for $\text{FCP}_m^{w,(n)}(\alpha)$ at a fixed miscoverage level α in the following theorem.

Theorem 4.1 (Fixed-α FCP bound). *Fix $\alpha \in (0,1)$ and let $\delta_{\text{cal}}, \delta_{\text{test}} \in (0,1)$. Define*

$$\Delta_{\text{fix}} := \varepsilon_w(\mu_w + a_n)^{-1}\sqrt{(2n)^{-1}\log(1/\delta_{\text{cal}})}, \text{ and } \varepsilon_{\text{fix}} := \sqrt{(2m)^{-1}\log(1/\delta_{\text{test}})}.$$

Then, under Assumption 2.2 and Assumption 2.3, we have

$$\mathbb{P}\left(\text{FCP}_m^{w,(n)}(\alpha) \le G_{w,n}\left(\alpha + \Delta_{\text{fix}}\right) + \varepsilon_{\text{fix}}\right) \ge 1 - \delta_{\text{cal}} - \delta_{\text{test}}.$$

The proof for Theorem 4.1 is given in Appendix B.1. We next strengthen this result by deriving a finite-sample bound that holds simultaneously for all $\alpha \in (0,1)$ as follows.

Theorem 4.2 (Uniform-in-α FCP bound). *Let $\delta_+, \delta_-, \delta_{\text{test}} \in (0,1)$, and let $d_n := \mu_w + a_n$. Define*

$$r_+ := \varepsilon_w(2n)^{-1/2}\left(\sqrt{\pi/4} + \sqrt{\log(1/\delta_+)}\right), r_- := \varepsilon_w(2n)^{-1/2}\sqrt{\log(1/\delta_-)},$$

$$\varepsilon_{\text{unif}} := \sqrt{\frac{\log(1/\delta_{\text{test}})}{2m}}, \text{ and } \Delta_{\text{unif}} := \begin{cases} (d_n r_+ + \mu_w r_-)(d_n(d_n - r_-))^{-1}, & \text{if } r_- < d_n, \\ 1, & \text{if } r_- \ge d_n. \end{cases},$$

Then, under Assumption 2.2 and Assumption 2.3, we have

$$\mathbb{P}\left(\forall \alpha \in (0,1), \quad \text{FCP}_m^{w,(n)}(\alpha) \le G_{w,n}\left(\alpha + \Delta_{\text{unif}}\right) + \varepsilon_{\text{unif}}\right) \ge 1 - \delta_+ - \delta_- - \delta_{\text{test}}.$$

The proof for Theorem 4.2 is given in Appendix B.2.

5 Practical Estimation and The Inverse FCP Problem

Although Theorems 4.1 and 4.2 provide finite-sample guarantees, their direct practical use is limited because the CDF $G_{w,n}$ is generally unknown. Moreover, $G_{w,n}$ need not be invertible, so these results do not directly allow us to solve the inverse FCP problem. We will now address these limitations.

5.1 Covariate Shift

We begin with the classical setting of *Covariate Shift*, one of the simplest forms of One-Phase Distribution Shift. In this setting, we take the weight function to be the oracle weight (see equation 5), equipped with a fixed test weight w_∞. Under this setting, the following result holds.

Lemma 5.1. *Assume $w = w_0$, the oracle weight function. Under Assumption 2.2, Assumption 2.3 and Assumption 2.4, for every $\eta \in (0,1)$, we have $G_{w_0,n}(\eta) \le \eta$.*

By applying Lemma 5.1 to the intractable CDF $G_{w_0,n}$ appearing in Theorems 4.1 and 4.2, one obtains explicit and tractable FCP guarantees for Weighted CP under Covariate Shift, both at a fixed miscoverage level α and uniformly over $\alpha \in (0,1)$. Moreover, since Lemma 5.1 bounds $G_{w_0,n}$ by the identity map, which is invertible, these guarantees can be inverted explicitly to obtain miscoverage levels satisfying the fixed and uniform inverse FCP objectives in equation 9 and equation 10, respectively. This yields the practical bounds under Covariate Shift, summarized in the second row of Table 1.

Conservativeness induced by Lemma 5.1. Applying Lemma 5.1 introduces an additional approximation error into the bounds of Theorems 4.1 and 4.2. We further discuss and quantify this error empirically in Appendix D.3 and show that it is negligible in practice.

Choice of $\boldsymbol{\delta_+, \delta_-, \delta_{\text{cal}}, \delta_{\text{test}}}$. In practice, the user specifies only an overall δ, with $\delta = \delta_{\text{cal}} + \delta_{\text{test}}$ for the fixed-α guarantee and $\delta = \delta_+ + \delta_- + \delta_{\text{test}}$ for the uniform-in-α guarantee. Under Covariate Shift, the characterization provided by Lemma 5.1 allows us to explicitly optimize the allocation of this overall δ among $\delta_+, \delta_-, \delta_{\text{cal}}$, and δ_{test}. We derive these optimal choices in Appendix C.4.

5.2 General One-Phase Distribution Shift

We now introduce practical estimators for the theoretical FCP bounds established in Theorem 4.1 and Theorem 4.2 under general OPDS, namely, for any distribution shift satisfying Assumption 2.2.

As mentioned above, under this general setting, the CDF $G_{w,n}$ is typically intractable and therefore cannot be evaluated directly in practice. To address this issue, we propose Algorithm 1, which constructs an empirical estimator of $G_{w,n}$, denoted by $\widehat{G}^{\text{IW}}_{w,n}$, using only the available calibration data together with the weight function w. To establish that this estimator provides a consistent approximation to the true $G_{w,n}$, we first introduce an additional assumption on w as follows.

Table 1: Summary of theoretical finite-sample FCP guarantees and their practical estimators. The first row presents the general bounds, while the second and third give the corresponding practical forms under Covariate Shift (CovShift) and General One-Phase Distribution Shift (GenOPDS).

Form	$\overline{\mathrm{FCP}}_{\mathrm{fix}}^{w,\delta,\alpha}$ (equation 7)	$\overline{\mathrm{FCP}}_{\mathrm{unif}}^{w,\delta,\alpha}$ (equation 8)	$\widehat{\mathrm{FCP}}_{\mathrm{fix}}^{w,\delta,\beta}$ (equation 9)	$\widehat{\mathrm{FCP}}_{\mathrm{unif}}^{w,\delta,\beta}$ (equation 10)
Finite-sample guarantees	$G_{w,n}(\alpha+\Delta_{\mathrm{fix}})+\varepsilon_{\mathrm{fix}}$	$G_{w,n}(\alpha+\Delta_{\mathrm{unif}})+\varepsilon_{\mathrm{unif}}$	$\left(G^{+}_{w,n}(\beta-\varepsilon_{\mathrm{fix}})\right.$ $\left.-\Delta_{\mathrm{fix}}-\eta\right)_{+}$	$\left(G^{+}_{w,n}(\beta-\varepsilon_{\mathrm{unif}})\right.$ $\left.-\Delta_{\mathrm{unif}}-\eta\right)_{+}$
CovShift	$\alpha+\Delta_{\mathrm{fix}}+\varepsilon_{\mathrm{fix}}$	$\alpha+\Delta_{\mathrm{unif}}+\varepsilon_{\mathrm{unif}}$	$\left(\beta-\Delta_{\mathrm{fix}}-\varepsilon_{\mathrm{fix}}\right)_{+}$	$\left(\beta-\Delta_{\mathrm{unif}}-\varepsilon_{\mathrm{unif}}\right)_{+}$
GenOPDS	$\widehat{G}^{\mathrm{IW}}_{w,n}(\alpha+\Delta_{\mathrm{fix}})$ $+\varepsilon_{\mathrm{fix}}$	$\widehat{G}^{\mathrm{IW}}_{w,n}(\alpha+\Delta_{\mathrm{unif}})$ $+\varepsilon_{\mathrm{unif}}$	$\left((\widehat{G}^{\mathrm{IW}}_{w,n})^{+}(\beta-\varepsilon_{\mathrm{fix}})\right.$ $\left.-\Delta_{\mathrm{fix}}-\eta\right)_{+}$	$\left((\widehat{G}^{\mathrm{IW}}_{w,n})^{+}(\beta-\varepsilon_{\mathrm{unif}})\right.$ $\left.-\Delta_{\mathrm{unif}}-\eta\right)_{+}$

Assumption 5.2. *The weight function w satisfies*

$$P_{\mathrm{test}}(T\le t)=\mu_w^{-1}\mathbb{E}_{\mathrm{cal}}\left[w(X)\mathbf{1}\{S\le t\}\right],\qquad \textit{for every } t\in\mathbb{R}.$$

Discussion of Assumption 5.2. This Assumption requires only that the weight function w sufficiently account for the Distribution Shift at the level of the score distribution. It is satisfied under Covariate Shift with the oracle importance weight and is natural in practice, where weighting schemes are designed or trained to precisely account for such Distribution Shifts.

We now establish the uniform consistency of Algorithm 1 in the following theorem.

Theorem 5.3 (Consistency of Algorithm 1)**.** *Suppose Assumptions 2.2, 2.3, and 5.2 hold, and let $\widehat{G}^{\mathrm{IW}}_{w,n}$ be the output of Algorithm 1. Then, we have*

$$\sup_{\beta\in(0,1)}\left|\widehat{G}^{\mathrm{IW}}_{w,n}(\beta)-G_{w,n}(\beta)\right|=O_{\mathbb{P}}\left(n^{-1}\right).$$

The proof of Theorem 5.3 is given in Appendix B.4. By replacing the unknown $G_{w,n}$ with its estimator $\widehat{G}^{\mathrm{IW}}_{w,n}$, we obtain practical plug-in estimates of the finite-sample FCP bounds under general OPDS. The resulting expressions are summarized in the first two columns of the third row of Table 1.

Remark 5.4. Algorithm 1 has computational complexity $\mathcal{O}(n\log n)$, and therefore introduces negligible overhead. When an independent labeled hold-out sample from the test distribution is available, $G_{w,n}$ can also be estimated under weaker assumptions; see Appendix C.2.

5.3 The Inverse FCP Problem

Theorem 4.1 and 4.2 address only the direct problem in equations 7 and 8, respectively. We now address the corresponding inverse FCP problems in equations 9 and 10, where the goal is to select a miscoverage that achieves a prescribed target FCP.

Since $G_{w,n}$ need not be invertible, for every real number $u<1$, we first define its generalized inverse by

$$G^{+}_{w,n}(u):=\inf\left\{\tau\in[0,1]:G_{w,n}(\tau)>u\right\}.$$

Using this generalized inverse together with the direct guarantees of Theorem 4.1 and Theorem 4.2, we obtain theoretical guarantees for the inverse FCP problem. The fixed-β and uniform-in-β results are given in Corollary 5.5 and Corollary 5.6, respectively.

Corollary 5.5. *Fix $\beta\in(0,1)$ and let Δ_{fix} and $\varepsilon_{\mathrm{fix}}$ be defined as in Theorem 4.1. Then, under Assumption 2.2 and Assumption 2.3, for any $\eta>0$, we have*

$$\mathbb{P}\left(\mathrm{FCP}_m^{w,(n)}\left[\left(G^{+}_{w,n}(\beta-\varepsilon_{\mathrm{fix}})-\Delta_{\mathrm{fix}}-\eta\right)_{+}\right]\le\beta\right)\ge 1-\delta_{\mathrm{cal}}-\delta_{\mathrm{test}}.$$

Corollary 5.6. *Let $\delta_{+},\delta_{-},\delta_{\mathrm{test}}\in(0,1)$, and let Δ_{unif} and $\varepsilon_{\mathrm{unif}}$ be defined as in Theorem 4.2. Then, under Assumption 2.2 and Assumption 2.3, for any $\eta>0$, we have*

$$\mathbb{P}\Big(\forall\beta\in(0,1),\quad \mathrm{FCP}_m^{w,(n)}\left[\left(G^{+}_{w,n}(\beta-\varepsilon_{\mathrm{unif}})-\Delta_{\mathrm{unif}}-\eta\right)_{+}\right]\le\beta\Big)\ge 1-\delta_{+}-\delta_{-}-\delta_{\mathrm{test}}.$$

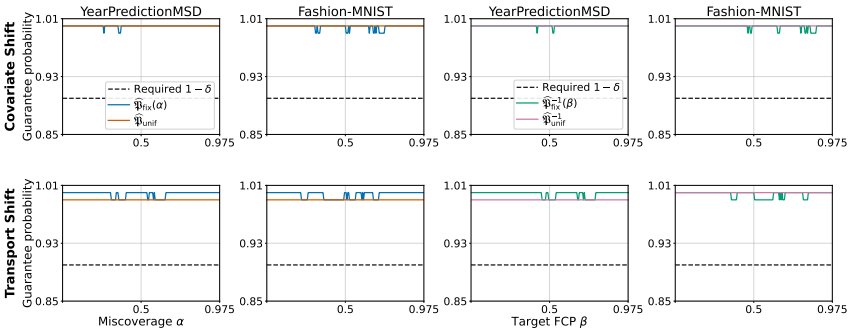

Figure 2: Empirical evaluation of the estimated direct and inverse FCP bounds summarized in Table 1 on the YearPredictionMSD and Fashion-MNIST datasets. The top row corresponds to Covariate Shift, while the second row corresponds to Score-Transport Shift (STS).

The proofs of Corollary 5.5 and Corollary 5.6 are provided in Appendix B.5 and Appendix B.6, respectively. In practice, the slack parameter η appearing in both corollaries may be chosen arbitrarily small (e.g., $\eta = 10^{-100}$), so that its contribution to the resulting miscoverage level is negligible.

Similar to $G_{w,n}$, its generalized inverse $G^{+}_{w,n}$ is also intractable and cannot be evaluated directly in practice. We therefore propose Algorithm 2, which uses only the available calibration data together with the weight function w to construct an estimator of $G^{+}_{w,n}$, denoted by $\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}$. Under Assumption 5.2, we establish the consistency of this estimator in the following theorem.

Theorem 5.7 (Consistency of Algorithm 2)**.** *Suppose Assumptions 2.2, 2.3, and 5.2 hold, and let* $\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}$ *be the output of Algorithm 2. Then, we have*

$$\sup_{u\in[0,1)} \left| \left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}(u) - G^{+}_{w,n}(u) \right| = O_{\mathbb{P}}\left(n^{-1}\right).$$

The proof of Theorem 5.7 is provided in Appendix B.7. By replacing the unknown generalized inverse $G^{+}_{w,n}$ in Corollary 5.5 and Corollary 5.6 with its estimator $\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}$, we obtain practical plug-in estimates for the inverse FCP problem, both for fixed β and uniformly over $\beta \in (0,1)$. The resulting expressions are summarized in the last two columns of the third row of Table 1.

6 Experimental Results

We provide numerical evaluations of the estimated FCP bounds summarized in Table 1 under various settings. We begin with a simple Covariate Shift (CS) setting, where we empirically validate Lemma 5.1 together with the optimal choices of δ_{+}, δ_{-}, δ_{cal}, and δ_{test} derived in Appendix C.4. We then consider the more challenging OPDS setting of Score-Transport Shift (STS), where we demonstrate the effectiveness of Algorithms 1 and 2. Finally, we conduct an asymptotic study to investigate the behavior of our proposed FCP bounds as the calibration and test sample sizes increase.

Overall Setup. We evaluate our proposed method on three classification datasets: Fashion-MNIST (Xiao et al., 2017), Electricity (Harries, 1999), and Adult (Becker & Kohavi, 1996); and three regression datasets: Allstate Claims Severity (Allstate Insurance Company, 2016), YearPredictionMSD (Bertin-Mahieux et al., 2011), and Diamonds (Wickham, 2016). All datasets are obtained through OpenML (Vanschoren et al., 2013). We consider two types of Distribution Shift: Covariate Shift and Score-Transport Shift (STS) (see Appendix D.1). For the asymptotic study, we additionally construct simulated datasets following Setting 2 of Barber et al. (2023). Across all experiments, we use histogram-based gradient-boosted trees (Friedman, 2001), implemented via scikit-learn (Pedregosa et al., 2011), as the base predictor. For Covariate Shift, we use the exponential tilting weight $w(x) \propto \exp\{\gamma z(x)\}$, where $z(x)$ is the standardized score-informed projection defined in Appendix D.1. Due to space constraints, further experimental details are deferred to Appendix D.

Ablation Study. We further conduct ablation studies on the underlying model $\hat{f}$, the weight function w, inference time, and efficiency in Appendices D.4, D.5, D.7, and D.6, respectively.

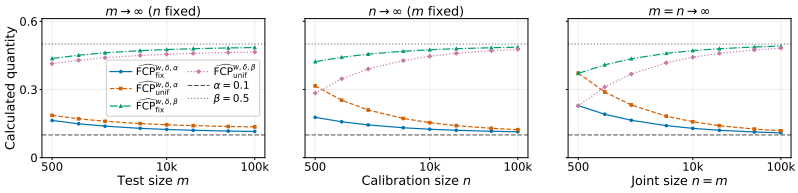

Figure 3: Asymptotic behavior of the direct and inverse FCP bounds in Table 1 under three regimes: fixed n with $m \to \infty$, fixed m with $n \to \infty$, and $m = n \to \infty$.

6.1 Verification of Theoretical Bounds

Setup. To validate the direct FCP guarantees, we fix m and n, vary $\alpha \in [0, 1]$, and sample k test subsets of size m. For each subset $r \in [k]$, we compute $\mathrm{FCP}_r(\alpha)$ via equation 6 and the corresponding fixed and uniform bounds. We report the empirical pass rates $\mathfrak{R}_{\mathrm{fix}}(\alpha) := k^{-1} \sum_{r=1}^{k} \mathbf{1}\left\{\mathrm{FCP}_r(\alpha) \le \overline{\mathrm{FCP}}^{w,\delta,\alpha}_{\mathrm{fix},r}\right\}$ and $\mathfrak{R}_{\mathrm{unif}} := k^{-1} \sum_{r=1}^{k} \mathbf{1}\left\{\mathrm{FCP}_r(\alpha) \le \overline{\mathrm{FCP}}^{w,\delta,\alpha}_{\mathrm{unif},r},\ \forall\alpha\right\}$, and compare them with $1 - \delta$. For the inverse guarantees, we analogously vary $\beta \in [0, 1]$, compute $\widehat{\mathrm{FCP}}^{w,\delta,\beta}_{\mathrm{fix},r}$ and $\widehat{\mathrm{FCP}}^{w,\delta,\beta}_{\mathrm{unif},r}$, and report $\mathfrak{R}^{-1}_{\mathrm{fix}}(\beta) := k^{-1} \sum_{r=1}^{k} \mathbf{1}\left\{\mathrm{FCP}_r\left(\widehat{\mathrm{FCP}}^{w,\delta,\beta}_{\mathrm{fix},r}\right) \le \beta\right\}$ and $\mathfrak{R}^{-1}_{\mathrm{unif}} := k^{-1} \sum_{r=1}^{k} \mathbf{1}\left\{\mathrm{FCP}_r\left(\widehat{\mathrm{FCP}}^{w,\delta,\beta}_{\mathrm{unif},r}\right) \le \beta,\ \forall\beta\right\}$, again comparing them with $1 - \delta$.

Results for CS. Results are given in the first row of Figure 2 and Appendix Figures 4, 5. Both proposed estimated bounds, $\overline{\mathrm{FCP}}^{w,\delta,\alpha}_{\mathrm{fix}}$ and $\overline{\mathrm{FCP}}^{w,\delta,\alpha}_{\mathrm{unif}}$, achieve $\mathfrak{R}_{\mathrm{fix}}(\alpha)$ and $\mathfrak{R}_{\mathrm{unif}}$ above the target level $1 - \delta$ across all α. This supports Lemma 5.1, together with the choices of δ_+, δ_-, δ_{cal}, and δ_{test} derived in Appendix C.4. For the inverse FCP problem, both $\widehat{\mathrm{FCP}}^{w,\delta,\beta}_{\mathrm{fix}}$ and $\widehat{\mathrm{FCP}}^{w,\delta,\beta}_{\mathrm{unif}}$ achieve $\mathfrak{R}^{-1}_{\mathrm{fix}}(\beta)$ and $\mathfrak{R}^{-1}_{\mathrm{unif}}$ above $1 - \delta$ across all levels β, supporting the proposed estimated inverse bounds.

Results for STS. Results are given in Figure 2 and Appendix Figures 4, 5. Both $\overline{\mathrm{FCP}}^{w,\delta,\alpha}_{\mathrm{fix}}$ and $\overline{\mathrm{FCP}}^{w,\delta,\alpha}_{\mathrm{unif}}$ achieve $\mathfrak{R}_{\mathrm{fix}}(\alpha)$ and $\mathfrak{R}_{\mathrm{unif}}$ above $1 - \delta$ across all α, while the inverse procedures achieve $\mathfrak{R}^{-1}_{\mathrm{fix}}(\beta)$ and $\mathfrak{R}^{-1}_{\mathrm{unif}}$ above $1 - \delta$ across all β. These results support the effectiveness of Algorithms 1 and 2 in estimating $G_{w,n}$ and $G^{+}_{w,n}$ and recovering the corresponding FCP guarantees under STS.

6.2 Asymptotic Study

Setup. To study the asymptotic behavior of the FCP bounds in Table 1, we fix a miscoverage α and simulate a Covariate Shift dataset following Setting 2 of Barber et al. (2023). We consider three asymptotic regimes: (i) fixed n with $m \to \infty$, (ii) fixed m with $n \to \infty$, and (iii) $m = n \to \infty$. For each pair (m, n) within each of the aforementioned regimes, we evaluate the direct bounds $\overline{\mathrm{FCP}}^{w,\delta,\alpha}_{\mathrm{fix}}$ and $\overline{\mathrm{FCP}}^{w,\delta,\alpha}_{\mathrm{unif}}$, together with the inverse bounds $\widehat{\mathrm{FCP}}^{w,\delta,\beta}_{\mathrm{fix}}$ and $\widehat{\mathrm{FCP}}^{w,\delta,\beta}_{\mathrm{unif}}$, as defined in Table 1.

Results. The results are shown in Figure 3. Across all regimes, the direct bounds decrease, and the inverse bounds increase as sample sizes grow. When n is fixed and $m \to \infty$, a residual gap remains due to calibration uncertainty. When m is fixed and $n \to \infty$, the bounds tighten substantially, particularly for the uniform guarantees, but a residual gap remains due to the non-vanishing test-sample uncertainty. When $m = n \to \infty$, all bounds converge toward their respective nominal levels. These behaviors are consistent with our theoretical bounds, in which the calibration- and test-sample slack terms vanish only when the corresponding sample sizes diverge. Overall, the results demonstrate the asymptotic tightening of both the direct and inverse FCP guarantees.

7 Conclusion

In this work, we show that existing False Coverage Proportion (FCP) bounds can fail under Distribution Shift (DS). We then derive finite-sample FCP guarantees under Covariate Shift and general one-phase DS, along with results for the inverse FCP problem. We further develop practical algorithms for estimating the quantities required to evaluate these bounds. Lastly, we validate the proposed guarantees across multiple datasets and shift settings, and study their asymptotic behavior.

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

Allstate Insurance Company. Allstate claims severity. Kaggle Competition, 2016. URL https://www.kaggle.com/c/allstate-claims-severity.

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 and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Liviu Aolaritei, Julie Zhu, Oliver Wang, Michael Jordan, and Youssef Marzouk. Conformal prediction under lévy-prokhorov distribution shifts: Robustness to local and global perturbations. *Advances in Neural Information Processing Systems*, 38:63811–63840, 2026.

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

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL https://doi.org/10.24432/C5XW20.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 591–596, Miami, Florida, USA, 2011.

Alexandre Blain, Bertrand Thirion, and Pierre Neuvial. False coverage proportion control for conformal prediction. In *Forty-second International Conference on Machine Learning*, 2025.

Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. Post hoc confidence bounds on false positives using reference families. 2020.

Dimitrios Boursinos and Xenofon Koutsoukos. Selective classification of sequential data using inductive conformal prediction. In *2022 IEEE international conference on assured autonomy (ICAA)*, pp. 46–55. IEEE, 2022.

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.

Alvaro Correia and Christos Louizos. Non-exchangeable conformal prediction with optimal transport: Tackling distribution shift with unlabeled data. *Advances in Neural Information Processing Systems*, 38:157924–157960, 2026.

Jean-Baptiste Fermanian, Pierre Humbert, and Gilles Blanchard. Transductive conformal inference for full ranking. *Advances in Neural Information Processing Systems*, 38:6061–6093, 2026.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

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.

Ulysse Gazin. Asymptotics for conformal inference. *arXiv preprint arXiv:2409.12019*, 2024.

Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics*, pp. 1504–1512. PMLR, 2024.

Ulysse Gazin, Ruth Heller, Ariane Marandon, and Etienne Roquain. Selecting informative conformal prediction sets with false coverage rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):909–929, 2025.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Michael Harries. Splice-2 comparative evaluation: Electricity pricing. Technical report, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia, 1999.

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*, 2025.

Huiqun Huang, Sihong He, and Fei Miao. Cuqds: conformal uncertainty quantification under distribution shift for trajectory prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 17422–17430, 2025.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pp. 14153–14177. PMLR, 2023.

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

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.

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.

Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin Bonilla, and Terry Lyons. Distribution regression for sequential data. In *International Conference on Artificial Intelligence and Statistics*, pp. 3754–3762. PMLR, 2021.

Yi Liu, Alexander W Levis, Sharon-Lise Normand, and Larry Han. Multi-source conformal inference under distribution shift. *Proceedings of machine learning research*, 235:31344, 2024.

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.

Thomas Mortier, Cas Decancq, and Diego G Miralles. Evaluating conformal prediction in weather forecasting. In *The 37th Benelux Conference on Artificial Intelligence and the 34th Belgian Dutch Conference on Machine Learning*, 2025.

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.

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

Harris Papadopoulos. A cross-conformal predictor for multi-label classification. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 241–250. Springer, 2014.

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.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. Pac prediction sets under covariate shift. *arXiv preprint arXiv:2106.09848*, 2021.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pp. 27907–27947. PMLR, 2023.

Mehrdad Pournaderi and Yu Xiang. Training-conditional coverage bounds under covariate shift. *arXiv preprint arXiv:2405.16594*, 2024.

Derck WE Prinzhorn, Thijmen Nijdam, Putri A van der Linden, and Alexander Timans. Conformal time series decomposition with component-wise exchangeability. *arXiv preprint arXiv:2406.16766*, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Ziang Song, Ying Jin, and Emmanuel J Candès. Everywhere valid bounds on false discovery proportions in conformal inference. *arXiv preprint arXiv:2605.20726*, 2026.

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.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198.

Janette Vazquez and Julio C Facelli. Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3):241–252, 2022.

Guillermo Villate-Castillo, Javier Del Ser, and Borja Sanz. A collaborative content moderation framework for toxicity detection based on conformalized estimates of annotation disagreement. *arXiv preprint arXiv:2411.04090*, 2024.

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. Transductive conformal predictors. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 348–360. Springer, 2013.

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

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.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. doi: 10.1007/978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

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.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL https://arxiv.org/abs/1708.07747.

Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. Wasserstein-regularized conformal prediction under general distribution shift. In *International Conference on Learning Representations*, volume 2025, pp. 15503–15528, 2025.

Fatima Rabia Yapicioglu, Meltem Aksoy, Tuwe Löfström, Fabio Vitali, and Alberto Rigenti. Conformasegment: A conformal prediction-based, uncertainty-aware, and model-agnostic explainability framework for time-series forecasting. In *World Conference on Explainable Artificial Intelligence*, pp. 218–242. Springer, 2025.

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
$\mathbb{R}_{>0}$	Set of positive real numbers
$\|\cdot\|_d$	Euclidean norm in $\mathbb{R}^d$
$\inf\{\cdot\}$	Infimum (greatest lower bound)
$[m]$	The set $\{1, 2, \ldots, m\}$
$[n, m]$	The set $\{n, n+1, \ldots, m\}$
$(x)_+$	The positive part of x, $\max\{x, 0\}$
$\mathbf{1}\{\cdot\}$	Indicator function
$\mathbb{E}[\cdot]$	Expectation of a Random Variable
$O(\cdot)$	Big-O notation
$O_{\mathbb{P}}(\cdot)$	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
$\mathcal{C}^{\alpha}(\cdot)$	Set-valued predictor at miscoverage level α
$\mathcal{C}^{w,\alpha}(\cdot)$	Weighted set-valued predictor at miscoverage level α and weight w
$s(\cdot)$	Score function
$p^{w,(n)}(x, y)$	Weighted p-value at datapoint (x, y).
$\mathrm{FCP}_m^{w,(n)}(\alpha)$	Weighted False Coverage Proportion at miscoverage level α.
$\delta(\cdot)$	Dirac delta function
$Q_{1-\alpha}(\cdot)$	$(1-\alpha)$-quantile
$w(\cdot)$	Calibration weight function
$w_0(\cdot)$	Oracle weight function
w_∞	Test weight
P_{cal}	Calibration Distribution
P_{test}	Test Distribution
F_{cal}	Calibration score C.D.F
F_{test}	Test score C.D.F
ε_w	upper bound of w
μ_w	Mean of the weight function w
$A_w(t)$	Weighted calibration distribution function $\mathbb{E}_{\mathrm{cal}}[w(X_1)\mathbf{1}\{S \le t\}]$
$F_{w,n}(t)$	Population weighted CDF $(A_w(t))/(\mu_w + a_n)$.
$G_{w,n}(\beta)$	Population CDF of the ideal weighted p-value under the test distribution, $P_{\mathrm{test}}(F_{w,n}(T) \ge 1-\beta)$
$G^+_{w,n}(u)$	Generalized Inverse of $G_{w,n}$
$\widehat{G}^{\mathrm{IW}}_{w,n}$	Estimation of $G_{w,n}$ via Algorithm 1
$\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^+$	Estimation of $G^+_{w,n}$ via Algorithm 2

Appendix of "Conformal Prediction under Distribution Shift: False Coverage Proportion Guarantees"

Table of Contents

A Background **17**

A.1 Conformal Prediction 17

A.2 False Coverage Proportion 22

A.3 Distribution Shift and Weighted Conformal Prediction 26

B Theoretical Proofs **30**

B.1 Proof of the Forward FCP Bound for Fixed Miscoverage - Theorem 4.1 30

B.2 Proof of the Uniform Forward FCP Bound over Miscoverage - Theorem 4.2 31

B.3 Proof of the Bound on $G_{w,n}$ under Covariate Shift - Lemma 5.1 34

B.4 Proof of the Consistency of Algorithm 1 - Theorem 5.3 34

B.5 Proof of the Inverse FCP Bound for Fixed Miscoverage- Corollary 5.5 37

B.6 Proof of the Inverse FCP Bound Uniformly over Miscoverage - Corollary 5.6 . . . 37

B.7 Proof of the Consistency of Algorithm 2 - Theorem 5.7 38

B.8 Proof of the Consistency of Algorithm 3 - Theorem C.2 40

B.9 Proof of the Consistency of Algorithm 4 - Theorem C.3 42

C Further Discussion and Analysis **43**

C.1 On the Convention of Fixed versus Adaptive Test Weights 43

C.2 Relaxing Assumption 5.2 Using Holdout Test Data 44

C.3 False Coverage Proportion under General Distribution Shift 48

C.4 Optimal Choice of Confidence Parameters under Covariate Shift 49

D Experiment Details **51**

D.1 Construction of Covariate and Score-Transport Shifts 51

D.2 More on Limitations of Existing FCP Bounds under Distribution Shift 53

D.3 Empirical Evaluation of Lemma 5.1 55

D.4 Ablation on Underlying Models $\hat{f}$ 55

D.5 Ablation on Weight Function 55

D.6 Ablation on Prediction Efficiency 56

D.7 Runtime Analysis 57

D.8 Computation Infrastructure 57

A Background

A.1 Conformal Prediction

Motivation. Modern machine learning methods have achieved remarkable predictive performance across a wide variety of applications, ranging from structured-data problems and large-scale recommender systems to scientific and engineering tasks. Nevertheless, even highly accurate models are subject to prediction error, while their outputs are often reported as point estimates with no accompanying guarantee on their reliability. In applications where predictions are used to support consequential decisions, such uncertainty can be particularly problematic. For example, an incorrect but seemingly confident prediction in healthcare decision support, automated financial systems, or safety-critical monitoring may result in substantial downstream costs.

A central challenge is that conventional notions of model confidence do not necessarily provide a faithful characterization of predictive uncertainty. Such scores may be poorly calibrated, depend strongly on modeling assumptions, and deteriorate when the distribution of the observed inputs differs from that encountered during training. This motivates the broader problem of equipping arbitrary predictive models with uncertainty quantification procedures that offer rigorous finite-sample guarantees under weak assumptions.

Conformal Prediction provides a principled solution to this problem by replacing point predictions with prediction sets designed to contain the unknown outcome with a user-specified probability. Rather than relying on model-specific confidence scores as calibrated probabilities, conformal methods compare the conformity of a new observation with that of observations in a reference sample and use this comparison to calibrate the resulting prediction set. The resulting guarantees are distribution-free and hold in finite samples, without requiring a parametric model or a specific distributional form. The central assumption underlying the standard conformal framework is exchangeability between the calibration and test observations, which we recall below.

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

$$(X_1, \dots, X_n) \overset{d}{=} (X_{\pi(1)}, \dots, X_{\pi(n)}). \tag{11}$$

This combination of finite-sample validity, distribution-free guarantees, and model-agnostic applicability makes Conformal Prediction an attractive framework for uncertainty quantification in modern machine learning systems, particularly when reliability is of primary importance.

Full (Transductive) Conformal Prediction (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021). Full Conformal Prediction provides the basic transductive formulation underlying many conformal methods. Consider the exchangeable observations

$$(X_1, Y_1), \dots, (X_n, Y_n), (X_0, Y_0), \tag{12}$$

where the first n labeled examples and the test covariate X_0 are observed, while Y_0 is unknown. For each candidate $y \in \mathcal{Y}$, augment the training sample with (X_0, y) and fit a predictor $\hat{f}^y$ using

$$(X_1, Y_1), \dots, (X_n, Y_n), (X_0, y). \tag{13}$$

The learning algorithm is assumed to be permutation-invariant with respect to its training observations. Given a score function $s(\cdot, \cdot; \cdot)$, define

$$s_i^y := s\big(X_i, Y_i; \hat{f}^y\big), \qquad i = 0, \dots, n, \tag{14}$$

where $i = 0$ denotes the candidate observation (X_0, y). Let

$$k := \lceil (n+1)(1-\alpha) \rceil, \tag{15}$$

and let $\widehat{q}^y$ be the k-th smallest value among $s_0^y, \dots, s_n^y$. The full Conformal Prediction set is

$$\mathcal{C}^\alpha(X_0) = \Big\{ y \in \mathcal{Y} : s_0^y \le \widehat{q}^y \Big\}. \tag{16}$$

Thus, a candidate y is retained whenever its score is not unusually large relative to the scores of the augmented sample. Equivalently, (X_0, y) is accepted if it is sufficiently conforming with the observed data under the predictor fitted using that candidate.

Algorithm 1 Estimation of $G_{w,n}(\cdot)$ without labeled test data

Require: Calibration pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, weight function w, additional weight w_∞.
Ensure: Estimate $\widehat{G}^{\mathrm{IW}}_{w,n}(\cdot)$ of $G_{w,n}(\cdot)$

1: Compute scores $S_i := s(X_i, Y_i)$ for $i \in [n]$
2: Compute the calibration weights

$$W_i := w(X_i), \qquad i = 1, \ldots, n.$$

3: Compute

$$\widehat{\mu}_w := \frac{1}{n}\sum_{i=1}^{n} W_i, \qquad a_n := \frac{w_\infty}{n}.$$

4: Sort the calibration scores:

$$S_{(1)} \le S_{(2)} \le \cdots \le S_{(n)}.$$

Let $W_{(i)}$ denote the weight corresponding to $S_{(i)}$.
5: Compute the cumulative weighted sums

$$C_0 := 0, \qquad C_k := \sum_{i=1}^{k} W_{(i)}, \qquad k = 1, \ldots, n.$$

6: **for** $r = 1, \ldots, n$ **do**
7: Find, with the convention that $S_{(0)} = -\infty$,

$$k_r := \max\{k \in \{0, 1, \ldots, n\} : S_{(k)} \le S_{(r)}\}.$$

8: Compute

$$\widehat{A}_w(S_{(r)}) := \frac{C_{k_r}}{n}.$$

9: Compute

$$\widehat{F}_{w,n}(S_{(r)}) := \frac{\widehat{A}_w(S_{(r)})}{\widehat{\mu}_w + a_n}.$$

10: **end for**
11: Return

$$\widehat{G}^{\mathrm{IW}}_{w,n} : \mathbb{R} \to [0, 1], \qquad \beta \mapsto \frac{\sum_{r=1}^{n} W_{(r)} \mathbf{1}\left\{\widehat{F}_{w,n}(S_{(r)}) \ge 1 - \beta\right\}}{\sum_{r=1}^{n} W_{(r)}}.$$

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*

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

The result follows from exchangeability of the augmented scores: when $y = Y_0$, the rank of the test score among $\{s_i^y\}_{i=0}^{n}$ is uniform.

Although full Conformal Prediction provides an exact finite-sample guarantee under exchangeability, it can be computationally expensive because the model must be refitted for each candidate y. This motivates computationally cheaper alternatives, most notably split Conformal Prediction.

Split (Inductive) Conformal Prediction. The main drawback of full Conformal Prediction is its computational cost, since the model must be refitted for every candidate response. Split Conformal Prediction avoids this expense by separating model training from calibration: the predictive model is fitted once, while a held-out calibration set is used to construct the uncertainty set.

Let $(X_i, Y_i)_{i=1}^{n}$ be exchangeable observations from an unknown distribution on $\mathcal{X} \times \mathcal{Y}$. For a target miscoverage level $\alpha \in (0, 1)$, the goal is to construct a set-valued predictor $\mathcal{C} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ satisfying

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0)\right) \ge 1 - \alpha, \tag{18}$$

Algorithm 2 Estimation of $\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}(\cdot)$ without labeled test data

Require: Calibration pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, weight function w, additional weight w_∞
Ensure: Step-function representation of $\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}(\cdot)$

1: Compute scores $S_i := s(X_i, Y_i)$ for $i \in [n]$
2: Compute the calibration weights

$$W_i \leftarrow w(X_i), \qquad i = 1, \ldots, n.$$

3: Compute

$$\widehat{\mu}_w := \frac{1}{n}\sum_{i=1}^{n} W_i, \qquad a_n := \frac{w_\infty}{n}.$$

4: Sort the calibration scores:

$$S_{(1)} \le S_{(2)} \le \cdots \le S_{(n)},$$

and let $W_{(r)}$ denote the weight corresponding to $S_{(r)}$.
5: Compute the cumulative weighted sums

$$C_0 \leftarrow 0, \qquad C_k \leftarrow \sum_{i=1}^{k} W_{(i)}, \qquad k = 1, \ldots, n.$$

6: **for** $r = 1, \ldots, n$ **do**
7: Find, with the convention $S_{(0)} = -\infty$,

$$k_r \leftarrow \max\left\{k \in \{0, \ldots, n\} : S_{(k)} \le S_{(r)}\right\},$$

8: Compute

$$\widehat{A}_w(S_{(r)}) := \frac{C_{k_r}}{n}.$$

9: Compute

$$\widehat{F}_{w,n}(S_{(r)}) := \frac{\widehat{A}_w(S_{(r)})}{\widehat{\mu}_w + a_n}.$$

10: Compute

$$\widehat{V}_r \leftarrow 1 - \widehat{F}_{w,n}(S_{(r)}).$$

11: **end for**
12: Reverse the order of the values $\widehat{V}_r$. For $j = 1, \ldots, n$, set

$$r_j \leftarrow n - j + 1, \qquad Q_j \leftarrow \widehat{V}_{r_j}.$$

13: Compute the reverse cumulative normalized weights

$$H_0 \leftarrow 0, \qquad H_j \leftarrow \frac{\sum_{\ell=1}^{j} W_{(r_\ell)}}{C_n}, \qquad j = 1, \ldots, n.$$

14: **return** the weighted step function

$$\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}(u) := \begin{cases} 0, & u \le 0, \\ Q_{k(u)}, & 0 < u < 1, \end{cases}$$

where

$$k(u) := \min\left\{j \in \{1, \ldots, n\} : H_j > u\right\}.$$

where (X_0, Y_0) is a test observation drawn from the same distribution.

Split Conformal Prediction partitions the available data into disjoint training and calibration subsets. A predictor $\hat{f} : \mathcal{X} \to \mathcal{Y}$ is fitted using only the training data and then held fixed. Let $\mathcal{I}_{\mathrm{cal}} \subset \{1, \ldots, n\}$

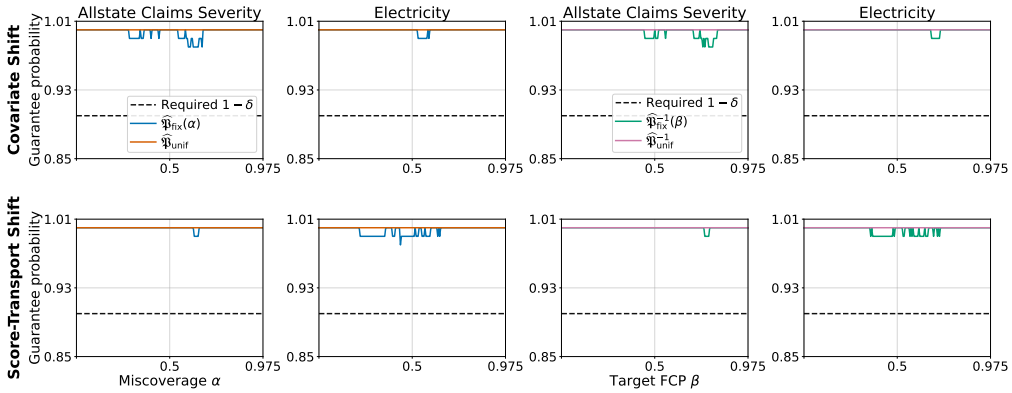

Figure 4: Empirical validation of the theoretical bounds for the direct and inverse FCP problems proposed in Table 1 on the Allstate Claims Severity and Electricity datasets. The top row corresponds to Covariate Shift, while the second row corresponds to score-transport shift (STS).

denote the calibration indices. Given a nonconformity score s, define

$$s_i := s\big(X_i, Y_i\big), \qquad i \in \mathcal{I}_{\text{cal}}. \tag{19}$$

These scores quantify how poorly each calibration response agrees with its corresponding prediction.

Let $n_{\text{cal}} := |\mathcal{I}_{\text{cal}}|$ and define $\hat{q}$ as the $\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil$ -th smallest calibration score. The split Conformal Prediction set at a new input x is then

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

Theorem A.3 (Split conformal coverage guarantee)**.** *If $(X_i, Y_i)_{i=1}^{n} \cup (X_0, Y_0)$ are exchangeable, then the prediction set $\mathcal{C}$ defined above satisfies*

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

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

The guarantee follows from exchangeability of the calibration and test scores: conditional on the fitted predictor, the test score has the same rank symmetry as the calibration scores. Hence, unlike full Conformal Prediction, split conformal achieves finite-sample marginal coverage without refitting the model for each candidate response.

Evaluation. Conformal Prediction methods are commonly evaluated along two dimensions: validity and efficiency. Validity concerns whether the prediction sets attain the desired coverage level, whereas efficiency reflects how concentrated or informative those sets are. These two criteria are typically summarized through coverage and prediction set size.

Coverage. Coverage measures how often the true response is contained in the reported prediction set. For a target miscoverage level $\alpha \in (0, 1)$, one seeks the guarantee

$$\mathbb{P}\left(Y_0 \in \mathcal{C}(X_0)\right) \ge 1 - \alpha, \tag{22}$$

for a new test point (X_0, Y_0) drawn from the same distribution as the training data. This finite-sample, distribution-free guarantee is the central validity property provided by Conformal Prediction.

In practice, coverage is estimated using an independent test set $(X_i, Y_i)_{i=1}^{n_{\text{test}}}$ according to

$$\widehat{\text{Cov}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\left\{ Y_i \in \mathcal{C}(X_i) \right\}. \tag{23}$$

This empirical quantity records the proportion of test observations for which the true outcome lies inside the corresponding prediction set. Values close to the nominal target $1 - \alpha$ indicate appropriate calibration, while persistent under-coverage signals a failure to provide the intended level of reliability.

Prediction set size. Coverage alone does not determine the practical usefulness of a conformal predictor, since arbitrarily large sets can trivially achieve high coverage. The size of the prediction

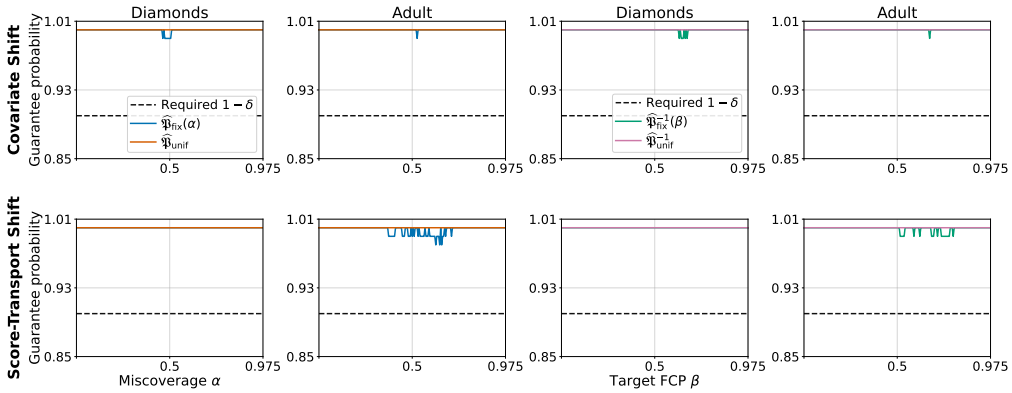

Figure 5: Empirical validation of the theoretical bounds for the direct and inverse FCP problems proposed in Table 1 on the Diamonds and Adult datasets. The top row corresponds to Covariate Shift, while the second row corresponds to score-transport shift (STS).

set therefore serves as a measure of efficiency: among methods with comparable coverage, smaller sets are generally preferable because they provide more precise uncertainty quantification.

The appropriate notion of size depends on the prediction task. For classification, it is typically the number of labels contained in the prediction set, whereas for regression it is commonly represented by interval length or, more generally, set volume. Over a test sample $\{X_i\}_{i=1}^{n_{\text{test}}}$, the empirical average size is

$$\widehat{\text{Size}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}(X_i)|. \tag{24}$$

Applications. Conformal Prediction has been successfully applied across a wide range of learning problems by augmenting point predictions with statistically valid prediction sets. Below, we briefly describe several representative applications, highlighting how Conformal Prediction is used, the form of the prediction sets, and the type of uncertainty control involved.

Multilabel Classification Papadopoulos (2014). In multilabel classification, Conformal Prediction is used to output a set of labels for each input rather than a single label vector. Prediction sets are constructed by conformalizing label-wise or joint nonconformity scores derived from probabilistic classifiers, ensuring that the true label set (or each true label) is contained with high probability. These methods often emphasize risk control, such as controlling the false discovery rate or expected set size, and can be adapted to label-conditional settings.

Disease (Sickness) Classification Vazquez & Facelli (2022). For disease classification tasks, Conformal Prediction produces a set of plausible diagnoses for a patient. The prediction set consists of disease labels whose scores satisfy a calibrated conformity threshold, guaranteeing that the true diagnosis is included with a prescribed probability. This setting primarily focuses on risk-sensitive decision making, allowing clinicians to identify ambiguous cases and defer decisions when uncertainty is high.

Polyp Segmentation Mossina & Friedrich (2025). In polyp segmentation, Conformal Prediction is applied to construct uncertainty-aware segmentation regions rather than a single binary mask. Prediction sets may take the form of pixel-wise confidence regions or spatial bands around predicted boundaries, providing coverage guarantees for the true polyp region. These sets are particularly useful for safety-critical applications, where missing polyp tissue carries a significant risk.

Weather Prediction Mortier et al. (2025). In weather forecasting, Conformal Prediction is used to generate prediction intervals or sets for continuous variables such as temperature or precipitation. These sets are typically constructed by conformalizing regression residuals or ensemble forecasts and provide valid coverage without strong distributional assumptions. This application often addresses Distribution Shift across time or locations through split or adaptive calibration.

Quantile Regression Romano et al. (2019). In quantile regression, Conformal Prediction is used to transform estimated conditional quantiles into valid prediction intervals. By conformalizing lower

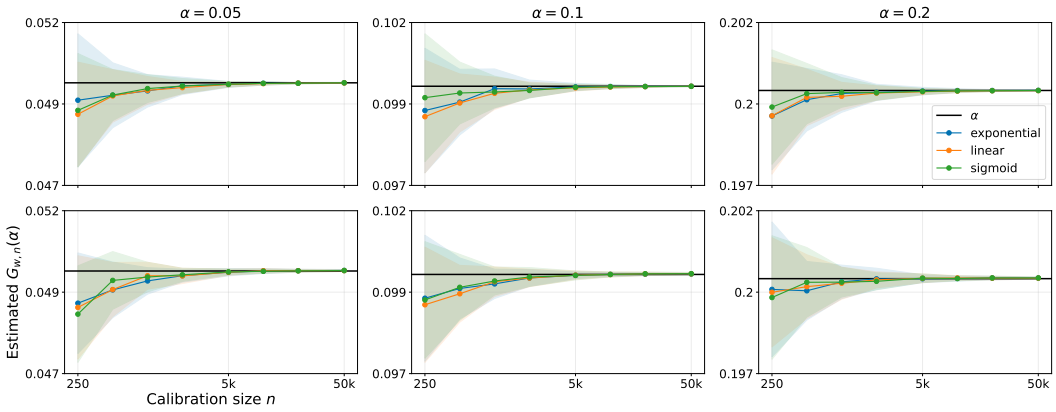

Figure 6: Empirical evaluation of Lemma 5.1 on the YearPredictionMSD (top) and Adult (bottom) datasets. The three columns correspond to $\alpha \in \{0.05, 0.1, 0.2\}$. Each colored curve shows the estimated $G_{w,n}(\alpha)$ under a different weight function as the calibration size n increases, while the black horizontal line indicates α.

and upper quantile estimates, the resulting intervals achieve guaranteed marginal coverage even under model misspecification. This approach is particularly effective in heteroskedastic settings and focuses on uncertainty calibration rather than pointwise accuracy.

Toxic Online Comment Identification Villate-Castillo et al. (2024). For toxic content detection, Conformal Prediction constructs prediction sets over toxicity labels to identify uncertain or borderline cases. When the conformal set contains multiple labels, the system may defer judgment or flag content for human review. This application emphasizes selective risk control, reducing harmful misclassifications in automated moderation systems.

Selective Classification Boursinos & Koutsoukos (2022). Selective classification naturally aligns with Conformal Prediction by allowing models to abstain when prediction sets are non-singleton or empty. Predictions are made only when the conformal set contains a single label, while abstention occurs otherwise. This framework provides explicit guarantees on the error rate of accepted predictions and enables principled risk-coverage trade-offs.

A.2 False Coverage Proportion

Motivation. While the marginal guarantee of Conformal Prediction introduced in Theorem A.3 provides rigorous protection for an individual future observation drawn at random, it does not directly describe the realized performance of the conformal procedure on a finite collection of test points. Suppose, for instance, that we are given a batch of m test observations. Even when every test point satisfies the marginal coverage guarantee, the fraction of points that are actually miscovered within this particular batch need not be close to α. The marginal guarantee controls the probability of miscoverage for a randomly selected test observation, or equivalently the expected fraction of miscovered points, but it does not by itself rule out a realization in which the empirical miscoverage is substantially larger. This distinction is particularly relevant when the same calibration sample is used to construct prediction sets for all test points, since the resulting coverage events are generally coupled through the common calibration data.

It is therefore natural to seek a stronger, batch-level statement that directly controls the fraction of miscovered observations among the realized test sample. Rather than asking only whether a randomly drawn test point is covered with probability at least $1 - \alpha$, we would like to guarantee, with high probability, that most of the m test observations are simultaneously covered. This viewpoint leads to the *transductive Conformal Prediction* setting introduced by Vovk (2013), in which a finite collection of test covariates is considered jointly and the quantity of interest is the empirical proportion of prediction sets that fail to contain their corresponding responses.

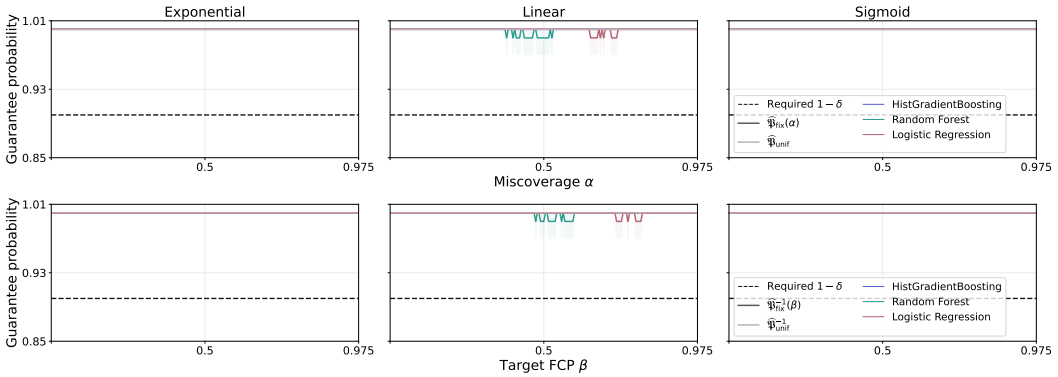

Figure 7: Ablation on the underlying predictive model on Adult. The top row compares empirical FCP with the fixed- and uniform-α bounds; the bottom row compares the empirical FCP of the inverse procedures with the target β. See Appendix D.4 for further experimental details.

More formally, let $\{(X_i, Y_i)\}_{i=1}^n$ denote the calibration sample, and consider a collection of m test observations

$$\{(X_{n+j}, Y_{n+j})\}_{j=1}^m.$$

For each test covariate X_{n+j}, let $\mathcal{C}^\alpha(X_{n+j})$ denote the Conformal Prediction set obtained at nominal miscoverage level α. We define the *False Coverage Proportion* (FCP) over the m test observations as

$$\mathrm{FCP}_m^{(n)}(\alpha) := \frac{1}{m}\sum_{j=1}^m \mathbf{1}\left\{Y_{n+j} \notin \mathcal{C}^\alpha(X_{n+j})\right\} = \frac{1}{m}\sum_{j=1}^m \mathbf{1}\left\{p_j^{(n)} \le \alpha\right\}, \tag{25}$$

where $p_j^{(n)}$ denotes the conformal p-value associated with the jth test observation. The second equality follows from the usual equivalence between exclusion from a Conformal Prediction set and the corresponding conformal p-value falling below the prescribed miscoverage level.

Primary theoretical objectives for FCP. Once the False Coverage Proportion has been identified as the relevant batch-level performance criterion, the next question is what kind of theoretical guarantees one should seek for it. Broadly speaking, there are two natural directions. In the *forward* problem, the miscoverage level α is fixed in advance, and the goal is to control the resulting realized False Coverage Proportion over the test batch. In the *inverse* problem, one instead starts from a desired upper bound β on the False Coverage Proportion and seeks to determine a conformal level that attains this target with high probability. In both cases, one may pursue either a guarantee at a single prescribed level, or a guarantee that holds uniformly over all levels. This leads to four fundamental inferential objectives.

1. Fixed-level direct FCP control. A first natural goal is to control the False Coverage Proportion at a single, pre-specified miscoverage level $\alpha \in (0,1)$. Concretely, one seeks an upper bound $\overline{\mathrm{FCP}}_{\mathrm{fix}}^{\delta,\alpha}$ such that

$$\mathbb{P}\left(\mathrm{FCP}_m^{(n)}(\alpha) \le \overline{\mathrm{FCP}}_{\mathrm{fix}}^{\delta,\alpha}\right) \ge 1-\delta, \qquad \text{for a fixed } \alpha \in (0,1). \tag{26}$$

This is the most basic form of batch-level validity. It is relevant when the user has already chosen a nominal conformal level α and wants a rigorous finite-sample statement about the realized proportion of miscovered test points. In other words, instead of only knowing that a randomly drawn test point is miscovered with probability at most α, one obtains a high-probability upper bound on the actual empirical miscoverage across the entire batch.

2. Uniform direct FCP control. In many applications, however, one does not wish to restrict attention to a single value of α. For instance, one may want to inspect the behavior of the procedure across many confidence levels, or choose α only after seeing a family of candidate prediction sets. In such situations, a stronger guarantee is needed, namely a bound $\overline{\mathrm{FCP}}_{\mathrm{unif}}^{\delta,\alpha}$ satisfying

$$\mathbb{P}\left(\forall \alpha \in (0,1),\ \mathrm{FCP}_m^{(n)}(\alpha) \le \overline{\mathrm{FCP}}_{\mathrm{unif}}^{\delta,\alpha}\right) \ge 1-\delta. \tag{27}$$

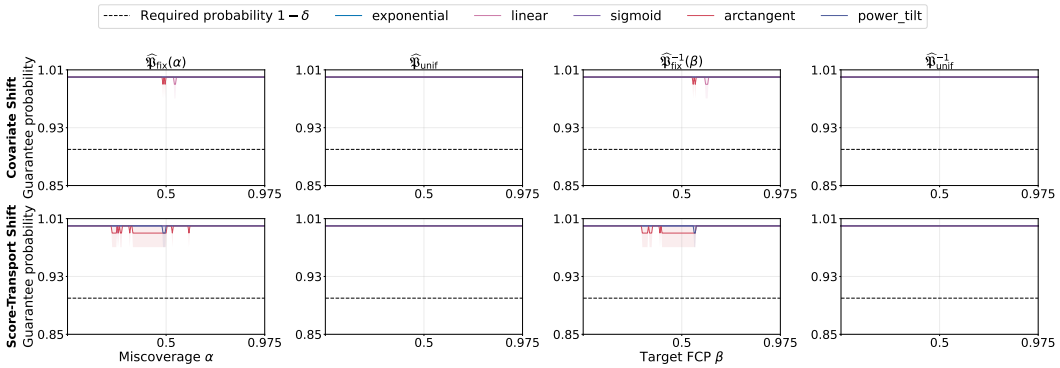

Figure 8: Sensitivity to the weight function on YearPredictionMSD. The top and bottom rows show Covariate Shift and Score-Transport Shift, respectively.

The need for such a uniform guarantee comes from the fact that fixed-α validity does not automatically justify simultaneous consideration of many levels. A uniform bound provides this stronger protection: with probability at least $1-\delta$, the entire False Coverage Proportion curve is controlled at once. This is especially useful when one wants robustness to post hoc tuning or wishes to understand the global behavior of the conformal procedure over the full range of miscoverage levels.

3. Fixed-target inverse FCP control. Beyond the forward problem, one may also consider the inverse question. Suppose that the practitioner specifies a tolerable False Coverage Proportion level $\beta \in (0,1)$. Rather than fixing α and asking what FCP results, the goal is now to choose a conformal level that guarantees the realized FCP does not exceed β with high probability. Formally, one seeks $\widehat{\mathrm{FCP}}^{\delta,\beta}_{\mathrm{fix}}$ such that

$$\mathbb{P}\Big(\mathrm{FCP}^{(n)}_m\big(\widehat{\mathrm{FCP}}^{\delta,\beta}_{\mathrm{fix}}\big) \le \beta\Big) \ge 1-\delta, \qquad \text{for a fixed } \beta \in (0,1). \tag{28}$$

This inverse formulation is important because, in practice, the quantity of real interest is often the tolerated batch-level error rate itself. A user may be willing to accept, say, at most a β fraction of miscovered points in the test batch, and therefore needs a principled rule for selecting the nominal conformal level accordingly. The inverse fixed-target problem provides exactly this type of calibration.

4. Uniform inverse FCP control. Finally, one may seek an inverse guarantee that holds simultaneously for all target levels $\beta \in (0,1)$. In this case, the objective is to construct $\widehat{\mathrm{FCP}}^{\delta,\beta}_{\mathrm{unif}}$ satisfying

$$\mathbb{P}\Big(\forall \beta \in (0,1),\ \mathrm{FCP}^{(n)}_m\big(\widehat{\mathrm{FCP}}^{\delta,\beta}_{\mathrm{unif}}\big) \le \beta\Big) \ge 1-\delta. \tag{29}$$

This is the strongest of the four objectives. It is needed when one wants simultaneous validity over an entire range of target FCP levels, rather than only at a single prescribed value. Such a guarantee supports flexible downstream use, since the user may later decide which target level β is most appropriate without invalidating the theoretical guarantee. Equivalently, it yields a uniformly valid inverse calibration rule across all admissible batch-level error tolerances.

Related work on FCP control. The study of coverage over multiple test observations has recently received increasing attention in conformal inference. Gazin et al. (2024) develop a general theory for transductive conformal inference by characterizing the joint distribution of conformal p-values under exchangeability. In particular, they show that this joint distribution admits a Pólya-urn representation and derive a finite-sample DKW-type concentration inequality for the empirical distribution function of the conformal p-values. Since the FCP can be written as the empirical distribution function of these p-values evaluated at the conformal level, their results yield high-probability upper bounds on the FCP that hold uniformly over the choice of the miscoverage level. Their framework also accommodates adaptive conformity scores that may depend symmetrically on the calibration and test covariates; however, the FCP guarantees fundamentally rely on exchangeability of the calibration and test scores, and therefore do not directly address Distribution Shift between the calibration and

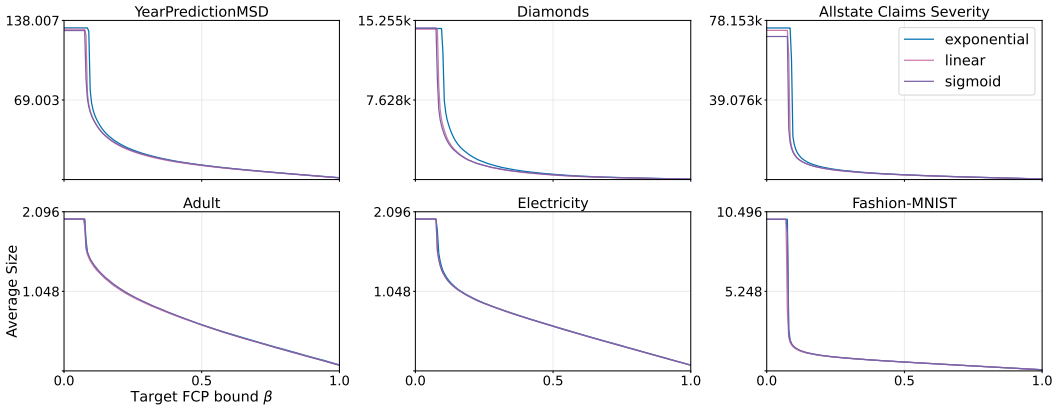

Figure 9: Prediction efficiency versus the target FCP bound β under Covariate Shift. The top row shows average prediction-interval length on regression datasets; the bottom row shows average prediction-set size on classification datasets. Each curve corresponds to a weight function and averages 100 repetitions. Smaller values indicate more efficient predictions.

test populations. Blain et al. (2025) build on this transductive characterization and propose CoJER, a procedure specifically designed to control the realized FCP in probability. Rather than relying on the usual marginal guarantee of split Conformal Prediction, CoJER adjusts the prediction intervals so that the proportion of miscovered test observations remains below a prescribed target with high probability, while also incorporating an aggregation mechanism to improve robustness with respect to the underlying predictive model. Thus, this work focuses primarily on constructing prediction sets that achieve a prescribed FCP target, corresponding to the inverse perspective considered in our formulation, again under the standard exchangeable setting.

Related ideas have also appeared in more specialized conformal inference problems. Fermanian et al. (2026) study transductive conformal inference for full ranking, where the objective is to construct prediction sets for the unknown ranks of several new items. Because the ranks of the calibration items themselves depend on the positions of the test items, standard conformal calibration cannot be applied directly. The authors instead derive distribution-free bounds on the unknown conformity scores and use recent results on the joint distribution of conformal p-values to construct rank prediction sets. Since prediction sets are produced simultaneously for multiple test items, they additionally impose high-probability control of their FCP, thereby adapting transductive FCP methodology to the ranking problem. Gazin et al. (2025), on the other hand, consider Conformal Prediction after a data-dependent selection step, where prediction sets are reported only for observations deemed sufficiently informative, for example because the resulting intervals are sufficiently short or exclude a designated null value. Their target is the *false coverage rate* (FCR), defined as the expectation of the FCP over the selected observations, and they introduce procedures such as InfoSP and InfoSCOP to control this quantity. This objective is therefore distinct from the high-probability control of the realized FCP studied here: FCR controls the average behavior of the random FCP, whereas our objective is to bound its realized value with probability at least $1 - \delta$. Song et al. (2026) study a closely related multiple-inference problem through the *false discovery proportion* (FDP), rather than the FCP. They construct finite-sample upper confidence envelopes for the FDP that are simultaneously valid over all conformal p-value rejection thresholds, thereby allowing the threshold to be selected after observing the data. Although the inferential target is FDP rather than prediction-set miscoverage, their work is closely connected at a technical level to uniform FCP control, since both problems require controlling an empirical error proportion generated by a dependent collection of conformal p-values.

Comparison with Gazin (2024). The work most closely related to ours is Gazin (2024), which develops a systematic asymptotic analysis of the FCP process. Under the exchangeable setting, they derive the limiting distribution of the FCP as both the calibration size n and test size m grow, and characterize how its fluctuations depend on the relative growth of the two sample sizes. Importantly, their analysis is also extended beyond exchangeability to settings with Distribution Shift between the calibration and test score distributions. In this case, the FCP no longer converges to the nominal

miscoverage level α, but instead to a distribution-dependent function that explicitly captures the shift. They further study weighted Conformal Prediction, including oracle and non-oracle weighting schemes, and characterize how weighting modifies both the limiting behavior and the asymptotic variance of the FCP. These results provide an important description of how Distribution Shift affects batch-level conformal coverage and are therefore particularly relevant to our setting.

There is, however, a fundamental distinction between these asymptotic characterizations and the guarantees developed in this work. The distribution-shift results of Gazin (2024) describe the limiting behavior of the FCP as the relevant sample sizes tend to infinity, and hence the resulting approximation becomes informative through a large-sample argument. In contrast, our results provide explicit finite-sample, high-probability guarantees for the FCP under Distribution Shift, valid for the realized calibration and test sample sizes n and m. This distinction is important in conformal inference, whose practical appeal rests heavily on finite-sample validity: in many applications, neither the calibration nor the test sample is sufficiently large for the accuracy of an asymptotic approximation to be taken for granted. Moreover, an asymptotic limit describes what the FCP approaches as the sample sizes increase, but does not by itself quantify how far the realized FCP at a given finite (n, m) may deviate from that limit. A finite-sample bound instead provides an explicit level of protection at the sample sizes actually available to the practitioner. Our results therefore complement the asymptotic theory of Gazin (2024) by providing non-asymptotic FCP guarantees under Distribution Shift, including both fixed-level and uniform control.

A further distinction concerns the practical computability of the resulting guarantees. Under Distribution Shift, the asymptotic characterization of Gazin (2024) depends on population-level quantities involving the test-score distribution. In the weighted setting, a central quantity takes the form

$$G_w(t) = 1 - F_{\text{test}} \circ \left(F^w_{\text{cal}}\right)^{-1}(1-t),$$

where F_{test} denotes the distribution of the conformity score under the test distribution and F^w_{cal} denotes the corresponding weighted calibration-score distribution. The limiting covariance additionally involves further distribution-dependent quantities derived from G_w. These terms are generally not directly observable in the standard prediction setting, since evaluating F_{test} requires knowledge of the distribution of test nonconformity scores, while the test scores themselves depend on the unknown test responses. Consequently, even when the asymptotic characterization is available theoretically, the corresponding quantities need not be directly computable from the observed calibration sample and unlabeled test covariates.

In contrast, we explicitly address this practical obstacle. In addition to deriving the theoretical FCP guarantees, Algorithms 1 and 2 provide estimators of the unknown distributional quantities required for the direct and inverse guarantees, respectively, without requiring the test responses to be observed. This makes the proposed bounds operational in the prediction setting considered here. To our knowledge, Gazin (2024) does not provide a corresponding general estimation procedure that renders their distribution-shift FCP characterization directly computable from calibration data and unlabeled test covariates alone.

Finally, our work also considers a complementary inferential direction that is not addressed in Gazin (2024). Their analysis studies the *direct* FCP problem: for a prescribed conformal miscoverage level α, the objective is to characterize the resulting false coverage proportion. We additionally study the *inverse FCP* problem, in which one instead begins with a desired batch-level tolerance β and seeks a conformal level α such that

$$\text{FCP}^{(n)}_m(\alpha) \leq \beta$$

with high probability. This inverse formulation is particularly relevant when the desired control on the realized batch-level error is specified in advance and the conformal level must be chosen accordingly. We derive both fixed-target and uniform inverse guarantees, together with a practical estimator for the quantities required to implement them. Thus, relative to Gazin (2024), our contribution is not only finite-sample rather than asymptotic, but also provides practically computable guarantees and addresses both the direct and inverse FCP problems under Distribution Shift.

A.3 Distribution Shift and Weighted Conformal Prediction

Motivation. A key strength of Conformal Prediction is that its standard finite-sample validity requires only the relatively mild assumption of exchangeability, introduced in Definition A.1. Nevertheless,

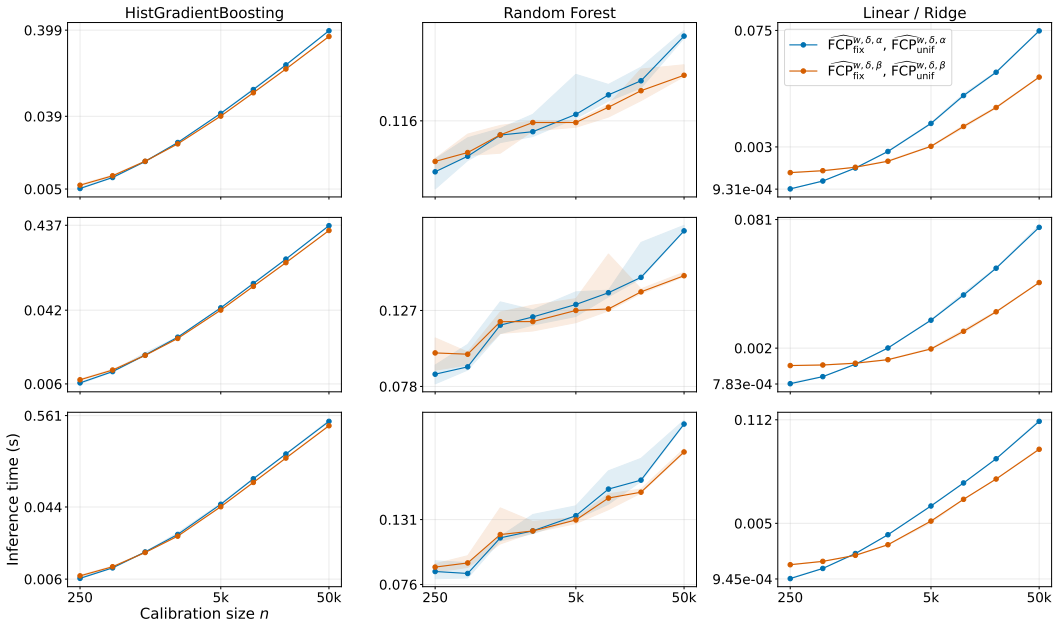

Figure 10: Runtime analysis of the proposed estimation procedures as a function of the calibration size n. The rows correspond to the Adult, Diamonds, and YearPredictionMSD datasets, respectively, while the columns correspond to histogram-based gradient boosting, random forest, and linear/ridge regression models.

this assumption can fail in many practical deployment settings. One common example is *temporal Distribution Shift*, where the calibration data are collected during an earlier time period while predictions are made at a later stage. Changes in population behavior, economic conditions, clinical practice, sensor characteristics, or other environmental factors may then cause the future test distribution to differ systematically from the distribution represented by the calibration sample. A second example arises under *population or domain shift*. A predictive model may be calibrated on data collected from one hospital, geographic region, demographic population, or experimental environment, but subsequently deployed on another population with different covariate characteristics. Even if the prediction task itself remains unchanged, such a discrepancy can invalidate the symmetry between calibration and test observations required by standard conformal arguments. A third important setting occurs when the test population is affected by *selection or sampling mechanisms*. For instance, the calibration sample may be drawn from a broad source population, whereas the deployed system is applied only to a non-random subset of individuals selected according to covariates, eligibility criteria, or operational constraints. The resulting test observations may therefore follow a systematically reweighted version of the calibration distribution.

These examples all violate the usual exchangeability assumption in essentially the same way: the calibration and test observations are no longer identically distributed. To study conformal inference in such settings, a growing line of work (Tibshirani et al., 2019; Barber et al., 2023; Gibbs & Candes, 2021; Aolaritei et al., 2026; Gibbs & Candès, 2024; Liu et al., 2024; Xu et al., 2025; Huang et al., 2025; Correia & Louizos, 2026) considers Conformal Prediction under *Distribution Shift*, where the calibration observations are drawn from a distribution P_{cal} while the test observations are drawn from a potentially different distribution P_{test}, with

$$P_{\text{cal}} \neq P_{\text{test}}.$$

The objective is then to recover meaningful coverage guarantees despite this mismatch between the calibration and test populations.

Covariate Shift. An important and widely studied form of Distribution Shift is *Covariate Shift*, first considered in the context of Conformal Prediction by Tibshirani et al. (2019). In this setting, the distribution of the covariates may change between calibration and deployment, while the conditional distribution of the response given the covariates remains unchanged. Thus, the shift arises from a

change in which regions of the covariate space are encountered more frequently at test time, rather than from a change in the underlying relationship between X and Y.

A representative example arises in medical prediction. Suppose that a model is calibrated using data collected from one hospital and subsequently deployed at another hospital. The two hospitals may serve populations with substantially different distributions of age, disease severity, or other patient characteristics, so that the marginal distribution of X differs between calibration and deployment. Nevertheless, after conditioning on these relevant patient characteristics, the distribution of the clinical outcome may remain approximately the same across hospitals. In such a case, the test population places different probability mass on different regions of the covariate space, but the conditional prediction problem within each region remains unchanged. This is precisely the structure captured by Covariate Shift.

Formally, let P_{cal} and P_{test} denote the joint distributions of (X, Y) for the calibration and test populations, respectively. Under Covariate Shift, we assume that

$$P_{\text{cal}}(Y \mid X) = P_{\text{test}}(Y \mid X), \tag{30}$$

while allowing the corresponding covariate marginals P^X_{cal} and P^X_{test} to differ. We further require

$$P^X_{\text{test}} \ll P^X_{\text{cal}},$$

so that the test covariate distribution is absolutely continuous with respect to the calibration covariate distribution. This condition ensures that the test distribution does not assign positive probability to regions of the covariate space that are completely absent from the calibration population; equivalently, the distributional discrepancy can be represented through a density ratio.

Under these assumptions, the discrepancy between the two covariate distributions is characterized by the oracle importance weight

$$w_0(x) := \frac{dP^X_{\text{test}}}{dP^X_{\text{cal}}}(x), \qquad x \in \mathcal{X}. \tag{31}$$

The quantity $w_0(x)$ measures how much more or less prevalent a covariate value x is under the test distribution relative to the calibration distribution. Calibration observations lying in regions that are more common at test time receive larger weights, whereas observations from regions that are overrepresented in the calibration population receive smaller weights. In this way, weighting rebalances the calibration sample so that, at the level of the covariates, it behaves as though it had been drawn from the test population.

This observation forms the basis of Weighted Conformal Prediction under Covariate Shift. Specifically, Tibshirani et al. (2019) show that when the calibration observations are weighted according to the oracle density ratio w_0, and the test point is assigned the corresponding weight $w_0(X_{\text{test}})$, the resulting weighted conformal procedure recovers the standard marginal coverage guarantee under the test distribution:

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}^{w_0,\alpha}(X_{\text{test}})\right) \geq 1 - \alpha. \tag{32}$$

Thus, although ordinary exchangeability between calibration and test observations is lost, the particular structure of Covariate Shift allows the discrepancy to be corrected through importance weighting. The conditional equality in equation 30 ensures that correcting the covariate distribution is sufficient: once the difference between P^X_{cal} and P^X_{test} is accounted for, the conditional law of the response is already aligned across the two populations.

Weighted Conformal Prediction. The use of importance weighting in Conformal Prediction is not restricted to the Covariate Shift setting described above. More generally, when exchangeability between the calibration and test observations is violated, one may assign different levels of importance to the calibration observations according to how relevant they are to the target test distribution. This broader perspective was developed by Barber et al. (2023), who study conformal inference beyond exchangeability and introduce weighted conformal constructions designed to accommodate distribution drift and other forms of non-exchangeability. Rather than requiring every calibration observation to contribute equally, their framework allows observations that are more representative of the test population to receive greater weight, while observations that are less representative receive smaller weight. The weights need not necessarily arise from the exact covariate density

ratio considered above; they may instead encode other information about the discrepancy between calibration and deployment, such as temporal recency or similarity to the target population. Their analysis shows how such weighting can substantially reduce the loss of coverage induced by violations of exchangeability, while recovering the usual conformal guarantees when exchangeability holds.

Inspired by the above motivation and following the weighted conformal formulation adopted in the asymptotic FCP analysis of Gazin (2024), we associate each calibration observation with a positive weight

$$w : \mathcal{X} \to \mathbb{R}_{>0},$$

and assign an additional positive mass $w_\infty > 0$ to the test point. Writing

$$S_k = s(X_k, Y_k), \qquad k \in [n],$$

for the calibration nonconformity scores, we define

$$W := w_\infty + \sum_{k=1}^{n} w(X_k)$$

as the total weight. The resulting weighted empirical measure on the calibration scores is

$$\frac{w_\infty}{W}\delta(\infty) + \sum_{k=1}^{n} \frac{w(X_k)}{W}\delta(S_k).$$

The atom at ∞ plays the same role as the usual test-point correction in split Conformal Prediction: since the test response is unavailable at prediction time, its conformity score is unknown and is represented conservatively by an additional mass at infinity.

For a test covariate X_{n+i}, the weighted Conformal Prediction set at level $\alpha \in (0, 1)$ is therefore defined as

$$\mathcal{C}^{w,\alpha}(X_{n+i}) := \left\{ y \in \mathcal{Y} : s(X_{n+i}, y) \le Q_{1-\alpha}\left(\frac{w_\infty}{W}\delta(\infty) + \sum_{k=1}^{n} \frac{w(X_k)}{W}\delta(S_k) \right) \right\}, \tag{33}$$

where $Q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the weighted empirical measure. Relative to ordinary split Conformal Prediction, the only modification is that the calibration scores no longer contribute uniformly: the influence of the kth calibration observation is proportional to $w(X_k)$.

It is often convenient to express the same construction through weighted conformal p-values. For a candidate response $y \in \mathcal{Y}$ at the ith test point, define

$$p^{w,(n)}(X_{n+i}, y) := \frac{w_\infty + \sum_{k=1}^{n} w(X_k)\mathbf{1}\{S_k \ge s(X_{n+i}, y)\}}{w_\infty + \sum_{k=1}^{n} w(X_k)}. \tag{34}$$

For the realized response Y_{n+i}, we use the shorthand

$$p_i^{w,(n)} := p^{w,(n)}(X_{n+i}, Y_{n+i}).$$

With this notation, the prediction set in equation 33 can equivalently be written as

$$\mathcal{C}^{w,\alpha}(X_{n+i}) = \left\{ y \in \mathcal{Y} : p^{w,(n)}(X_{n+i}, y) > \alpha \right\}. \tag{35}$$

Hence, the ith test point is miscovered exactly when $p_i^{w,(n)} \le \alpha$, a representation that will be used throughout our analysis of the FCP.

We specify that, in more general formulations of Weighted Conformal Prediction, the test weight may depend on the test covariate, leading to a point-specific quantity $w_{\infty,i}$ for each X_{n+i}; under Covariate Shift, for example, one may take $w_{\infty,i} = w_0(X_{n+i})$. In the present work, as in the asymptotic FCP framework of Gazin (2024), we instead restrict attention to a fixed positive test weight w_∞ that is shared across all m test observations. The calibration weights $w(X_k)$ remain observation-dependent, while the test contribution is held fixed across the batch. This convention aligns our finite-sample formulation with the asymptotic setting of Gazin (2024) and ensures that all test points are evaluated against the same weighted calibration distribution.

B THEORETICAL PROOFS

B.1 PROOF OF THE FORWARD FCP BOUND FOR FIXED MISCOVERAGE - THEOREM 4.1

Proof. Let us first denote $d_n := \mu_w + a_n$ and

$$\overline{A}_n(t) := \frac{1}{n}\sum_{k=1}^{n} w(X_k)\mathbf{1}\{S_k \le t\}, \qquad \overline{W}_n := \frac{1}{n}\sum_{k=1}^{n} w(X_k).$$

Then, the weighted empirical CDF $\widehat{F}_n^w$ and its marginal counterpart $F_{w,n}$ can be written as

$$\widehat{F}_{w,n}(t) = \frac{\overline{A}_n(t)}{a_n + \overline{W}_n}, \qquad F_{w,n}(t) = \frac{A_w(t)}{a_n + \mu_w} = \frac{A_w(t)}{d_n}.$$

If $\alpha + \Delta_{\text{fix}} \ge 1$, then $G_{w,n}\left(\alpha + \Delta_{\text{fix}}\right) = 1$, and the desired result is trivial. Hence, assume from now on that

$$\alpha + \Delta_{\text{fix}} < 1.$$

Define $c_\alpha := 1 - \alpha - \Delta_{\text{fix}}$, and

$$\tau_\alpha := \sup\left\{t \in \mathbb{R} : F_{w,n}(t) < c_\alpha\right\} \in \overline{\mathbb{R}}.$$

Here, $F_{w,n}$ and A_w are extended to $\overline{\mathbb{R}}$ by their natural limits at $\pm\infty$. By the definition of τ_α, the continuity and monotonicity of $F_{w,n}$ on $\mathbb{R}$, and its natural limits at $\pm\infty$,

$$F_{w,n}(\tau_\alpha) \le c_\alpha.$$

Equivalently, $A_w(\tau_\alpha) \le d_n\left(1 - \alpha - \Delta_{\text{fix}}\right)$. For each $k \in [n]$, define

$$Y_k := w(X_k)\left[\mathbf{1}\{S_k \le \tau_\alpha\} - (1-\alpha)\right].$$

Then

$$\frac{1}{n}\sum_{k=1}^{n} Y_k = \overline{A}_n(\tau_\alpha) - (1-\alpha)\overline{W}_n.$$

Moreover,

$$\mathbb{E}[Y_k] = A_w(\tau_\alpha) - (1-\alpha)\mu_w.$$

We next determine the range of Y_k. If $S_k \le \tau_\alpha$, then

$$Y_k = \alpha w(X_k) \in [0, \alpha\varepsilon_w].$$

If $S_k > \tau_\alpha$, then

$$Y_k = -(1-\alpha)w(X_k) \in [-(1-\alpha)\varepsilon_w, 0].$$

Therefore, $Y_k \in [-(1-\alpha)\varepsilon_w, \alpha\varepsilon_w]$, whose interval length is exactly ε_w. By Hoeffding's inequality, for every $r > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k - \mathbb{E}[Y_k] \ge r\right) \le \exp\left(-\frac{2nr^2}{\varepsilon_w^2}\right).$$

Choose

$$r := \varepsilon_w\sqrt{\frac{\log(1/\delta_{\text{cal}})}{2n}} = d_n\Delta_{\text{fix}}.$$

Then, with probability at least $1 - \delta_{\text{cal}}$,

$$\frac{1}{n}\sum_{k=1}^{n} Y_k < \mathbb{E}[Y_k] + d_n\Delta_{\text{fix}}.$$

On this event,

$$\begin{aligned}
&\overline{A}_n(\tau_\alpha) - (1-\alpha)\left(a_n + \overline{W}_n\right) \\
&= \frac{1}{n}\sum_{k=1}^{n} Y_k - (1-\alpha)a_n \\
&< A_w(\tau_\alpha) - (1-\alpha)\mu_w + d_n\Delta_{\text{fix}} - (1-\alpha)a_n \\
&= A_w(\tau_\alpha) - (1-\alpha)d_n + d_n\Delta_{\text{fix}} \\
&\le d_n\left(1 - \alpha - \Delta_{\text{fix}}\right) - (1-\alpha)d_n + d_n\Delta_{\text{fix}} \\
&= 0.
\end{aligned}$$

Hence $\overline{A}_n(\tau_\alpha) < (1-\alpha)\left(a_n + \overline{W}_n\right)$, or equivalently, $\widehat{F}_n^w(\tau_\alpha) < 1-\alpha$. Since $\widehat{F}_n^w$ is nondecreasing, for every $t \le \tau_\alpha$,

$$\widehat{F}_{w,n}(t) < 1 - \alpha.$$

Therefore,

$$\left\{\widehat{F}_n^w(T_i) \ge 1 - \alpha\right\} \subseteq \{T_i > \tau_\alpha\}.$$

By the definition of τ_α,

$$T_i > \tau_\alpha \quad \Longrightarrow \quad F_{w,n}(T_i) \ge 1 - \alpha - \Delta_{\text{fix}}.$$

Consequently, on the preceding calibration event,

$$\begin{aligned}\mathbb{P}_{\text{test}}\left(p_i^{w,(n)} \le \alpha \mid \mathcal{D}_{\text{cal}}\right) &= \mathbb{P}_{\text{test}}\left(\widehat{F}_n^w(T_i) \ge 1 - \alpha \mid \mathcal{D}_{\text{cal}}\right) \\ &\le \mathbb{P}_{\text{test}}\left(F_{w,n}(T_i) \ge 1 - \alpha - \Delta_{\text{fix}}\right) \\ &= G_{w,n}\left(\alpha + \Delta_{\text{fix}}\right).\end{aligned}$$

Define the calibration event

$$E_{\text{cal}} := \left\{\mathbb{P}_{\text{test}}\left(p_i^{w,(n)} \le \alpha \mid \mathcal{D}_{\text{cal}}\right) \le G_{w,n}\left(\alpha + \Delta_{\text{fix}}\right)\right\}.$$

The preceding argument gives

$$\mathbb{P}(E_{\text{cal}}) \ge 1 - \delta_{\text{cal}}.$$

Now condition on the calibration sample. The random variables

$$\mathbf{1}\left\{p_i^{w,(n)} \le \alpha\right\}, \qquad i = 1, \dots, m,$$

are conditionally i.i.d. Bernoulli random variables. Let

$$q_n(\alpha) := \mathbb{P}_{\text{test}}\left(p_i^{w,(n)} \le \alpha \mid \mathcal{D}_{\text{cal}}\right).$$

By Hoeffding's inequality,

$$\mathbb{P}\left(\text{FCP}_m^{w,(n)}(\alpha) > q_n(\alpha) + \varepsilon_{\text{fix}} \;\middle|\; \mathcal{D}_{\text{cal}}\right) \le \delta_{\text{test}}.$$

Define

$$E_{\text{test}} := \left\{\text{FCP}_m^{w,(n)}(\alpha) \le q_n(\alpha) + \varepsilon_{\text{fix}}\right\}.$$

Taking expectations over the calibration sample gives

$$\mathbb{P}(E_{\text{test}}^c) \le \delta_{\text{test}}.$$

On $E_{\text{cal}} \cap E_{\text{test}}$, $\text{FCP}_m^{w,(n)}(\alpha) \le G_{w,n}\left(\alpha + \Delta_{\text{fix}}\right) + \varepsilon_{\text{fix}}$. Finally, by the union bound,

$$\mathbb{P}\left(E_{\text{cal}} \cap E_{\text{test}}\right) \ge 1 - \delta_{\text{cal}} - \delta_{\text{test}}.$$

This proves the theorem. □

B.2 Proof of the Uniform Forward FCP Bound over Miscoverage - Theorem 4.2

Proof. Let us first denote

$$\overline{A}_n(t) := \frac{1}{n}\sum_{k=1}^{n} w(X_k)\mathbf{1}\{S_k \le t\}, \qquad \overline{W}_n := \frac{1}{n}\sum_{k=1}^{n} w(X_k).$$

Then the weighted empirical CDF $\widehat{F}_n^w$ and its marginal counterpart $F_{w,n}$ can be written as

$$\widehat{F}_{w,n}(t) = \frac{\overline{A}_n(t)}{a_n + \overline{W}_n}, \qquad F_{w,n}(t) = \frac{A_w(t)}{a_n + \mu_w} = \frac{A_w(t)}{d_n}.$$

We first derive a one-sided uniform concentration bound for $\overline{A}_n$. Let $U_1, \dots, U_n$ be independent $\text{Unif}(0,1)$ random variables, independent of the calibration sample. Define

$$Q_k := \begin{cases} S_k, & U_k \le w(X_k)/\varepsilon_w, \\ +\infty, & U_k > w(X_k)/\varepsilon_w. \end{cases}$$

For $t \in \mathbb{R}$, define $\widetilde{F}_n(t) := \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\{Q_k \le t\}$, and let $H(t) := \mathbb{P}(Q_1 \le t)$. Conditionally on the calibration sample,

$$\mathbb{E}\left[\mathbf{1}\{Q_k \le t\} \mid X_k, Y_k\right] = \frac{w(X_k)}{\varepsilon_w}\mathbf{1}\{S_k \le t\}.$$

Therefore,

$$\mathbb{E}\left[\widetilde{F}_n(t) \mid \mathcal{D}_{\text{cal}}\right] = \frac{\overline{A}_n(t)}{\varepsilon_w}.$$

Moreover,

$$H(t) = \mathbb{E}\left[\frac{w(X_1)}{\varepsilon_w}\mathbf{1}\{S_1 \le t\}\right] = \frac{A_w(t)}{\varepsilon_w}.$$

Define $Z_n^+ := \frac{1}{\varepsilon_w}\sup_{t\in\mathbb{R}}\left\{\overline{A}_n(t) - A_w(t)\right\}$. Since

$$\overline{A}_n(t) - A_w(t) \longrightarrow 0 \qquad \text{as } t \to -\infty,$$

we have $Z_n^+ \ge 0$. Also define

$$V_n^+ := \sup_{t\in\mathbb{R}}\left\{\widetilde{F}_n(t) - H(t)\right\}.$$

Since both $\widetilde{F}_n(t)$ and $H(t)$ converge to zero as $t \to -\infty$, we also have $V_n^+ \ge 0$. For every $t \in \mathbb{R}$,

$$\frac{\overline{A}_n(t) - A_w(t)}{\varepsilon_w} = \mathbb{E}\left[\widetilde{F}_n(t) - H(t) \mid \mathcal{D}_{\text{cal}}\right].$$

Hence,

$$\begin{aligned} Z_n^+ &= \sup_{t\in\mathbb{R}} \mathbb{E}\left[\widetilde{F}_n(t) - H(t) \mid \mathcal{D}_{\text{cal}}\right] \\ &\le \mathbb{E}\left[\sup_{t\in\mathbb{R}}\left\{\widetilde{F}_n(t) - H(t)\right\} \mid \mathcal{D}_{\text{cal}}\right] \\ &= \mathbb{E}\left[V_n^+ \mid \mathcal{D}_{\text{cal}}\right]. \end{aligned}$$

Taking expectations on both sides gives

$$\mathbb{E}[Z_n^+] \le \mathbb{E}[V_n^+].$$

Since $Q_1, \ldots, Q_n$ are i.i.d. with distribution function H, the one-sided Dvoretzky–Kiefer–Wolfowitz inequality gives, for every $x > 0$,

$$\mathbb{P}\left(V_n^+ > x\right) \le \exp(-2nx^2).$$

Because $V_n^+ \ge 0$, the tail-integration identity yields

$$\begin{aligned} \mathbb{E}[V_n^+] &= \int_0^\infty \mathbb{P}(V_n^+ > x)\, dx \\ &\le \int_0^\infty \exp(-2nx^2)\, dx \\ &= \sqrt{\frac{\pi}{8n}}. \end{aligned}$$

Consequently,

$$\mathbb{E}[Z_n^+] \le \sqrt{\frac{\pi}{8n}}.$$

We next apply McDiarmid's inequality. Replacing one calibration observation can change

$$\frac{\overline{A}_n(t)}{\varepsilon_w}$$

by at most $1/n$, uniformly over $t \in \mathbb{R}$. Therefore, replacing one observation can change Z_n^+ by at most $1/n$. It follows that

$$\mathbb{P}\left(Z_n^+ > \mathbb{E}[Z_n^+] + \sqrt{\frac{\log(1/\delta_+)}{2n}}\right) \le \delta_+.$$

Combining this with the bound on $\mathbb{E}[Z_n^+]$, we obtain

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}\left\{\overline{A}_n(t) - A_w(t)\right\} > r_+\right) \le \delta_+.$$

Define $E_+ := \left\{\forall t \in \mathbb{R}, \quad \overline{A}_n(t) \le A_w(t) + r_+\right\}$. Then

$$\mathbb{P}(E_+) \ge 1 - \delta_+.$$

We now control the denominator. Since

$$0 < w(X_k) \le \varepsilon_w,$$

Hoeffding's inequality gives

$$\mathbb{P}\left(\overline{W}_n - \mu_w < -r_-\right) \le \exp\left(-\frac{2nr_-^2}{\varepsilon_w^2}\right) = \delta_-.$$

Define $E_- := \left\{\overline{W}_n \ge \mu_w - r_-\right\}$. Then $\mathbb{P}(E_-) \ge 1 - \delta_-$. Suppose first that $r_- < d_n$. On the event $E_+ \cap E_-$, for every $t \in \mathbb{R}$,

$$\begin{aligned}\widehat{F}_{w,n}(t) &= \frac{\overline{A}_n(t)}{}a_n + \overline{W}_n \\ &\le \frac{A_w(t) + r_+}{}d_n - r_-.\end{aligned}$$

Therefore,

$$\begin{aligned}\widehat{F}_{w,n}(t) - F_{w,n}(t) &\le \frac{A_w(t) + r_+}{d_n - r_-} - \frac{A_w(t)}{d_n} \\ &= \frac{d_n r_+ + A_w(t) r_-}{d_n(d_n - r_-)}.\end{aligned}$$

Since $A_w(t) \le \mu_w$, we obtain

$$\widehat{F}_{w,n}(t) - F_{w,n}(t) \le \frac{d_n r_+ + \mu_w r_-}{d_n(d_n - r_-)} = \Delta_{\mathrm{unif}}.$$

Thus, on $E_+ \cap E_-$,

$$\sup_{t\in\mathbb{R}}\left\{\widehat{F}_{w,n}(t) - F_{w,n}(t)\right\} \le \Delta_{\mathrm{unif}}.$$

Now condition on the calibration sample and define $G_n(\alpha) := \mathbb{P}_{\mathrm{test}}\left(p_i^{w,(n)} \le \alpha \mid \mathcal{D}_{\mathrm{cal}}\right)$. Using

$$p_i^{w,(n)} = 1 - \widehat{F}_n^w(T_i) \qquad \text{almost surely,}$$

we have

$$G_n(\alpha) = \mathbb{P}_{\mathrm{test}}\left(\widehat{F}_n^w(T_i) \ge 1 - \alpha \mid \mathcal{D}_{\mathrm{cal}}\right).$$

On $E_+ \cap E_-$, simultaneously for every $\alpha \in (0,1)$,

$$\widehat{F}_n^w(T_i) \ge 1 - \alpha \quad \Longrightarrow \quad F_{w,n}(T_i) \ge 1 - \alpha - \Delta_{\mathrm{unif}}.$$

Therefore, $G_n(\alpha) \le G_{w,n}\left(\alpha + \Delta_{\mathrm{unif}}\right)$ simultaneously for all $\alpha \in (0,1)$. If $r_- \ge d_n$, then $\Delta_{\mathrm{unif}} = 1$ by definition and hence $G_{w,n}\left(\alpha + \Delta_{\mathrm{unif}}\right) = 1$. In this case, the preceding inequality holds trivially. Define the calibration event

$$E_{\mathrm{cal}} := \left\{\forall \alpha \in (0,1), \quad G_n(\alpha) \le G_{w,n}\left(\alpha + \Delta_{\mathrm{unif}}\right)\right\}.$$

By the union bound,

$$\mathbb{P}(E_{\mathrm{cal}}) \ge 1 - \delta_+ - \delta_-.$$

It remains to control the empirical test-sample fluctuation. Conditionally on $\mathcal{D}_{\mathrm{cal}}$, the random variables

$$p_1^{w,(n)}, \ldots, p_m^{w,(n)}$$

are i.i.d. with distribution function G_n. By the one-sided DKW inequality,

$$\mathbb{P}\left(\sup_{\alpha\in(0,1)}\left\{\mathrm{FCP}_m^{w,(n)}(\alpha)-G_n(\alpha)\right\}>\varepsilon_{\mathrm{unif}}\,\middle|\,\mathcal{D}_{\mathrm{cal}}\right)\le\exp\left(-2m(\varepsilon_{\mathrm{unif}})^2\right)=\delta_{\mathrm{test}}.$$

Define

$$E_{\mathrm{test}}:=\left\{\forall\alpha\in(0,1),\quad \mathrm{FCP}_m^{w,(n)}(\alpha)\le G_n(\alpha)+\varepsilon_{\mathrm{unif}}\right\}.$$

Taking expectations over the calibration sample gives

$$\mathbb{P}(E_{\mathrm{test}}^c)\le\delta_{\mathrm{test}}.$$

Finally, on $E_{\mathrm{cal}}\cap E_{\mathrm{test}}$, simultaneously for every $\alpha\in(0,1)$,

$$\mathrm{FCP}_m^{w,(n)}(\alpha)\le G_{w,n}\left(\alpha+\Delta_{\mathrm{unif}}\right)+\varepsilon_{\mathrm{unif}}.$$

Another union bound gives

$$\begin{aligned}\mathbb{P}\left(E_{\mathrm{cal}}\cap E_{\mathrm{test}}\right)&\ge 1-\mathbb{P}(E_{\mathrm{cal}}^c)-\mathbb{P}(E_{\mathrm{test}}^c)\\&\ge 1-\delta_+-\delta_--\delta_{\mathrm{test}}.\end{aligned}$$

This proves the theorem. □

B.3 Proof of the Bound on $G_{w,n}$ under Covariate Shift - Lemma 5.1

Proof. By the definition of the oracle weight function, for every $t\in\mathbb{R}$,

$$A_{w_0}(t)=\mathbb{E}_{\mathrm{cal}}\left[w_0(X_1)\mathbf{1}\{S\le t\}\right]=\mathbb{P}_{\mathrm{test}}(T\le t)=F_{\mathrm{test}}(t).$$

Moreover, $\mu_{w_0}=\mathbb{E}_{\mathrm{cal}}[w_0(X_1)]=1$. Therefore,

$$F_{w_0,n}(t)=\frac{A_{w_0}(t)}{\mu_{w_0}+a_n}=\frac{F_{\mathrm{test}}(t)}{1+a_n}.$$

Hence, by the definition of $G_{w,n}$,

$$G_{w_0,n}(\eta)=\mathbb{P}_{\mathrm{test}}\left(F_{w_0,n}(T)\ge 1-\eta\right)=\mathbb{P}_{\mathrm{test}}\left(F_{\mathrm{test}}(T)\ge(1-\eta)(1+a_n)\right).$$

Under Assumption 2.2, F_{test} is continuous, the probability integral transform gives

$$F_{\mathrm{test}}(T)\sim\mathrm{Unif}(0,1).$$

Therefore,

$$G_{w_0,n}(\eta)=\left[1-(1-\eta)(1+a_n)\right]_+\le\eta$$

This proves the claim. □

B.4 Proof of the Consistency of Algorithm 1 - Theorem 5.3

Proof. Recall that

$$A_w(t):=\mathbb{E}_{\mathrm{cal}}\left[w(X)\mathbf{1}\{S\le t\}\right],\qquad \mu_w:=\mathbb{E}_{\mathrm{cal}}[w(X)],\qquad \widehat{\mu}_w:=\frac{1}{n}\sum_{i=1}^n w(X_i),$$

Define

$$r_n(\delta):=\varepsilon_w\sqrt{\frac{\log(2/\delta)}{2n}}$$

and the normalized weighted calibration score CDF

$$H_w(t):=\frac{A_w(t)}{\mu_w}.$$

By Assumption 5.2,

$$H_w(t)=F_{\mathrm{test}}(t)\qquad\text{for every } t\in\mathbb{R}.$$

Moreover, Assumption 2.2 implies that F_{test} is continuous. Therefore, by the probability integral transform,

$$H_w(T) = F_{\text{test}}(T) \sim \text{Unif}(0,1).$$

Let $\lambda_n := \frac{\mu_w}{\mu_w + a_n}$. Since

$$F_{w,n}(t) = \frac{A_w(t)}{\mu_w + a_n} = \lambda_n H_w(t),$$

we obtain, for every $\beta \in (0,1)$,

$$\begin{aligned} G_{w,n}(\beta) &= P_{\text{test}}\left(F_{w,n}(T) \geq 1-\beta\right) \\ &= P_{\text{test}}\left(\lambda_n H_w(T) \geq 1-\beta\right) \\ &= \left[1 - \frac{1-\beta}{\lambda_n}\right]_+, \end{aligned}$$

where $[x]_+ := \max\{x, 0\}$. We next characterize the output of Algorithm 1. By the continuity of F_{cal}, the calibration scores are distinct almost surely. Let

$$S_{(1)} < S_{(2)} < \cdots < S_{(n)}$$

denote the ordered calibration scores, and let $W_{(r)}$ denote the weight corresponding to $S_{(r)}$. Define

$$C_r := \sum_{j=1}^{r} W_{(j)}, \qquad P_r := \frac{C_r}{C_n}, \qquad P_0 := 0.$$

Also define $\widehat{\lambda}_n := \frac{\widehat{\mu}_w}{\widehat{\mu}_w + a_n}$. Since $C_n = n\widehat{\mu}_w$, Algorithm 1 satisfies

$$\widehat{F}_{w,n}(S_{(r)}) = \frac{C_r/n}{\widehat{\mu}_w + a_n} = \widehat{\lambda}_n P_r.$$

For $\beta \in (0,1)$, set $q_\beta := \frac{1-\beta}{\widehat{\lambda}_n}$. If $q_\beta > 1$, then

$$\widehat{G}^{\text{IW}}_{w,n}(\beta) = \left[1 - \frac{1-\beta}{\widehat{\lambda}_n}\right]_+ = 0.$$

Suppose now that $q_\beta \leq 1$, and define

$$r_\beta := \min\left\{r \in [n] : P_r \geq q_\beta\right\}.$$

Then $P_{r_\beta - 1} < q_\beta \leq P_{r_\beta}$, and Algorithm 1 returns

$$\widehat{G}^{\text{IW}}_{w,n}(\beta) = \sum_{r=r_\beta}^{n} \frac{W_{(r)}}{C_n} = 1 - P_{r_\beta - 1}.$$

It follows that

$$\begin{aligned} 0 &\leq \widehat{G}^{\text{IW}}_{w,n}(\beta) - \left[1 - \frac{1-\beta}{\widehat{\lambda}_n}\right]_+ \\ &= q_\beta - P_{r_\beta - 1} \\ &\leq P_{r_\beta} - P_{r_\beta - 1} \\ &= \frac{W_{(r_\beta)}}{C_n}. \end{aligned}$$

Consequently,

$$\sup_{\beta \in (0,1)} \left| \widehat{G}^{\text{IW}}_{w,n}(\beta) - \left[1 - \frac{1-\beta}{\widehat{\lambda}_n}\right]_+ \right| \leq \max_{1 \leq i \leq n} \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j)}.$$

For any $\lambda, \widetilde{\lambda} \in (0,1]$, a direct calculation gives

$$\sup_{\beta \in (0,1)} \left| \left[1 - \frac{1-\beta}{\widetilde{\lambda}}\right]_+ - \left[1 - \frac{1-\beta}{\lambda}\right]_+ \right| = \frac{|\widetilde{\lambda} - \lambda|}{\max\{\widetilde{\lambda}, \lambda\}}.$$

Therefore,

$$\sup_{\beta\in(0,1)}\left|\widehat{G}^{\mathrm{IW}}_{w,n}(\beta)-G_{w,n}(\beta)\right| \le \max_{1\le i\le n}\frac{w(X_i)}{\sum_{j=1}^n w(X_j)}+\frac{|\widehat{\lambda}_n-\lambda_n|}{\max\{\widehat{\lambda}_n,\lambda_n\}}. \tag{36}$$

By Assumption 2.3, $0 < w(X_i) \le \varepsilon_w$. Hence, Hoeffding's inequality gives

$$\mathbb{P}\left(|\widehat{\mu}_w-\mu_w|>r_n(\delta)\right)\le\delta.$$

Consider the event

$$\mathcal{E}_n(\delta):=\left\{|\widehat{\mu}_w-\mu_w|\le r_n(\delta)\right\},$$

which has probability at least $1-\delta$. On $\mathcal{E}_n(\delta)$,

$$\widehat{\mu}_w\ge\mu_w-r_n(\delta)>0,$$

and thus

$$\max_{1\le i\le n}\frac{w(X_i)}{\sum_{j=1}^n w(X_j)}\le\frac{\varepsilon_w}{n\big(\mu_w-r_n(\delta)\big)}.$$

Furthermore,

$$\left|\widehat{\lambda}_n-\lambda_n\right|=\frac{a_n\,|\widehat{\mu}_w-\mu_w|}{(\widehat{\mu}_w+a_n)(\mu_w+a_n)}.$$

Let

$$M_n:=\max\{\widehat{\mu}_w,\mu_w\},\qquad m_n:=\min\{\widehat{\mu}_w,\mu_w\}.$$

Since $u\mapsto u/(u+a_n)$ is increasing,

$$\max\{\widehat{\lambda}_n,\lambda_n\}=\frac{M_n}{M_n+a_n}.$$

Therefore,

$$\frac{|\widehat{\lambda}_n-\lambda_n|}{\max\{\widehat{\lambda}_n,\lambda_n\}}=\frac{a_n|\widehat{\mu}_w-\mu_w|}{M_n(m_n+a_n)}\le\frac{a_nr_n(\delta)}{\mu_w\big(\mu_w-r_n(\delta)\big)},$$

where the final inequality follows from

$$M_n\ge\mu_w,\qquad m_n\ge\mu_w-r_n(\delta)$$

on $\mathcal{E}_n(\delta)$. Substituting these two bounds into equation 36 proves that, with probability at least $1-\delta$,

$$\sup_{\beta\in(0,1)}\left|\widehat{G}^{\mathrm{IW}}_{w,n}(\beta)-G_{w,n}(\beta)\right|\le\frac{\varepsilon_w}{n\big(\mu_w-r_n(\delta)\big)}+\frac{a_nr_n(\delta)}{\mu_w\big(\mu_w-r_n(\delta)\big)}.$$

Finally, since $\mu_w > 0$ and $\varepsilon_w < \infty$ are fixed,

$$r_n(\delta)=O(n^{-1/2}).$$

It follows that

$$\sup_{\beta\in(0,1)}\left|\widehat{G}^{\mathrm{IW}}_{w,n}(\beta)-G_{w,n}(\beta)\right|=O_{\mathbb{P}}\left(\frac{1}{n}+\frac{a_n}{\sqrt{n}}\right).$$

Since w_∞ is fixed, $a_n = O(n^{-1})$, and hence

$$\sup_{\beta\in(0,1)}\left|\widehat{G}^{\mathrm{IW}}_{w,n}(\beta)-G_{w,n}(\beta)\right|=O_{\mathbb{P}}(n^{-1}).$$

This proves the result. □

B.5 Proof of the Inverse FCP Bound for Fixed Miscoverage- Corollary 5.5

Proof. Fix $\beta \in (0,1)$ and $\eta > 0$. Consider

$$G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \Delta_{\text{fix}} - \eta.$$

If

$$G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \Delta_{\text{fix}} - \eta \le 0,$$

then, by our convention for nonpositive miscoverage levels,

$$\text{FCP}_m^{w,(n)}\Big[\big(G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \Delta_{\text{fix}} - \eta\big)_+\Big] = 0 \le \beta,$$

and the claim holds trivially.

It therefore remains to consider the case

$$G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \Delta_{\text{fix}} - \eta > 0.$$

In particular, this implies $\beta - \varepsilon_{\text{fix}} \ge 0$. By the definition

$$G^+_{w,n}(u) := \inf\{\gamma \in [0,1) : G_{w,n}(\gamma) > u\},$$

for any $\eta > 0$ such that $G^+_{w,n}(u) - \eta \ge 0$, we have

$$G_{w,n}\big(G^+_{w,n}(u) - \eta\big) \le u.$$

Indeed, if

$$G_{w,n}\big(G^+_{w,n}(u) - \eta\big) > u,$$

then $G^+_{w,n}(u) - \eta$ would belong to the set whose infimum defines $G^+_{w,n}(u)$, contradicting

$$G^+_{w,n}(u) - \eta < G^+_{w,n}(u).$$

Taking $u = \beta - \varepsilon_{\text{fix}}$ therefore gives

$$G_{w,n}\big(G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \eta\big) \le \beta - \varepsilon_{\text{fix}}.$$

Now apply Theorem 4.1 at the fixed miscoverage level

$$G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \Delta_{\text{fix}} - \eta.$$

With probability at least $1 - \delta_{\text{cal}} - \delta_{\text{test}}$,

$$\begin{aligned}
&\text{FCP}_m^{w,(n)}\big[G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \Delta_{\text{fix}} - \eta\big] \\
&\quad \le G_{w,n}\big(G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \Delta_{\text{fix}} - \eta + \Delta_{\text{fix}}\big) + \varepsilon_{\text{fix}} \\
&\quad = G_{w,n}\big(G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \eta\big) + \varepsilon_{\text{fix}} \\
&\quad \le \beta - \varepsilon_{\text{fix}} + \varepsilon_{\text{fix}} = \beta.
\end{aligned}$$

Hence,

$$\mathbb{P}\Big(\text{FCP}_m^{w,(n)}\big[G^+_{w,n}(\beta - \varepsilon_{\text{fix}}) - \Delta_{\text{fix}} - \eta\big] \le \beta\Big) \ge 1 - \delta_{\text{cal}} - \delta_{\text{test}},$$

which proves the result. □

B.6 Proof of the Inverse FCP Bound Uniformly over Miscoverage - Corollary 5.6

Proof. By Theorem 4.2, with probability at least $1 - \delta_+ - \delta_- - \delta_{\text{test}}$, we have simultaneously for all $\alpha \in (0,1)$ that

$$\text{FCP}_m^{w,(n)}(\alpha) \le G_{w,n}(\alpha + \Delta_{\text{unif}}) + \varepsilon_{\text{unif}}.$$

We work on this event.

Fix any $\beta \in (0,1)$. If

$$G^+_{w,n}(\beta - \varepsilon_{\text{unif}}) - \Delta_{\text{unif}} - \eta \le 0,$$

then, by our convention for nonpositive miscoverage levels,

$$\mathrm{FCP}_m^{w,(n)}\Big[\big(G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\Delta_{\mathrm{unif}}-\eta\big)_+\Big]=0\le\beta.$$

It therefore remains to consider the case

$$G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\Delta_{\mathrm{unif}}-\eta>0.$$

Since $\Delta_{\mathrm{unif}}\ge 0$, this implies

$$G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\eta>0.$$

By the definition

$$G_{w,n}^+(u):=\inf\{\gamma\in[0,1):G_{w,n}(\gamma)>u\},$$

we have

$$G_{w,n}\big(G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\eta\big)\le\beta-\varepsilon_{\mathrm{unif}}.$$

Indeed, otherwise

$$G_{w,n}\big(G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\eta\big)>\beta-\varepsilon_{\mathrm{unif}},$$

so $G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\eta$ would belong to the set defining $G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})$, contradicting the definition of the infimum.

Applying the uniform bound from Theorem 4.2 at

$$\alpha=G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\Delta_{\mathrm{unif}}-\eta,$$

we obtain

$$\begin{aligned}&\mathrm{FCP}_m^{w,(n)}\big[G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\Delta_{\mathrm{unif}}-\eta\big]\\&\quad\le G_{w,n}\big(G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\Delta_{\mathrm{unif}}-\eta+\Delta_{\mathrm{unif}}\big)+\varepsilon_{\mathrm{unif}}\\&\quad=G_{w,n}\big(G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\eta\big)+\varepsilon_{\mathrm{unif}}\\&\quad\le\beta-\varepsilon_{\mathrm{unif}}+\varepsilon_{\mathrm{unif}}=\beta.\end{aligned}$$

Since the event in Theorem 4.2 holds simultaneously for all $\alpha\in(0,1)$, the preceding argument holds simultaneously for every $\beta\in(0,1)$. Therefore,

$$\mathbb{P}\Big(\forall\beta\in(0,1),\quad\mathrm{FCP}_m^{w,(n)}\big[G_{w,n}^+(\beta-\varepsilon_{\mathrm{unif}})-\Delta_{\mathrm{unif}}-\eta\big]\le\beta\Big)\ge 1-\delta_+-\delta_--\delta_{\mathrm{test}}.$$

□

B.7 Proof of the Consistency of Algorithm 2 - Theorem 5.7

Proof. Recall that

$$A_w(t):=\mathbb{E}_{\mathrm{cal}}\left[w(X)\mathbf{1}\{S\le t\}\right],\qquad H_w(t):=\frac{A_w(t)}{\mu_w},$$

Define

$$r_n(\delta):=\varepsilon_w\sqrt{\frac{\log(2/\delta)}{2n}}\text{ and }\lambda_n:=\frac{\mu_w}{\mu_w+a_n}.$$

By Assumption 5.2,

$$H_w(t)=F_{\mathrm{test}}(t)$$

for every $t\in\mathbb{R}$. Moreover, Assumption 2.2 implies that F_{test} is continuous. Therefore, by the probability integral transform,

$$H_w(T)=F_{\mathrm{test}}(T)\sim\mathrm{Unif}(0,1).$$

Since

$$F_{w,n}(t)=\frac{A_w(t)}{\mu_w+a_n}=\lambda_nH_w(t),$$

we have, for every $\beta \in (0,1)$,

$$\begin{aligned} G_{w,n}(\beta) &= \mathbb{P}_{\text{test}}\left(F_{w,n}(T) \geq 1-\beta\right) \\ &= \mathbb{P}_{\text{test}}\left(\lambda_n H_w(T) \geq 1-\beta\right) \\ &= \left[1 - \frac{1-\beta}{\lambda_n}\right]_+ . \end{aligned}$$

It follows from the definition

$$G^+_{w,n}(u) := \inf\left\{\eta \in [0,1) : G_{w,n}(\eta) > u\right\}$$

that

$$G^+_{w,n}(u) = 1 - \lambda_n(1-u), \qquad u \in [0,1).$$

We next characterize the output of Algorithm 2. By the continuity of F_{cal}, the calibration scores are distinct almost surely. Let

$$S_{(1)} < S_{(2)} < \cdots < S_{(n)}$$

denote the ordered calibration scores, and let $W_{(r)}$ be the weight corresponding to $S_{(r)}$. Define

$$C_r := \sum_{j=1}^{r} W_{(j)}, \qquad P_r := \frac{C_r}{C_n}, \qquad P_0 := 0,$$

and

$$\widehat{\lambda}_n := \frac{\widehat{\mu}_w}{\widehat{\mu}_w + a_n}, \qquad \widehat{\mu}_w := \frac{1}{n}\sum_{i=1}^{n} w(X_i).$$

Since $C_n = n\widehat{\mu}_w$, we have

$$\widehat{F}_{w,n}(S_{(r)}) = \frac{C_r/n}{\widehat{\mu}_w + a_n} = \widehat{\lambda}_n P_r.$$

For $u \in [0,1)$, define

$$r(u) := \min\left\{r \in [n] : P_r \geq 1-u\right\}.$$

By the minimality of $r(u)$,

$$P_{r(u)-1} < 1 - u \leq P_{r(u)}.$$

The reverse cumulative weight associated with $S_{(r)}$ in Algorithm 2 is $1 - P_{r-1}$. Hence, the first reverse cumulative weight that is strictly larger than u corresponds to $r(u)$. Therefore, Algorithm 2 returns

$$\left(\widehat{G}^{\text{IW}}_{w,n}\right)^+(u) = 1 - \widehat{F}_{w,n}(S_{(r(u))}) = 1 - \widehat{\lambda}_n P_{r(u)}.$$

Consequently,

$$\begin{aligned} 0 &\leq \left[1 - \widehat{\lambda}_n(1-u)\right] - \left(\widehat{G}^{\text{IW}}_{w,n}\right)^+(u) \\ &= \widehat{\lambda}_n\left(P_{r(u)} - (1-u)\right) \\ &\leq \widehat{\lambda}_n\left(P_{r(u)} - P_{r(u)-1}\right) \\ &\leq \widehat{\lambda}_n \max_{1\leq r\leq n} \frac{W_{(r)}}{C_n}. \end{aligned}$$

Using $C_n = n\widehat{\mu}_w$ gives

$$\widehat{\lambda}_n \max_{1\leq r\leq n} \frac{W_{(r)}}{C_n} = \frac{\max_{1\leq i\leq n} w(X_i)}{n(\widehat{\mu}_w + a_n)}.$$

Furthermore,

$$\begin{aligned} &\left|\left[1 - \widehat{\lambda}_n(1-u)\right] - G^+_{w,n}(u)\right| \\ &\quad = (1-u)\left|\widehat{\lambda}_n - \lambda_n\right| \leq \left|\widehat{\lambda}_n - \lambda_n\right|. \end{aligned}$$

The triangle inequality therefore yields

$$\sup_{u\in[0,1)} \left|\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}(u) - G^{+}_{w,n}(u)\right| \leq \frac{\max_{1\leq i\leq n} w(X_i)}{n(\widehat{\mu}_w + a_n)} + \left|\widehat{\lambda}_n - \lambda_n\right|.$$

By Assumption 2.3,

$$0 < w(X_i) \leq \varepsilon_w.$$

Hoeffding's inequality gives

$$\mathbb{P}\left(|\widehat{\mu}_w - \mu_w| > r_n(\delta)\right) \leq \delta.$$

Consider the event

$$E_n(\delta) := \left\{|\widehat{\mu}_w - \mu_w| \leq r_n(\delta)\right\},$$

which has probability at least $1-\delta$. On $E_n(\delta)$,

$$\widehat{\mu}_w \geq \mu_w - r_n(\delta) > 0.$$

It follows that

$$\frac{\max_{1\leq i\leq n} w(X_i)}{n(\widehat{\mu}_w + a_n)} \leq \frac{\varepsilon_w}{n\big(\mu_w - r_n(\delta) + a_n\big)}.$$

Moreover,

$$\begin{aligned}\left|\widehat{\lambda}_n - \lambda_n\right| &= \left|\frac{\widehat{\mu}_w}{\widehat{\mu}_w + a_n} - \frac{\mu_w}{\mu_w + a_n}\right| \\ &= \frac{a_n\left|\widehat{\mu}_w - \mu_w\right|}{(\widehat{\mu}_w + a_n)(\mu_w + a_n)} \\ &\leq \frac{a_n r_n(\delta)}{\big(\mu_w - r_n(\delta) + a_n\big)(\mu_w + a_n)}.\end{aligned}$$

Combining the preceding inequalities proves that, with probability at least $1-\delta$,

$$\sup_{u\in[0,1)} \left|\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}(u) - G^{+}_{w,n}(u)\right| \leq \frac{\varepsilon_w}{n\big(\mu_w - r_n(\delta) + a_n\big)} + \frac{a_n r_n(\delta)}{\big(\mu_w - r_n(\delta) + a_n\big)(\mu_w + a_n)}.$$

Finally, since $\mu_w > 0$ and $\varepsilon_w < \infty$ are fixed,

$$r_n(\delta) = O(n^{-1/2}).$$

Therefore,

$$\sup_{u\in[0,1)} \left|\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}(u) - G^{+}_{w,n}(u)\right| = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{a_n}{\sqrt{n}}\right).$$

Since w_∞ is fixed, $a_n = w_\infty/n = O(n^{-1})$, and hence

$$\sup_{u\in[0,1)} \left|\left(\widehat{G}^{\mathrm{IW}}_{w,n}\right)^{+}(u) - G^{+}_{w,n}(u)\right| = O_{\mathbb{P}}(n^{-1}).$$

This proves the result. □

B.8 Proof of the Consistency of Algorithm 3 - Theorem C.2

Proof. Recall that

$$A_w(t) := \mathbb{E}_{\mathrm{cal}}\left[w(X)\mathbf{1}\{S \leq t\}\right], \qquad \mu_w := \mathbb{E}_{\mathrm{cal}}[w(X)],$$

and define

$$\widehat{A}_w(t) := \frac{1}{n}\sum_{i=1}^{n} w(X_i)\mathbf{1}\{S_i \leq t\}, \qquad \widehat{\mu}_w := \frac{1}{n}\sum_{i=1}^{n} w(X_i).$$

Let

$$a_n := \frac{w_\infty}{n}.$$

Then

$$F_{w,n}(t) = \frac{A_w(t)}{\mu_w + a_n}, \qquad \widehat{F}_{w,n}(t) = \frac{\widehat{A}_w(t)}{\widehat{\mu}_w + a_n}.$$

By Assumption 2.3, the class

$$\{(x,y) \mapsto w(x)\mathbf{1}\{s(x,y) \le t\} : t \in \mathbb{R}\}$$

is a uniformly bounded VC-subgraph class. Hence,

$$\sup_{t\in\mathbb{R}} \left|\widehat{A}_w(t) - A_w(t)\right| = O_{\mathbb{P}}(n^{-1/2}),$$

while, by the law of large numbers and Hoeffding's inequality,

$$|\widehat{\mu}_w - \mu_w| = O_{\mathbb{P}}(n^{-1/2}).$$

Since $\mu_w > 0$ and $a_n \ge 0$, it follows that

$$D_n := \sup_{t\in\mathbb{R}} \left|\widehat{F}_{w,n}(t) - F_{w,n}(t)\right| = O_{\mathbb{P}}(n^{-1/2}). \tag{37}$$

We next characterize the output of Algorithm 3. Let $T_1, \dots, T_h$ denote the holdout test scores. By construction,

$$\widehat{G}^{\mathrm{HO}}_{w,n}(\beta) = \frac{1}{h}\sum_{j=1}^{h} \mathbf{1}\left\{\widehat{F}_{w,n}(T_j) \ge 1-\beta\right\}.$$

Conditionally on the calibration sample, define

$$\widetilde{G}_{w,n}(\beta) := P_{\mathrm{test}}\left(\widehat{F}_{w,n}(T) \ge 1-\beta\right).$$

Since the holdout test observations are i.i.d. from P_{test} and independent of the calibration sample, the Dvoretzky–Kiefer–Wolfowitz inequality yields

$$\sup_{\beta\in(0,1)} \left|\widehat{G}^{\mathrm{HO}}_{w,n}(\beta) - \widetilde{G}_{w,n}(\beta)\right| = O_{\mathbb{P}}(h^{-1/2}). \tag{38}$$

It remains to compare $\widetilde{G}_{w,n}$ with $G_{w,n}$. Recall that

$$G_{w,n}(\beta) = P_{\mathrm{test}}\left(F_{w,n}(T) \ge 1-\beta\right).$$

On the event $\{D_n \le d\}$, for every $\beta \in (0,1)$,

$$\{F_{w,n}(T) \ge 1-\beta+d\} \subseteq \left\{\widehat{F}_{w,n}(T) \ge 1-\beta\right\} \subseteq \{F_{w,n}(T) \ge 1-\beta-d\}.$$

Therefore,

$$G_{w,n}(\beta-d) \le \widetilde{G}_{w,n}(\beta) \le G_{w,n}(\beta+d),$$

where $G_{w,n}$ is extended in the natural way outside $(0,1)$. Consequently,

$$\sup_{\beta\in(0,1)} \left|\widetilde{G}_{w,n}(\beta) - G_{w,n}(\beta)\right| \le \omega_{G_{w,n}}(D_n), \tag{39}$$

where

$$\omega_{G_{w,n}}(d) := \sup_{\substack{\beta,\beta'\in[0,1]\\ |\beta-\beta'|\le d}} |G_{w,n}(\beta) - G_{w,n}(\beta')|$$

denotes the modulus of continuity.

To control this term, write

$$H_w(t) := \frac{A_w(t)}{\mu_w}, \qquad \lambda_n := \frac{\mu_w}{\mu_w + a_n}.$$

Then

$$F_{w,n}(t) = \lambda_n H_w(t),$$

and hence

$$G_{w,n}(\beta) = P_{\text{test}}\left(\lambda_n H_w(T) \geq 1 - \beta\right).$$

Since w_∞ is fixed, $a_n = O(n^{-1})$ and therefore $\lambda_n \to 1$. By assumption, $H_w(T)$ has a continuous distribution. Its distribution function is therefore uniformly continuous on $[0, 1]$, which implies

$$\omega_{G_{w,n}}(d) \longrightarrow 0 \qquad \text{as } d \downarrow 0,$$

uniformly for all sufficiently large n. Together with equation 37, this gives

$$\sup_{\beta\in(0,1)} \left|\widetilde{G}_{w,n}(\beta) - G_{w,n}(\beta)\right| = o_{\mathbb{P}}(1).$$

Combining this with equation 38 yields

$$\sup_{\beta\in(0,1)} \left|\widehat{G}^{\text{HO}}_{w,n}(\beta) - G_{w,n}(\beta)\right| = o_{\mathbb{P}}(1),$$

provided $n, h \to \infty$.

This proves the result. □

B.9 Proof of the Consistency of Algorithm 4 - Theorem C.3

Proof. Recall that

$$A_w(t) := \mathbb{E}_{\text{cal}}\left[w(X)\mathbf{1}\{S \leq t\}\right], \qquad \mu_w := \mathbb{E}_{\text{cal}}[w(X)],$$

and define

$$\widehat{A}_w(t) := \frac{1}{n}\sum_{i=1}^{n} w(X_i)\mathbf{1}\{S_i \leq t\}, \qquad \widehat{\mu}_w := \frac{1}{n}\sum_{i=1}^{n} w(X_i).$$

Let

$$a_n := \frac{w_\infty}{n}.$$

Then

$$F_{w,n}(t) = \frac{A_w(t)}{\mu_w + a_n}, \qquad \widehat{F}_{w,n}(t) = \frac{\widehat{A}_w(t)}{\widehat{\mu}_w + a_n}.$$

As in the proof for Algorithm 3, Assumption 2.3 and uniform convergence of the weighted empirical score distribution give

$$D_n := \sup_{t\in\mathbb{R}} \left|\widehat{F}_{w,n}(t) - F_{w,n}(t)\right| = O_{\mathbb{P}}(n^{-1/2}). \tag{40}$$

For the holdout test scores $T_1, \ldots, T_h$, define

$$V_j := 1 - F_{w,n}(T_j), \qquad \widehat{V}_j := 1 - \widehat{F}_{w,n}(T_j).$$

It follows immediately from equation 40 that

$$\max_{1\leq j\leq h} |\widehat{V}_j - V_j| \leq D_n.$$

Let

$$V_{(1)} \leq \cdots \leq V_{(h)}, \qquad \widehat{V}_{(1)} \leq \cdots \leq \widehat{V}_{(h)}$$

denote the corresponding order statistics. Since perturbing every observation by at most D_n perturbs every order statistic by at most D_n,

$$\max_{1\leq j\leq h} \left|\widehat{V}_{(j)} - V_{(j)}\right| \leq D_n. \tag{41}$$

Let $\widetilde{G}_{w,n,h}$ denote the empirical CDF of $V_1, \ldots, V_h$. By the construction of Algorithm 4,

$$(\widehat{G}^{\mathrm{HO}}_{w,n})^+(u) = \widehat{V}_{(k(u))}, \qquad k(u) := \min\left\{ j \in [h] : \frac{j}{h} > u \right\},$$

whereas

$$\widetilde{G}^+_{w,n,h}(u) = V_{(k(u))}.$$

Therefore, from equation 41,

$$\sup_{u\in[0,1)} \left| (\widehat{G}^{\mathrm{HO}}_{w,n})^+(u) - \widetilde{G}^+_{w,n,h}(u) \right| \le D_n = O_{\mathbb{P}}(n^{-1/2}). \tag{42}$$

It remains to compare the empirical inverse $\widetilde{G}^+_{w,n,h}$ with the population inverse $G^+_{w,n}$. Define

$$H_w(t) := \frac{A_w(t)}{\mu_w}, \qquad \lambda_n := \frac{\mu_w}{\mu_w + a_n}.$$

Then

$$F_{w,n}(t) = \lambda_n H_w(t),$$

and hence

$$V_j = 1 - \lambda_n H_w(T_j) = (1-\lambda_n) + \lambda_n U_j, \qquad U_j := 1 - H_w(T_j).$$

Let K denote the CDF of $U = 1 - H_w(T)$ and K^+ its generalized inverse. Since the holdout observations are i.i.d. from P_{test}, the variables $U_1, \ldots, U_h$ are i.i.d. with CDF K. Moreover,

$$G^+_{w,n}(u) = (1-\lambda_n) + \lambda_n K^+(u).$$

Let $\widehat{K}_h$ be the empirical CDF of $U_1, \ldots, U_h$. The same affine transformation gives

$$\widetilde{G}^+_{w,n,h}(u) = (1-\lambda_n) + \lambda_n \widehat{K}^+_h(u).$$

Consequently,

$$\sup_{u\in[0,1)} \left| \widetilde{G}^+_{w,n,h}(u) - G^+_{w,n}(u) \right| \le \sup_{u\in[0,1)} \left| \widehat{K}^+_h(u) - K^+(u) \right|. \tag{43}$$

By the Dvoretzky–Kiefer–Wolfowitz inequality,

$$\sup_{v\in[0,1]} |\widehat{K}_h(v) - K(v)| = O_{\mathbb{P}}(h^{-1/2}).$$

Under the holdout regularity assumption, both K and K^+ are continuous. Hence, uniform convergence of $\widehat{K}_h$ to K implies uniform convergence of their generalized inverses:

$$\sup_{u\in[0,1)} \left| \widehat{K}^+_h(u) - K^+(u) \right| = o_{\mathbb{P}}(1).$$

Combining this with equation 42 and equation 43 yields

$$\sup_{u\in[0,1)} \left| (\widehat{G}^{\mathrm{HO}}_{w,n})^+(u) - G^+_{w,n}(u) \right| = o_{\mathbb{P}}(1).$$

This proves the result. □

C Further Discussion and Analysis

C.1 On the Convention of Fixed versus Adaptive Test Weights

Trade-off between fixed and adaptive test weights. Previous work on Conformal Prediction under Distribution Shift typically assigns data-dependent weights to the calibration observations, so that each calibration point contributes according to its relevance to the target distribution. For the test observations, however, two conventions are commonly used. The first is an *adaptive* test weight,

where the weight depends on the particular test covariate (Barber et al., 2023; Tibshirani et al., 2019), leading to a value $w_{\infty,i}$ for each X_{n+i}. The second is a *fixed* test weight, where the same value w_∞ is assigned to every test observation (Gazin, 2024; Prinzhorn et al., 2024). These two choices reflect different design priorities. An adaptive test weight allows the conformal procedure to account for heterogeneity across the test covariates and can therefore provide a more locally tailored correction. Under Covariate Shift, for example, one may take $w_{\infty,i} = w_0(X_{n+i})$ (see equation 5), so that the test contribution directly reflects the local density ratio at X_{n+i}. By contrast, a fixed test weight treats the test observations symmetrically at the level of the additional conformal mass, so that all test points are evaluated relative to the same weighted calibration distribution. This sacrifices some local adaptivity, but yields a common reference distribution across the entire test batch and leads to a considerably simpler joint structure for the corresponding conformal p-values.

Moreover, for adaptive test weights, the additional flexibility they provide comes at the cost of a more complicated joint analysis across the test batch. When the test weight varies from one observation to another, each test point is effectively associated with a different weighted construction, including its own normalization and potentially its own decision threshold. As a consequence, the resulting quantities are no longer generated under a common reference mechanism, and their dependence structure becomes heterogeneous across the batch. This makes it more difficult to obtain a unified probabilistic characterization, since the analysis must account simultaneously for variation induced by the data and variation induced by the test-specific weights. If the weights are estimated rather than known exactly, an additional layer of randomness is introduced, and estimation errors can propagate differently across test observations.

Therefore, the choice between fixed and adaptive test weights reflects a trade-off between local adaptivity and theoretical tractability: adaptive weights can better tailor the procedure to individual test points, whereas fixed weights yield a more stable and analytically transparent structure for joint theoretical analysis.

The Choice of Fixed Test Weight in Our Work. In this work, our primary objective is to develop finite-sample theoretical guarantees for the FCP under Distribution Shift. We therefore adopt a fixed test weight w_∞ shared across all test observations. This convention aligns naturally with the formulation used in prior theoretical analyses of FCP, in particular the asymptotic framework of Gazin (2024), and provides a common weighted reference distribution across the entire test batch. As a result, the resulting conformal construction admits a cleaner joint probabilistic structure, which is well suited to the development of explicit finite-sample bounds and facilitates a direct comparison between our results and the existing asymptotic theory.

The use of a fixed test weight does sacrifice some degree of local adaptation, since the contribution of a test point does not vary with its individual covariate value. In the present setting, however, this limitation is compatible with our main goal: we are primarily concerned with obtaining rigorous batch-level guarantees under Distribution Shift, rather than optimizing the prediction set separately for each test observation. Moreover, the calibration weights remain data dependent and can still encode substantial information about the discrepancy between the calibration and test distributions. We therefore view the fixed-weight formulation as an appropriate compromise between adaptability and a transparent theoretical framework for FCP analysis.

Extending the present results to adaptive test weights would nevertheless be a natural and practically relevant direction for future work. Such an extension could retain the local responsiveness of test-specific weighting while seeking to preserve finite-sample control at the batch level, but would require a more refined treatment of the additional heterogeneity introduced across test observations.

C.2 Relaxing Assumption 5.2 Using Holdout Test Data

In Section 5, we introduced Assumption 5.2 to establish the consistency of Algorithms 1 and 2, which in turn makes the direct and inverse FCP bounds practically estimable from observable data. For completeness, we recall the assumption below.

Assumption 5.2. The weight function w satisfies

$$P_{\text{test}}(T \le t) = \mu_w^{-1}\mathbb{E}_{\text{cal}}[w(X)\mathbf{1}\{S \le t\}], \qquad \text{for every } t \in \mathbb{R}.$$

While Assumption 5.2 is already relatively mild, as it only requires the weighting scheme to capture the calibration–test discrepancy sufficiently well at the level of the score distribution, it may still be

Algorithm 3 Estimation of $G_{w,n}(\cdot)$ with labeled holdout test data

Require: Calibration pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, holdout test pairs $(X_1^{\mathrm{ho}}, Y_1^{\mathrm{ho}}), \ldots, (X_h^{\mathrm{ho}}, Y_h^{\mathrm{ho}})$, weight function w, additional weight w_∞.

Ensure: Estimate $\widehat{G}_{w,n}^{\mathrm{HO}}(\cdot)$ of $G_{w,n}(\cdot)$.

1: Compute calibration scores

$$S_i := s(X_i, Y_i), \qquad i \in [n].$$

2: Compute holdout test scores

$$T_j := s(X_j^{\mathrm{ho}}, Y_j^{\mathrm{ho}}), \qquad j \in [h].$$

3: Compute the calibration weights

$$W_i := w(X_i), \qquad i = 1, \ldots, n.$$

4: Compute

$$\widehat{\mu}_w := \frac{1}{n} \sum_{i=1}^{n} W_i, \qquad a_n := \frac{w_\infty}{n}.$$

5: Sort the calibration scores:

$$S_{(1)} \le S_{(2)} \le \cdots \le S_{(n)},$$

and let $W_{(i)}$ denote the weight corresponding to $S_{(i)}$.

6: Compute the cumulative weighted sums

$$C_0 := 0, \qquad C_k := \sum_{i=1}^{k} W_{(i)}, \qquad k = 1, \ldots, n.$$

7: **for** $j = 1, \ldots, h$ **do**

8: Find, with the convention $S_{(0)} = -\infty$,

$$k_j := \max\left\{k \in \{0, 1, \ldots, n\} : S_{(k)} \le T_j\right\}.$$

9: Compute

$$\widehat{A}_w(T_j) := \frac{C_{k_j}}{n}.$$

10: Compute

$$\widehat{F}_{w,n}(T_j) := \frac{\widehat{A}_w(T_j)}{\widehat{\mu}_w + a_n}.$$

11: **end for**

12: **return**

$$\widehat{G}_{w,n}^{\mathrm{HO}} : \mathbb{R} \to [0, 1], \qquad \beta \mapsto \frac{1}{h} \sum_{j=1}^{h} \mathbf{1}\left\{\widehat{F}_{w,n}(T_j) \ge 1 - \beta\right\}.$$

restrictive when such a transport relation cannot be justified directly from the available calibration data. In this section, we show that when a small amount of additional holdout data from the test distribution is available, Assumption 5.2 can be replaced by a substantially weaker requirement.

Role of Assumption 5.2. The need for Assumption 5.2 arises from the quantities targeted by Algorithms 1 and 2. Both procedures aim to estimate $G_{w,n}$ and its generalized inverse $G_{w,n}^{+}$, where $G_{w,n}$ is the population CDF of the ideal weighted conformal p-value under the test distribution. Consequently, evaluating $G_{w,n}$ fundamentally requires information about the distribution of conformity scores for observations drawn from P_{test}.

At calibration time, however, the responses of the future test observations are unavailable. Their conformity scores therefore cannot be computed, and the test-score distribution cannot be estimated directly from the test sample. Algorithms 1 and 2 circumvent this difficulty by using the weighted calibration-score distribution as a surrogate for the unavailable test-score distribution. More precisely,

Algorithm 4 Estimation of $(G_{w,n})^+(\cdot)$ with labeled holdout test data

Require: Calibration pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, holdout test pairs $(X_1^{\mathrm{ho}}, Y_1^{\mathrm{ho}}), \ldots, (X_h^{\mathrm{ho}}, Y_h^{\mathrm{ho}})$, weight function w, additional weight w_∞.

Ensure: Step-function representation of $(\widehat{G}_{w,n}^{\mathrm{HO}})^+(\cdot)$.

1: Compute calibration scores
$$S_i := s(X_i, Y_i), \qquad i \in [n].$$
2: Compute holdout test scores
$$T_j := s(X_j^{\mathrm{ho}}, Y_j^{\mathrm{ho}}), \qquad j \in [h].$$
3: Compute the calibration weights
$$W_i \leftarrow w(X_i), \qquad i = 1, \ldots, n.$$
4: Compute
$$\widehat{\mu}_w := \frac{1}{n}\sum_{i=1}^{n} W_i, \qquad a_n := \frac{w_\infty}{n}.$$
5: Sort the calibration scores:
$$S_{(1)} \le S_{(2)} \le \cdots \le S_{(n)},$$
and let $W_{(i)}$ denote the weight corresponding to $S_{(i)}$.
6: Compute the cumulative weighted sums
$$C_0 \leftarrow 0, \qquad C_k \leftarrow \sum_{i=1}^{k} W_{(i)}, \qquad k = 1, \ldots, n.$$
7: **for** $j = 1, \ldots, h$ **do**
8: Find, with the convention $S_{(0)} = -\infty$,
$$k_j \leftarrow \max\left\{k \in \{0, \ldots, n\} : S_{(k)} \le T_j\right\}.$$
9: Compute
$$\widehat{A}_w(T_j) \leftarrow \frac{C_{k_j}}{n}.$$
10: Compute
$$\widehat{F}_{w,n}(T_j) \leftarrow \frac{\widehat{A}_w(T_j)}{\widehat{\mu}_w + a_n}.$$
11: Compute
$$\widehat{V}_j \leftarrow 1 - \widehat{F}_{w,n}(T_j).$$
12: **end for**
13: Sort
$$\widehat{V}_{(1)} \le \widehat{V}_{(2)} \le \cdots \le \widehat{V}_{(h)}.$$
14: For $j = 1, \ldots, h$, set
$$Q_j \leftarrow \widehat{V}_{(j)}, \qquad H_j \leftarrow \frac{j}{h}, \qquad H_0 \leftarrow 0.$$
15: **return** the step function
$$(\widehat{G}_{w,n}^{\mathrm{HO}})^+(u) := Q_{k(u)}, \qquad k(u) := \min\left\{j \in \{1, \ldots, h\} : H_j > u\right\}, \qquad u < 1.$$

the calibration observations are reweighted through $w(X)$ so that their weighted score distribution is intended to reproduce the score distribution that would be observed under P_{test}.

Assumption 5.2 is precisely the condition that justifies this substitution. It requires

$$P_{\mathrm{test}}(T \le t) = \mu_w^{-1}\mathbb{E}_{\mathrm{cal}}[w(X)\mathbf{1}\{S \le t\}], \qquad t \in \mathbb{R},$$

and hence ensures that the population-weighted CDF constructed from the calibration distribution coincides with the target score CDF under the test distribution. Under this relation, the empirical weighted calibration scores provide a consistent proxy for the otherwise inaccessible test-score distribution, which in turn enables consistent estimation of $G_{w,n}$ and $G^{+}_{w,n}$. Without such a relation, increasing the calibration sample size alone would not generally eliminate the discrepancy between the weighted calibration-score distribution and the test-score distribution, so the resulting estimators could retain a systematic bias.

Relaxing Assumption 5.2 with Holdout Test Data. As discussed above, the role of Assumption 5.2 is to ensure that the weighting scheme is sufficiently accurate to recover the test-score distribution from the calibration data alone. We now consider the setting in which, in addition to the calibration sample, an independent labeled holdout sample from the test distribution is available:

$$\mathcal{D}_{\mathrm{ho}} := \left\{(X_1^{\mathrm{ho}}, Y_1^{\mathrm{ho}}), \ldots, (X_h^{\mathrm{ho}}, Y_h^{\mathrm{ho}})\right\} \overset{\text{i.i.d.}}{\sim} P_{\mathrm{test}},$$

independently of the calibration sample. We denote the corresponding holdout conformity scores by

$$T_j := s(X_j^{\mathrm{ho}}, Y_j^{\mathrm{ho}}), \qquad j \in [h].$$

Such a setting is common in practical machine-learning applications. For instance, when a model is transferred to a new hospital, geographic region, or customer population, a relatively small labeled validation set from the deployment population may be collected before full deployment. Similarly, in systems that are already deployed, labels may become available retrospectively for a subset of previously encountered test observations and can be reserved as an independent audit or calibration sample. A related situation arises in temporal deployment, where a small batch of recently labeled observations from the current environment is available while the main calibration sample was collected at an earlier time. In each case, the amount of labeled target-domain data may be too limited to replace the original calibration sample entirely, but it can nevertheless provide direct information about the test distribution.

With the additional holdout sample $\mathcal{D}_{\mathrm{ho}}$, we introduce Algorithms 3 and 4 as holdout-based counterparts of Algorithms 1 and 2, for estimating $G_{w,n}$ and $G^{+}_{w,n}$, respectively. The key difference is conceptually simple. In the original procedures, the weighted calibration scores are used as a surrogate for the unavailable test-score distribution, with the weight function w providing the transport from calibration to test. By contrast, the holdout-based procedures can estimate the relevant test-side distribution directly from the labeled holdout scores $T_1, \ldots, T_h$. Thus, the weight function is still used to construct the weighted conformal quantity $\widehat{F}_{w,n}$ from the calibration sample, but it no longer needs to reproduce the test-score distribution itself. This removes the main reason for imposing Assumption 5.2. However, establishing the consistency of Algorithms 3 and 4 still requires a mild regularity condition on the transformed test-score distribution. In particular, we impose the following assumption.

Assumption C.1. *Let* $H_w(t) := \frac{\mathbb{E}_{\mathrm{cal}}[w(X)\mathbf{1}\{S \le t\}]}{\mu_w}$, *and let* K_w *denote the CDF of* $H_w(T)$ *for* $T \sim P_{\mathrm{test}}$. *Both* K_w *and its generalized inverse* K_w^{+} *are continuous on* $[0, 1)$.

Assumption C.1 is substantially weaker than Assumption 5.2. Recall that Assumption 5.2 requires the weighted calibration-score distribution to exactly reproduce the test-score distribution,

$$H_w(t) := \frac{\mathbb{E}_{\mathrm{cal}}[w(X)\mathbf{1}\{S \le t\}]}{\mu_w} = F_{\mathrm{test}}(t), \qquad \forall t \in \mathbb{R}.$$

Discussion of Assumption C.1. Thus, it imposes a global transport condition on the weight function w: after reweighting, the entire calibration-score distribution must coincide with the test-score distribution. In contrast, Assumption C.1 does not require H_w to approximate, or even be close to, F_{test}. It only imposes mild regularity on the one-dimensional transformed random variable $H_w(T)$, namely that its distribution and generalized inverse contain no discontinuities. Consequently, the weight function may substantially misspecify the calibration–test shift while Assumption C.1 continues to hold.

This condition is satisfied in many standard continuous-score settings. For example, suppose that the test score T has a continuous distribution and that H_w is continuous and strictly increasing on the support of T. Then the transformation $H_w(T)$ also has a continuous distribution with no gaps in its support, so that both its CDF and generalized inverse are continuous.

Under Assumption C.1, we establish the consistency of Algorithms 3 and 4 for estimating $G_{w,n}$ and $G^{+}_{w,n}$, respectively. The corresponding results are stated in the following two theorems.

Theorem C.2 (Consistency of Algorithm 3)**.** *Suppose Assumptions 2.2, 2.3, and Assumption C.1 hold, and suppose that the h labeled holdout test observations are i.i.d. from P_{test} and independent of the calibration sample. Let $\widehat{G}^{\text{HO}}_{w,n}$ be the output of Algorithm 3. Then, as $n, h \to \infty$,*

$$\sup_{\beta\in(0,1)} \left|\widehat{G}^{\text{HO}}_{w,n}(\beta) - G_{w,n}(\beta)\right| = o_{\mathbb{P}}(1).$$

Consequently, Algorithm 3 is uniformly consistent for $G_{w,n}$.

Theorem C.3 (Consistency of Algorithm 4)**.** *Suppose Assumptions 2.2, 2.3, and Assumption C.1 hold. Suppose further that the h labeled holdout test observations are i.i.d. from P_{test} and independent of the calibration sample. Let $(\widehat{G}^{\text{HO}}_{w,n})^{+}$ be the output of Algorithm 4. Then, as $n, h \to \infty$,*

$$\sup_{u\in[0,1)} \left|(\widehat{G}^{\text{HO}}_{w,n})^{+}(u) - G^{+}_{w,n}(u)\right| = o_{\mathbb{P}}(1).$$

Consequently, Algorithm 4 is uniformly consistent for $G^{+}_{w,n}$.

The proofs for Theorem C.2 and Theorem C.3 are given in Appendix B.8 and Appendix B.9, respectively.

C.3 False Coverage Proportion under General Distribution Shift

Previous work on Conformal Prediction under Distribution Shift has largely focused on the setting of *One-Phase Distribution Shift*, in which the calibration sample is drawn from a single distribution P_{cal} and the test observations are drawn from another distribution P_{test}, subject to Assumption 2.2. While this framework captures many practically important forms of deployment mismatch, it does not exhaust the ways in which Distribution Shift can arise. In particular, more general settings may involve multiple sources or phases of shift, heterogeneous test populations, or changes that cannot be represented by a single calibration-to-test transport relation. In this section, we briefly introduce several such distribution-shift settings and discuss how the preceding FCP analysis can be extended to accommodate them.

Other forms of Distribution Shifts. Beyond the One-Phase Distribution-Shift setting, Distribution Shift may exhibit substantially more complex structures, particularly when observations are collected sequentially over time. We briefly describe several representative forms of such shifts.

Gradual distribution drift. In this setting, the data-generating distribution evolves continuously over time rather than switching directly from one fixed distribution to another. Consequently, the discrepancy between the distributions of two observations typically increases with their temporal separation, and more recent observations tend to be more representative of the current test distribution. Such drift may arise, for example, from slowly changing population characteristics, environmental conditions, or underlying model parameters. Unlike One-Phase Distribution Shift, there is generally no single calibration distribution and test distribution; instead, the observations may be generated from a sequence of gradually varying distributions.

Changepoint shift. In a changepoint setting, the data distribution remains approximately stable within certain time intervals but changes abruptly at one or more unknown changepoints. Observations within the same segment may therefore follow a common or similar distribution, while observations lying on opposite sides of a changepoint can have substantially different distributions. In particular, calibration data collected before a recent changepoint may no longer be representative of the current test distribution, whereas observations collected after the changepoint may remain informative. Multiple changepoints can further produce a sequence of distinct distributional regimes over time.

Time-series dependence. In sequential data, distributional nonstationarity may also coexist with temporal dependence among observations. For example, the distribution of a current observation may depend on previous observations or on an evolving latent state, so that the resulting data are neither identically distributed nor independent across time. This differs from the standard distribution-shift setting, where calibration and test observations are typically assumed to be independent draws from their respective distributions. Such temporal dependence introduces an additional source of

nonexchangeability, since observations may differ not only in their marginal distributions but also through their dependence structure.

Overall, changepoint shift can be viewed as a piecewise-constant counterpart of gradual distribution drift, while One-Phase Distribution Shift corresponds to the special case with a single changepoint separating the calibration and test distributions.

Extension of FCP analysis to other Distribution Shifts. Extending the FCP analysis developed in Sections 4 and 5 to the more general distribution-shift settings described above is both theoretically interesting and practically important. Such an extension would allow finite-sample FCP guarantees to accommodate richer forms of nonstationarity beyond One-Phase Distribution Shift, including gradual drift, multiple changepoints, and temporally dependent observations. A complete treatment of these settings, however, would require additional assumptions and technical developments that are beyond the scope of the present work. We therefore restrict attention here to a brief discussion of several tools and analytical strategies that appear promising for extending our framework, without attempting to establish fully rigorous guarantees for these more general settings. We present these extensions below.

Extension to gradual distribution drift. A natural extension is to replace the single calibration and test distributions by a sequence of time-varying distributions, while allowing the calibration weights to depend on temporal distance so that more recent observations receive greater importance. The population quantities $F_{w,n}$ and $G_{w,n}$ in our current analysis would then become time-dependent, for example $G_{w,n,j}$ for the j-th test observation. Under suitable regularity conditions controlling how rapidly the distribution may drift, one could seek to bound the FCP in terms of an average of these test-specific quantities, together with corresponding calibration and test-sample concentration errors.

Extension to changepoint shift. The piecewise-stationary structure of changepoint shift suggests a more direct extension. If the most recent changepoint is known, observations from the current regime can be isolated and our One-Phase Distribution-Shift analysis may be applied locally to that segment. When the changepoints are unknown, this suggests combining changepoint detection with the present FCP framework, with additional error terms accounting for uncertainty in the estimated changepoint locations. More generally, multiple changepoints could be handled by applying the analysis separately within different approximately stationary regimes.

Extension to time-series dependence. The main additional difficulty under temporal dependence is that the concentration arguments used in our current analysis rely on independence of the calibration observations and, conditionally on the calibration sample, of the test miscoverage indicators. Extending the theory to dependent sequences would therefore require replacing these arguments with concentration tools adapted to dependent data, such as martingale- or mixing-based inequalities. With suitable assumptions controlling the strength of temporal dependence, these tools could play the same role as the independent-sample concentration bounds in our present analysis and yield analogous FCP guarantees with dependence-adjusted error terms.

C.4 Optimal Choice of Confidence Parameters under Covariate Shift

Our main theoretical results in Section 4 provide FCP guarantees at both a fixed miscoverage level and uniformly over α, for a prescribed overall failure probability δ. However, the bounds in Theorems 4.1 and 4.2 are stated in terms of separate confidence parameters arising from different concentration steps: the fixed-α guarantee holds with failure probability $\delta_{\text{cal}} + \delta_{\text{test}}$, whereas the uniform guarantee holds with failure probability $\delta_+ + \delta_- + \delta_{\text{test}}$. In practice, the user typically specifies only the overall level δ, so that

$$\delta = \delta_{\text{cal}} + \delta_{\text{test}} \qquad \text{or} \qquad \delta = \delta_+ + \delta_- + \delta_{\text{test}},$$

while the allocation among the individual components remains to be chosen.

Under general OPDS, optimizing this allocation is difficult because the resulting bounds depend on the generally unknown function $G_{w,n}$. Under Covariate Shift, however, Lemma 5.1 gives $G_{w,n}(\eta) \leq \eta$, reducing the fixed and uniform bounds to

$$\alpha + \Delta_{\text{fix}} + \varepsilon_{\text{fix}} \qquad \text{and} \qquad \alpha + \Delta_{\text{unif}} + \varepsilon_{\text{unif}},$$

respectively. The confidence parameters can therefore be chosen explicitly to minimize these upper bounds subject to the corresponding constraint on the total failure probability δ. We derive such choices below.

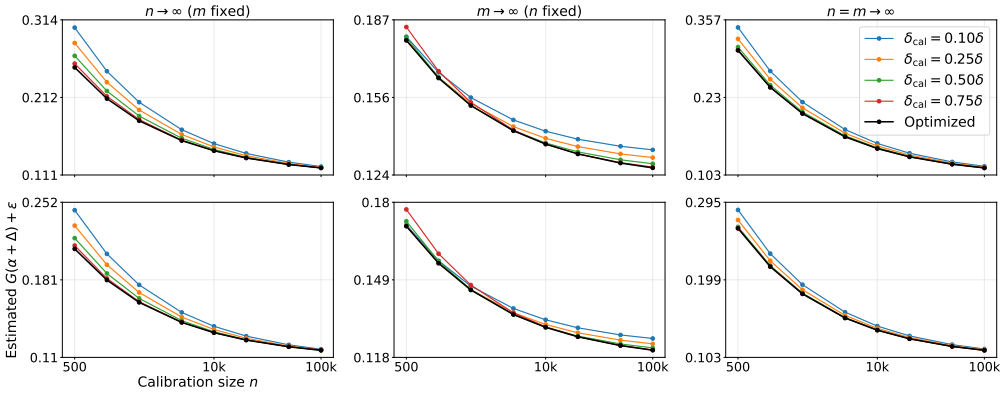

Figure 11: Empirical comparison of different allocations of the overall failure probability δ under Covariate Shift for Theorem 4.1. Results are shown for the Adult and YearPredictionMSD datasets across the three asymptotic regimes: fixed n with $m \to \infty$, fixed m with $n \to \infty$, and $m = n \to \infty$, for the fixed-α FCP guarantees. The black curve corresponds to the allocation proposed in Remark C.4.

Remark C.4 (Choice of $\delta_+, \delta_-, \delta_{\text{cal}}$ and δ_{test} under Covariate Shift)**.** For the fixed-α guarantee in Theorem 4.1, we parameterize the confidence allocation as

$$\delta_{\text{cal}} = a\delta, \qquad \delta_{\text{test}} = (1-a)\delta, \qquad a \in (0,1).$$

An optimal allocation is then obtained by setting $a = a_0$, where

$$a_0 \in \arg\min_{a\in(0,1)} \left\{ \frac{\varepsilon_w}{1+a_n}\sqrt{\frac{-\log(a\delta)}{2n}} + \sqrt{\frac{-\log((1-a)\delta)}{2m}} \right\}.$$

For $\delta < e^{-1/2} \approx 0.606$, the objective above is strictly convex in a, and hence the minimizer a_0 is unique. For the uniform-in-α guarantee in Theorem 4.2, define $t := \delta^{-1}\exp\left\{-2nd_n^2/\varepsilon_w^2\right\}$. Suppose $t < 1/2$. We consider the symmetric allocation

$$\delta_+ = \delta_- = a\delta, \qquad \delta_{\text{test}} = (1-2a)\delta, \qquad a \in (t, 1/2),$$

where the lower constraint $a > t$ ensures that the denominator appearing in Δ_{unif} remains positive. An optimal allocation is obtained by taking $a = a_0$, where

$$a_0 \in \arg\min_{a\in(t,0.5)} \left\{ \frac{\varepsilon_w d_n\left(\sqrt{\frac{\pi}{4}} + \sqrt{\log\frac{1}{a\delta}}\right) + \varepsilon_w\sqrt{\log\frac{1}{a\delta}}}{d_n\left(d_n\sqrt{2n} - \varepsilon_w\sqrt{\log\frac{1}{a\delta}}\right)} + \sqrt{\frac{\log\frac{1}{(1-2a)\delta}}{2m}} \right\}.$$

For $\delta < e^{-1/2} \approx 0.606$, this objective is likewise strictly convex in a, so the optimal allocation is uniquely determined.

Under the confidence allocation specified in Remark C.4, the resulting FCP upper bounds under Covariate Shift are minimized for both Theorem 4.1 and Theorem 4.2. This follows directly by substituting the corresponding choices of $\delta_{\text{cal}}, \delta_{\text{test}}$ and $\delta_+, \delta_-, \delta_{\text{test}}$ into the respective bounds and solving the resulting one-dimensional optimization problems. Since the argument is straightforward, we omit the proof for brevity.

Empirical validation. We next empirically evaluate the effectiveness of the confidence allocation proposed in Remark C.4. We consider the Adult and YearPredictionMSD datasets and, for each dataset, study the same three asymptotic regimes considered in Section 6: (i) fixed n with $m \to \infty$, (ii) fixed m with $n \to \infty$, and (iii) $m = n \to \infty$. For each dataset and regime configuration, we evaluate the FCP upper bounds in Theorem 4.1 and Theorem 4.2 under several admissible allocations of the overall failure probability δ, and compare them with the allocation prescribed by Remark C.4.

Results. The results for the fixed-α and uniform-in-α guarantees are presented in Figures 11 and 12, respectively. Across both datasets and all three asymptotic regimes, the allocation proposed in

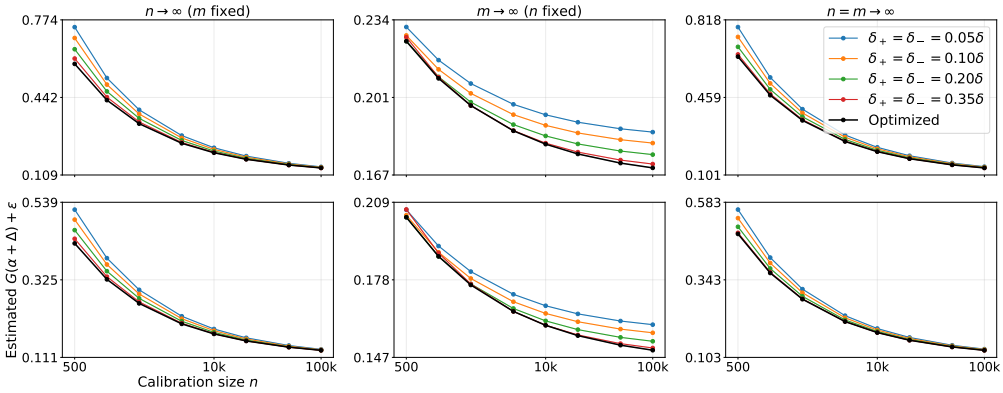

Figure 12: Empirical comparison of different allocations of the overall failure probability δ under Covariate Shift for Theorem 4.2. Results are shown for the Adult and YearPredictionMSD datasets across the three asymptotic regimes: fixed n with $m \to \infty$, fixed m with $n \to \infty$, and $m = n \to \infty$, for the uniform-in-α FCP guarantees. The black curve corresponds to the allocation proposed in Remark C.4.

Remark C.4, shown by the black curve, consistently yields the smallest FCP upper bound among the considered choices. This behavior is observed for both the fixed-α and uniform-in-α guarantees, providing empirical support for the proposed confidence-allocation strategy.

D Experiment Details

D.1 Construction of Covariate and Score-Transport Shifts

In this section, we describe the experimental setups used to generate Covariate Shift and Score-Transport Shift from a given dataset. Suppose we are given a complete dataset $\mathcal{D}$. Our goal is to partition $\mathcal{D}$ into a training set $\mathcal{D}_{\text{train}}$, a calibration set $\mathcal{D}_{\text{cal}}$, and a test set $\mathcal{D}_{\text{test}}$ such that the distributions underlying $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{test}}$ satisfy, respectively, the desired Covariate Shift or Score-Transport Shift structure.

Covariate Shift. We follow the covariate-shift sampling scheme of Tibshirani et al. (2019). We divide the data into disjoint training and auxiliary sets and a remaining source pool, from which calibration and test observations are sampled. After fitting $\widehat{f}$ on the training set, we compute auxiliary nonconformity scores: $s(x, y) = |y - \widehat{f}(x)|$ for regression and

$$s(x, y) = \log \frac{\max_{k \neq y} \widehat{p}(k \mid x)}{\widehat{p}(y \mid x)}$$

for classification. We regress these scores on standardized auxiliary features using Ridge regression and normalize the coefficient vector to obtain a direction a. We define $z(x)$ as the projection of standardized x onto a, centered and scaled using the auxiliary set. Thus, $z(x)$ serves as a one-dimensional, feature-only proxy for expected nonconformity, or predictive difficulty, rather than an observed score that requires the true label. The main experiments use two raw weights,

$$r_{\exp}(x) = e^{0.35 z(x)} \quad \text{and} \quad r_{\text{lin}}(x) = \max\{\varepsilon,\, 1 + 0.35 z(x)\},$$

where $\varepsilon = 10^{-8}$ ensures positivity. For each choice, we clip the raw weights at their 99.5th percentile and normalize them to have mean one over the source pool; we denote the resulting weight by w_0. Calibration observations are sampled uniformly from this pool, whereas test observations are sampled with probabilities proportional to the corresponding weights.

Score-Transport Shift. While Covariate Shift changes the distribution of X but preserves the conditional outcome law, we also consider Score-Transport Shift (STS), which changes the distribution of nonconformity scores within feature-defined groups. We split the feature space into five strata using the 20th, 40th, 60th, and 80th percentiles of $z(X)$ on the auxiliary set. These cutoffs are fixed before

sampling calibration and test observations. Let $g(X) \in [5]$ be the resulting stratum, p_h its proportion in the source pool, and $H_h(t) = P_{\text{cal}}(S \le t \mid g(X) = h)$ its score CDF. For each weighting scheme, the corresponding w_0 sets the test stratum probability q_h to the fraction of total source-pool weight carried by stratum h.

Calibration observations are drawn uniformly from the source pool. For each test observation, we draw h according to q_h and then select an observation uniformly within that stratum. With probability $1-\rho$, we retain its score; with probability ρ, we instead draw a score from the next stratum $\pi(h)$, cycling from stratum 5 back to 1. The resulting conditional test-score CDF is

$$P_{\text{test}}(T \le t \mid g(X) = h) = (1-\rho)H_h(t) + \rho H_{\pi(h)}(t).$$

Accordingly, we weight a calibration observation at x by

$$w_\rho(x) = \frac{(1-\rho)q_{g(x)} + \rho q_{\pi^{-1}(g(x))}}{p_{g(x)}}.$$

This choice matches the weighted calibration-score distribution to the simulated test-score distribution. When $\rho = 0$, the construction is a stratum-coarsened Covariate Shift; for $\rho > 0$, it also changes the conditional score distribution.

We next evaluate the proposed bounds under a general one-phase Distribution Shift that satisfies Assumption 2.2 but violates the Covariate Shift condition in Assumption 2.4. The construction is designed so that the weight function remains interpretable and adequately transports the calibration-score distribution to the test-score distribution.

Stratification. Let

$$C = g(X) \in [K] := \{1, \dots, K\}$$

denote an X-dependent stratum. The mapping $g : \mathcal{X} \to [K]$ is constructed using only the training or an independent auxiliary dataset and is subsequently held fixed. For example, the strata may correspond to quantiles of model uncertainty, quantiles of a selected covariate, clusters in feature space, or distance to the training data.

For each stratum $h \in [K]$, define the conditional calibration-score CDF

$$H_h(t) := P_{\text{cal}}\big(S \le t \mid C = h\big), \qquad S = s(X, Y).$$

We further denote the calibration and test stratum probabilities by

$$p_h := P_{\text{cal}}(C = h), \qquad q_h := P_{\text{test}}(C = h),$$

and assume that $p_h > 0$ and $q_h > 0$ for every $h \in [K]$.

Conditional shift. Let

$$\pi : [K] \to [K]$$

be a fixed nonidentity permutation, and let $\rho \in [0, 1]$ control the strength of the conditional score shift. We construct the synthetic test-score distribution such that

$$P_{\text{test}}\big(T \le t \mid C = g\big) = (1-\rho)H_g(t) + \rho H_{\pi(g)}(t). \tag{44}$$

Thus, a test score associated with stratum g follows the calibration-score distribution of stratum g with probability $1-\rho$, and that of stratum $\pi(g)$ with probability ρ.

When $\rho = 0$, the conditional score distribution is unchanged and the construction reduces to Covariate Shift. In contrast, when $\rho > 0$ and $H_{\pi(g)} \ne H_g$ for at least one stratum g, the conditional score distribution changes, and the construction no longer guarantees that

$$P_{\text{test}}(Y \mid X) = P_{\text{cal}}(Y \mid X).$$

Thus, Assumption 2.4 is not ensured under STS. We emphasize that STS is constructed directly at the score level and should therefore be interpreted as a synthetic stress test of the score- and p-value-based FCP machinery, rather than as an explicit generative model for the joint test distribution of (X, Y). The calibration scores and synthetic test scores are sampled i.i.d. from their respective fixed distributions, and the two samples are independent. This construction allows us to evaluate the proposed guarantees under controlled departures from Covariate Shift while retaining the score-level probabilistic structure underlying our analysis.

Choice of weight function. We define the general-shift weight by

$$w_\rho(x) := \frac{\widetilde{q}^{(\rho)}_{g(x)}}{p_{g(x)}} = \frac{(1-\rho)q_{g(x)} + \rho q_{\pi^{-1}(g(x))}}{p_{g(x)}}. \tag{45}$$

This weight accounts for both the change in the marginal frequency of the strata and the reassignment of conditional residual distributions between strata.

The mean calibration weight is

$$\begin{aligned}\mu_{w_\rho} &= \mathbb{E}_{\text{cal}}\left[w_\rho(X)\right] \\ &= \sum_{h=1}^{K} p_h \frac{\widetilde{q}^{(\rho)}_h}{p_h} = \sum_{h=1}^{K} \widetilde{q}^{(\rho)}_h = 1.\end{aligned} \tag{46}$$

Moreover, the corresponding weighted calibration-score distribution satisfies

$$\begin{aligned}A_{w_\rho}(t) &= \mathbb{E}_{\text{cal}}\left[w_\rho(X)\mathbf{1}\{S \le t\}\right] \\ &= \sum_{h=1}^{K} p_h \frac{\widetilde{q}^{(\rho)}_h}{p_h} H_h(t) \\ &= \sum_{h=1}^{K} \widetilde{q}^{(\rho)}_h H_h(t) = F_{\text{test}}(t).\end{aligned} \tag{47}$$

Consequently,

$$\frac{A_{w_\rho}(t)}{\mu_{w_\rho}} = F_{\text{test}}(t), \qquad \text{for every } t \in \mathbb{R}. \tag{48}$$

Hence, w_ρ exactly transports the calibration-score distribution to the test-score distribution.

Since K is finite and $p_h > 0$ for all $h \in [K]$, the weight is uniformly bounded:

$$0 < w_\rho(x) \le \epsilon_{w_\rho} := \max_{h\in[K]} \frac{\widetilde{q}^{(\rho)}_h}{p_h} < \infty.$$

Therefore, w_ρ satisfies the boundedness requirement in Assumption 2.3. Lastly, for all test observations, we use a fixed test-point weight w_∞ depending on the specific experiment.

D.2 More on Limitations of Existing FCP Bounds under Distribution Shift

In this section, we describe in detail the experimental setup used in Section 3 to study the limitations of existing FCP bounds under Distribution Shift. We also conduct additional experiments to further illustrate these limitations across different datasets and shift settings.

Setup. We compare DKW Gazin et al. (2024) and CoJER Blain et al. (2025) on YearPredictionMSD and Diamonds for regression, and Adult and Fashion-MNIST for classification. We use histogram-based gradient-boosted trees as the base predictor. After fitting the predictor on the training set, we hold out an auxiliary sample that is independent of the calibration and test source pool. We regress the auxiliary nonconformity scores on the corresponding standardized features using Ridge regression and normalize the resulting coefficient vector to obtain the projection direction. The standardized projection onto this direction defines $z(x)$.

We consider the exponential and linear raw weights

$$r_{\exp}(x) = e^{0.35z(x)} \quad \text{and} \quad r_{\text{lin}}(x) = \max\{\varepsilon,\, 1 + 0.35z(x)\},$$

where $\varepsilon = 10^{-8}$. Each raw weight is clipped at its 99.5th percentile and normalized to have mean one over the source pool. Calibration observations are sampled uniformly from this pool, whereas test observations are sampled with probabilities proportional to the corresponding weight.

We set $n = 10{,}000$ and $m = 500$ and independently sample $k = 100$ calibration-test pairs for each dataset and weight. We consider

$$\delta \in \{0.01, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2\}$$

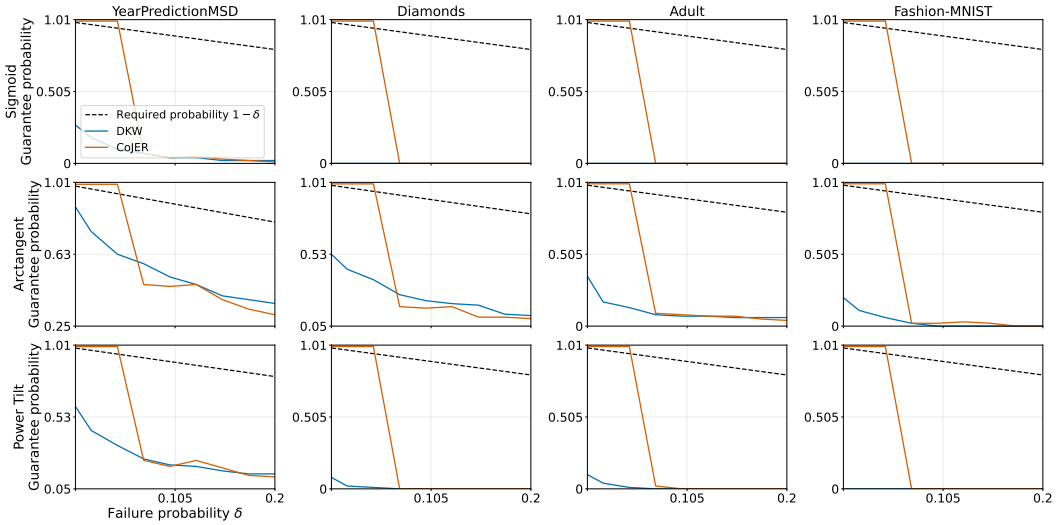

Figure 13: Forward-pass comparison of DKW and CoJER under Covariate Shift for the sigmoid, arctangent, and power-tilt weights (rows) on YearPredictionMSD, Diamonds, Adult, and Fashion-MNIST (columns). The solid curves show the empirical uniform pass rates over the complete miscoverage grid, while the dashed black curve denotes the required probability $1 - \delta$. DKW and CoJER are evaluated using ordinary unweighted conformal p-values.

and evaluate each bound over a grid of 201 miscoverage levels $\alpha \in [0, 1]$. For each repetition, we compute empirical FCP using ordinary *unweighted* conformal p-values. For a method $\mathcal{M} \in \{\mathrm{DKW}, \mathrm{CoJER}\}$, we report

$$\mathfrak{R}_{\mathcal{M}}(\delta) = \frac{1}{k} \sum_{r=1}^{k} \mathbf{1}\left\{\mathrm{FCP}_r(\alpha) \leq B^{\delta}_{\mathcal{M}}(\alpha),\ \forall \alpha\right\},$$

and compare this empirical pass rate with the required probability $1 - \delta$. CoJER is calibrated using 1,000 template simulations and 2,000 calibration simulations.

Extra Experiments. We repeat the baseline comparison on the original four datasets: Adult, Fashion-MNIST, YearPredictionMSD, and Diamonds. In addition to the exponential and linear weights, we consider three additional weight functions: sigmoid, arctangent, and power-tilt, respectively defined as

$$r_{\mathrm{sig}}(x) = \varepsilon + \frac{2}{1 + \exp(-z(x))},$$
$$r_{\mathrm{atan}}(x) = 1 + 2\left(\frac{1}{2} + \frac{\arctan z(x)}{\pi}\right),$$
$$r_{\mathrm{pow}}(x) = (1 + |z(x)|)^{0.5\ \mathrm{sign}(z(x))}.$$

We use the same predictor and experimental settings, and compare the forward DKW Gazin et al. (2024) and CoJER Blain et al. (2025) procedures across 201 target FCP levels.

Result. The results are presented in Figure 13. Across the three additional weight functions, the empirical pass rates of DKW remain substantially below the required probability $1 - \delta$. CoJER attains or approaches the required level only at the smallest values of δ, after which its pass rate decreases sharply. This deterioration is particularly pronounced on Adult and Fashion-MNIST, where both methods have pass rates close to zero over most of the displayed range. The same general behavior is observed on the regression datasets, although the arctangent shift on YearPredictionMSD produces comparatively higher pass rates. Overall, these experiments corroborate and extend the findings of Section 3: under the considered covariate shifts, neither DKW nor CoJER reliably controls the ordinary empirical FCP at the prescribed confidence level.

D.3 Empirical Evaluation of Lemma 5.1

While Theorems 4.1 and 4.2 provide the desired FCP bounds for the fixed-α and uniform-over-α settings, both depend on the CDF $G_{w,n}$, which is generally intractable in practice. Under general One-Phase Distribution Shift, we therefore rely on Algorithm 1 to estimate this quantity. Under Covariate Shift, however, Lemma 5.1 provides the simpler bound $G_{w_0,n}(\alpha) \leq \alpha$, yielding an explicit bound that is linear in α. This simplification also enables the confidence budget δ to be allocated explicitly, as discussed in Appendix C.4. Given the central role of Lemma 5.1, in this section we empirically assess the tightness and practical effectiveness of its bound.

Setup. We conduct experiments on Adult and YearPredictionMSD at $\alpha \in \{0.05, 0.1, 0.2\}$. We consider three weights:

$$w_{\exp}(x) \propto e^{0.35z(x)}, \qquad w_{\text{lin}}(x) \propto \max\{\varepsilon,\, 1 + 0.35z(x)\}, \qquad w_{\text{sig}}(x) \propto \frac{2}{1 + e^{-z(x)}},$$

where $z(x)$ is a standardized score-informed Ridge projection and $\varepsilon > 0$ ensures positive weights. The weights are clipped at the 99.5th percentile and normalized to mean one. We use histogram-based gradient-boosted trees as the base predictor and vary the calibration size from 250 to 50,000. At each size, we apply Algorithm 1 to estimate $G_{w,n}(\alpha)$ and compare it with α.

Result. The results are shown in Figure 6. Across all datasets, miscoverage levels, and weight functions, the estimated $G_{w,n}(\alpha)$ is generally at or below α, in agreement with Lemma 5.1, which guarantees that the underlying quantity satisfies $G_{w,n}(\alpha) \leq \alpha$. Because $G_{w,n}(\alpha)$ is estimated from finite samples, the estimates exhibit sampling variability, so their uncertainty bands may occasionally extend above α. As the calibration size increases, however, the estimated values move progressively closer to α while the corresponding uncertainty bands shrink, indicating both convergence of the estimator and decreasing estimation variability. Overall, the estimates concentrate increasingly tightly around α, demonstrating the empirical tightness of the bound. These results support the practical use of Lemma 5.1 as a simple surrogate for the intractable quantity $G_{w,n}(\alpha)$ under Covariate Shift.

D.4 Ablation on Underlying Models $\hat{f}$

In the main experiments in Section 6, we use histogram-based gradient-boosted trees as the base predictor for both regression and classification. Because the resulting nonconformity scores depend on the fitted predictor, it is important to examine whether the observed FCP behavior depends on this choice. We therefore study the sensitivity of our method to the underlying predictive model.

Setup. We examine the effect of the underlying predictor under Covariate Shift on Adult. We compare histogram-based gradient boosting, random forests, and logistic regression under the exponential, linear, and sigmoid weighting schemes. For each weight, we construct the induced test distribution using histogram-based gradient boosting as the reference predictor and keep this distribution fixed across all predictors. Within each repetition, the same calibration and test observations are therefore used for every predictor, ensuring that differences in the resulting pass rates can be attributed to the choice of the underlying predictive model.

We set $n = 10{,}000$, $m = 500$, and $\delta = 0.1$, and repeat each experiment 100 times. We evaluate the direct and inverse guarantees over grids of 201 values of α and β, respectively, and report their empirical fixed and uniform pass rates relative to the required probability $1 - \delta$.

Result. The results are shown in Figure 7. On Adult, the direct procedures achieve empirical pass rates $\mathfrak{R}_{\text{fix}}(\alpha)$ and $\mathfrak{R}_{\text{unif}}$ above the target confidence level $1 - \delta$ across the displayed miscoverage levels and predictive models. Likewise, the inverse procedures achieve empirical pass rates $\mathfrak{R}^{-1}_{\text{fix}}(\beta)$ and $\mathfrak{R}^{-1}_{\text{unif}}$ above $1 - \delta$ across the considered target FCP levels and predictive models. As shown in Figure 7, this behavior is consistent across the different predictor classes considered, supporting the empirical validity of both the direct and inverse FCP guarantees beyond a single underlying predictive model.

D.5 Ablation on Weight Function

As shown in Section D.1, the choice of weight function is a key component in constructing data under Covariate Shift and Score-Transport Shift. In particular, the weight function determines how

the calibration distribution is reweighted toward the test distribution and therefore directly controls the nature and severity of the induced shift. Different choices of weights can substantially alter the relative likelihood of samples across the covariate space, leading to markedly different calibration-test discrepancies and, consequently, different FCP behavior. It is therefore important that the proposed method remain reliable across a broad range of weighting schemes. In this section, we empirically evaluate the robustness of our method with respect to the choice of weight function.

Setup. We use YearPredictionMSD with a histogram-based gradient-boosting predictor and the score-informed projection $z(x)$ defined in Section D.1. The projection direction is fitted using nonconformity scores from an auxiliary sample that is independent of the calibration and test source pool. We consider five raw weight functions:

$$
\begin{aligned}
r_{\exp}(x) &= \exp\{0.35z(x)\},\\
r_{\mathrm{lin}}(x) &= \max\{\varepsilon,\, 1 + 0.35z(x)\},\\
r_{\mathrm{sig}}(x) &= \varepsilon + \frac{2}{1+\exp\{-z(x)\}},\\
r_{\mathrm{atan}}(x) &= 1 + 2\left(\frac{1}{2} + \frac{\arctan(z(x))}{\pi}\right),\\
r_{\mathrm{pow}}(x) &= (1+|z(x)|)^{0.5\ \mathrm{sgn}(z(x))},
\end{aligned}
$$

where $\varepsilon = 10^{-8}$. Each raw weight is clipped at its 99.5th percentile and normalized to have mean one over the source pool.

For each weighting scheme, we conduct experiments under both Covariate Shift and Score-Transport Shift. For STS, we partition the source pool into five strata according to $z(x)$ and set the transport probability to $\rho = 0.5$. Each weight induces its own test distribution. We set $n = 10{,}000$, $m = 500$, and $\delta = 0.1$, and independently sample 100 calibration-test pairs for each weight and shift setting. We evaluate the direct and inverse guarantees over grids of 201 values of α and β, respectively, and report the corresponding empirical fixed and uniform pass rates against the required confidence level $1 - \delta$.

Result. The results are shown in Figure 8. Across both Covariate Shift and STS, the direct procedures attain empirical pass rates $\mathfrak{R}_{\mathrm{fix}}(\alpha)$ and $\mathfrak{R}_{\mathrm{unif}}$ above the target confidence level $1 - \delta$ for all five weight functions. Similarly, the inverse procedures attain pass rates $\mathfrak{R}^{-1}_{\mathrm{fix}}(\beta)$ and $\mathfrak{R}^{-1}_{\mathrm{unif}}$ above $1 - \delta$ across the same settings. These results empirically support the validity of both the direct and inverse FCP guarantees under the considered distribution-shift scenarios.

D.6 Ablation on Prediction Efficiency

In this work, we have primarily focused on the validity of the proposed FCP guarantees. However, another important consideration in Conformal Prediction is the *efficiency* of the resulting prediction sets, measured by their average size over the test samples. Coverage guarantees alone are not sufficient to characterize the practical quality of a conformal method, since overly large prediction sets can achieve high coverage while providing little useful information. More efficient prediction sets are therefore those that remain valid while being as small and informative as possible. Therefore, in this section, we empirically analyze the relationship between FCP control and prediction-set efficiency.

Setup. We use the six datasets in Figure 9 with a histogram-based gradient-boosting predictor and exponential, linear and sigmoid weights. For each of 201 target bounds $\beta \in [0, 1]$, Algorithm 3 selects a conformal level $\alpha(\beta)$. We then evaluate the resulting prediction sets under Covariate Shift using 10,000 calibration and 500 test observations, averaged over 100 repetitions. Efficiency is measured by interval length for regression and set cardinality for classification.

Result. The result is given in Figure 9. Prediction sets become smaller as the permitted FCP bound β increases, illustrating the trade-off between tighter FCP control and efficiency. The curves differ across weight functions, particularly at stringent bounds, but exhibit the same overall trend.

D.7 Runtime Analysis

While Sections 4 and 5 develop theoretical guarantees for both the direct and inverse FCP problems, their practical implementation requires evaluating the CDF $G_{w,n}$ and its generalized inverse, which are generally intractable. We therefore rely on Algorithms 1 and 2 to estimate these quantities, with computational complexity $O(n \log n)$, where n denotes the calibration sample size. Although this cost is modest since the calibration set is typically much smaller than the training set, FCP control is performed at inference time on a batch of test samples, where computational efficiency remains important. In this section, we therefore empirically quantify the runtime of the proposed FCP bounds and their associated estimation procedures.

Setup. We measure runtime on Adult, Diamonds, and YearPredictionMSD using histogram-based gradient boosting, random forests, and a linear model (logistic regression for classification and ridge regression for regression). We use the exponential weight and vary the calibration size over $n \in \{250, 500, 1000, 2000, 5000, 10000, 20000, 50000\}$, with $m = 500$ and $\delta = 0.1$. For each dataset-model pair, we time the direct (fixed- and uniform-α) and inverse (fixed- and uniform-β) procedures, including calibration-score computation, construction of the weighted calibration structure, and evaluation of the corresponding bounds. Model training and weight fitting are excluded. Each setting has one warm-up run followed by 20 timed repetitions; we report the median and the 10th-90th percentile range.

Result. The results are shown in Figure 10. The runtime grows approximately linearly with the calibration size n, consistent with the computational complexity of the proposed algorithms. Moreover, across all datasets and underlying models, the runtime remains below 0.6 seconds for calibration sizes up to 50,000. This indicates that applying Algorithms 1 and 2 introduces only a modest computational overhead, supporting their practical use at inference time.

D.8 Computation Infrastructure

All experiments were run on a CPU-only machine with an 8-core processor and 8 GB of RAM. No GPU acceleration was used.