SPLIT CONFORMAL INTERVAL LENGTH UNDER HUBER CONTAMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing analyses of interval length in split conformal prediction largely focus on clean data. However, how contamination at the training and calibration stages alters interval length relative to the corresponding clean population oracle remains understudied. In this paper, we consider Huber contamination and derive a finite-sample upper bound that holds uniformly over contaminating distributions, with leading terms scaling as $L(\delta_1 + n_1^{-1/2}) + (n_2^{-1/2} + \delta_2)/((1-\delta_2)r_0)$, in addition to a clean-sample estimation term. We further establish a local minimax lower bound that matches an attainable fixed-confidence upper rate, up to constant factors, on a regular asymmetric location submodel. For calibration contamination, we obtain an exact worst-case characterization of the population interval-length deviation. We also extend the upper-bound results to conformalized quantile regression and adaptive residual scores. Experiments on synthetic and real-world datasets illustrate the distinct effects of training and calibration contamination and are consistent with our theoretical analysis.

1 INTRODUCTION

Modern regression methods often provide accurate point predictions. However, point predictions alone do not quantify predictive uncertainty, which can be important for downstream decision making. For example, uncertainty in wind and solar generation is relevant to reserve scheduling, balancing, and dispatch in power systems (Pinson, 2013; Hong et al., 2016). *Conformal prediction* (CP) provides a general framework for constructing prediction sets with distribution-free, finite-sample marginal coverage guarantees (Vovk et al., 2005). Among its variants, split conformal prediction (SCP) fits a predictive model on a training split and calibrates the prediction set using held-out data (Papadopoulos et al., 2002). Given a target miscoverage level $\alpha \in (0, 1)$, split conformal prediction constructs a prediction set C_α such that, for a new test observation $(\mathbf{X}_{\text{new}}, Y_{\text{new}})$,

$$\mathbb{P}(Y_{\text{new}} \in C_\alpha(\mathbf{X}_{\text{new}})) \geq 1 - \alpha,$$

provided that the calibration observations and the test observation are exchangeable.

In regression, the informativeness of a conformal predictor is commonly measured by prediction-interval length, with shorter intervals preferred at the target coverage level. Existing analyses of conformal efficiency, however, largely focus on clean data (Lei et al., 2018; Sesia & Candes, 2020), while a separate line of work has studied the robustness of conformal prediction under various contamination and nonexchangeability models (Barber et al., 2023; Li et al., 2024; Zargarbashi et al., 2024; Yang et al., 2026; Clarkson et al., 2024).

The effect of contamination on prediction-interval length remains less understood. We characterize its deviation from the clean population oracle in terms of the training and calibration contamination levels, uniformly over contaminating distributions. In split conformal prediction, the two stages affect interval length through different mechanisms: training contamination perturbs the fitted predictor, while calibration contamination shifts the score quantile.

To address this question, we study split conformal regression under Huber contamination in both the training and calibration samples, with contamination levels δ_1, δ_2 and sample sizes n_1, n_2, while allowing the contaminating distributions to be arbitrary. Our main contributions are as follows:

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.

- **Uniform upper bounds under joint contamination.** For absolute-residual scores and $\delta_2 < \alpha$, under standard stability and regularity conditions, we show that the prediction-interval length deviates from its clean oracle at the rate

$$O_P\left(\eta_{n_1}(P_0) + L\big(\delta_1 + n_1^{-1/2}\big) + \frac{n_2^{-1/2} + \delta_2}{(1-\delta_2)r_0}\right),$$

  uniformly over contaminating distributions. Here $\eta_{n_1}(P_0)$ is the clean sample estimation error, while L and r_0^{-1} capture the training- and calibration-side sensitivities, respectively. We further extend the interval-length analysis to conformalized quantile regression and adaptive residual scores, and derive a corresponding bound on set-level discrepancy. For calibration contamination, we additionally obtain an exact worst-case characterization of interval-length inflation and contraction over contaminating distributions.

- **A minimax lower bound on a location submodel.** On a regular asymmetric location submodel, we establish a minimax lower bound of order

$$L_\star\big(\delta_1 + n_1^{-1/2}\big) + \frac{n_2^{-1/2} + \delta_2}{r_\star}.$$

  Under suitable local comparability conditions on the same submodel, this lower bound matches the upper rate up to constant factors for absolute-residual split conformal prediction.

- **Empirical characterization.** Across synthetic and real-world regression datasets, we empirically characterize the distinct effects of training and calibration contamination and compare their behavior across absolute-residual, adaptive, and CQR scores. The results illustrate the stage-specific mechanisms predicted by our theory and their different sensitivity to contamination.

2 Related Work

Conformal prediction and efficiency analysis. Conformal prediction constructs prediction sets with finite-sample marginal coverage guarantees for exchangeable observations (Papadopoulos et al., 2002; Vovk et al., 2005; 2009). The efficiency of a conformal predictor characterizes the informativeness of its prediction sets and can be evaluated using different criteria (Vovk et al., 2016). In regression, efficiency is commonly measured by prediction-set length or volume, often relative to a population-level oracle benchmark. Prior work has studied such efficiency guarantees for nonparametric prediction bands (Lei & Wasserman, 2014), conformalized quantile regression and distributional conformal prediction (Sesia & Candes, 2020; Chernozhukov et al., 2021), and conditional-density-based prediction regions in high dimensions (Izbicki et al., 2022). Other work derives finite-sample bounds on prediction-set size (Dhillon et al., 2024; Zecchin et al., 2024; Yao et al., 2026) or directly optimizes set length or volume (Kiyani et al., 2024; Braun et al., 2025; Le Bars & Humbert, 2025; Gao et al., 2025). Efficiency improvements have also been developed for data-reuse procedures such as cross-conformal prediction through more effective aggregation across folds (Gasparin & Ramdas, 2025). Our analysis complements this literature by studying efficiency under training and calibration contamination. When $\delta_1 = \delta_2 = 0$, our decomposition reduces to the standard clean-data setting of Lei et al. (2018), in which only finite-sample training and calibration errors remain. Under contamination, additional population-level shifts in the predictor and score quantile may arise and need not vanish asymptotically.

Conformal prediction under contamination. Robustness of conformal prediction has been studied under departures from exchangeability and various forms of data contamination. Barber et al. (2023) analyze coverage under general nonexchangeability, including Huber contamination (Huber, 1964), and derive a multiplicative clean-coverage correction of $\alpha' = \alpha(1-\delta_2)$ under δ_2-Huber contamination. In classification, prior work considers training poisoning, test-time perturbations, calibration poisoning, and random label noise (Li et al., 2024; Zargarbashi et al., 2024; Einbinder et al., 2024). Yang et al. (2026) propose EnCP, a partition-and-vote defense against bounded poisoning of training or calibration samples. Their method modifies the conformal procedure through an ensemble construction, whereas we study unmodified split conformal regression under distributional contamination. The work most closely related to our calibration-contamination setting is Clarkson

et al. (2024), who bound the absolute difference between the expected clean and contaminated calibration order statistics by $\delta_2 C(n_2, k_{n_2}) W_1(\Pi_{P_0}, \Pi_{Q_2})$, where W_1 denotes the 1-Wasserstein distance between the clean and contaminating score distributions. This bound may be less informative in the worst-case setting: $W_1(\Pi_{P_0}, \Pi_{Q_2})$ may be unbounded, whereas our high-probability bound is uniform over Q_2; moreover, for $k_{n_2} = \lceil (n_2+1)(1-\alpha) \rceil$, $C(n_2, k_{n_2}) = \Theta(\sqrt{n_2})$. Our bound instead separates a vanishing $O(n_2^{-1/2})$ sampling term from an $O(\delta_2/r_0)$ population-contamination term and is uniform over Q_2.

3 Preliminaries and Setups

3.1 Preliminaries on Split Conformal Prediction

We consider regression with feature space $\mathcal{X} \subseteq \mathbb{R}^d$ and response space $\mathcal{Y} = \mathbb{R}$, and let P_0 denote the clean distribution on $\mathcal{X} \times \mathcal{Y}$. Fix a target miscoverage level $\alpha \in (0,1)$. Let A denote a training source, either a population distribution or a realized training sample, and let $\mu(\cdot; A)$ denote the corresponding population or fitted predictor. For the absolute-residual nonconformity score, define

$$R_A(\mathbf{x}, y) := |y - \mu(\mathbf{x}; A)|, \qquad (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}.$$

For any distribution B on $\mathcal{X} \times \mathcal{Y}$, let

$$F_{A,B}(t) := \mathbb{P}_{(X,Y)\sim B}\big(R_A(X,Y) \le t\big), \qquad q(A; B) := F_{A,B}^{-1}(1-\alpha),$$

where $F_{A,B}^{-1}(u) := \inf\{t \in \mathbb{R} : F_{A,B}(t) \ge u\}$, $u \in (0,1]$. When A is a realized training sample, $F_{A,B}$ and $q(A;B)$ are understood conditionally on A.

Let $\mathcal{D}_1$ and $\mathcal{D}_2$ denote training and calibration samples of sizes n_1 and n_2, respectively, with their sampling distributions specified below. Write $\mathcal{D}_2 = \{(X_i, Y_i)\}_{i=1}^{n_2}$. For a training source A, define the calibration scores

$$R_{A,i} := R_A(X_i, Y_i), \qquad i = 1, \dots, n_2,$$

and let $R_{A,(1)} \le \cdots \le R_{A,(n_2)}$ denote their order statistics. Set $k_{n_2} := \lceil (n_2+1)(1-\alpha) \rceil$ and define the split-conformal calibration quantile by

$$\widehat{q}(A; \mathcal{D}_2) := \begin{cases} R_{A,(k_{n_2})}, & k_{n_2} \le n_2, \\ +\infty, & k_{n_2} = n_2 + 1. \end{cases}$$

The split conformal prediction interval based on $\mathcal{D}_1$ is

$$C_\alpha(x; \mathcal{D}_1, \mathcal{D}_2) := \big[\mu(x; \mathcal{D}_1) - \widehat{q}(\mathcal{D}_1; \mathcal{D}_2),\ \mu(x; \mathcal{D}_1) + \widehat{q}(\mathcal{D}_1; \mathcal{D}_2)\big].$$

If, conditional on $\mathcal{D}_1$, the calibration observations and a new test observation $(X_{\text{new}}, Y_{\text{new}})$ are exchangeable, then

$$\mathbb{P}(Y_{\text{new}} \in C_\alpha(X_{\text{new}}; \mathcal{D}_1, \mathcal{D}_2) \,|\, \mathcal{D}_1) \ge 1 - \alpha.$$

For the absolute-residual score, the interval length is $2\widehat{q}(\mathcal{D}_1; \mathcal{D}_2)$. The corresponding clean population oracle is

$$C_\alpha^0(x) := \big[\mu(x; P_0) - q(P_0; P_0),\ \mu(x; P_0) + q(P_0; P_0)\big].$$

Our analysis compares the fitted interval under contamination with this clean oracle.

To control the sampling error of the fitted clean predictor, we use the following condition.

Assumption 3.1 (Sampling stability (Lei et al., 2018)). *For every fixed training distribution P, there exist deterministic sequences $\eta_n(P) = o(1)$ and $\rho_n(P) = o(1)$ such that, for $\mathcal{D}_n \sim P^n$ and all sufficiently large n,*

$$\mathbb{P}_{\mathcal{D}_n \sim P^n}\left(\|\mu(\cdot; \mathcal{D}_n) - \mu(\cdot; P)\|_\infty \ge \eta_n(P)\right) \le \rho_n(P),$$

where $\|g\|_\infty := \sup_{\mathbf{x} \in \mathcal{X}} |g(\mathbf{x})|$.

3.2 Training and Calibration Contamination

We model contamination in the training and calibration stages separately using Huber's contamination model (Huber, 1964):

$$P_1 := (1-\delta_1)P_0 + \delta_1 Q_1, \qquad P_2 := (1-\delta_2)P_0 + \delta_2 Q_2,$$

where $\delta_1, \delta_2 \in [0,1)$ are the contamination levels and Q_1, Q_2 are arbitrary distributions on $\mathcal{X} \times \mathcal{Y}$. For the training stage, we construct a clean dataset $\mathcal{D}_0 \sim P_0^{n_1}$ and a contaminated dataset $\mathcal{D}_1 \sim P_1^{n_1}$ on a common probability space through Bernoulli replacement. Specifically, each observation in $\mathcal{D}_0$ is independently replaced by a draw from Q_1 with probability δ_1. Hence, $\mathcal{D}_0$ and $\mathcal{D}_1$ differ in at most K observations, where $K \sim \mathrm{Binomial}(n_1, \delta_1)$. The calibration dataset $\mathcal{D}_2 \sim P_2^{n_2}$ is independent of $(\mathcal{D}_0, \mathcal{D}_1)$.

Remark 3.1 (Fixed-fraction replacement)**.** *An alternative is to replace exactly $k = \lfloor n\delta \rfloor$ uniformly selected observations. Unlike the Bernoulli replacement induced by the Huber contamination model, the number of replacements is then fixed rather than random. Our analysis extends to this fixed-fraction scheme: the same contamination-dependent terms are retained, while the finite-sample fluctuation due to the random replacement count is removed. Appendix E gives the corresponding bound and an empirical comparison of the two schemes.*

For any fixed A, the score CDF under P_2 satisfies

$$F_{A,P_2}(t) = (1-\delta_2)F_{A,P_0}(t) + \delta_2 F_{A,Q_2}(t).$$

The identity compares score distributions while holding the fitted predictor fixed. For our interval-length analyses, we assume $\alpha \in (0, 1/2)$ to simplify the two-sided calibration bounds. For the clean oracle $A = P_0$, define

$$q_-(\delta_2) := F^{-1}_{P_0,P_0}\left(\frac{1-\alpha-\delta_2}{1-\delta_2}\right), \qquad q_+(\delta_2) := F^{-1}_{P_0,P_0}\left(\frac{1-\alpha}{1-\delta_2}\right). \tag{1}$$

For the calibration-side analysis, we assume $\delta_2 \in [0, \alpha)$.

Assumption 3.2 (Local score-density regularity)**.** *The clean oracle score CDF F_{P_0,P_0} is absolutely continuous on an open interval $\mathcal{I} \supset [q_-(\delta_2), q_+(\delta_2)]$, with density f_{P_0,P_0} satisfying*

$$\inf_{t\in\mathcal{I}} f_{P_0,P_0}(t) \geq r_0 > 0.$$

4 Robustness Analysis Under Contamination

4.1 Finite-Sample Robustness under Joint Contamination

To isolate the effects of contamination from finite-sample variation at the training and calibration stages, we use the following decomposition of the quantile error:

$$\begin{aligned}|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2) - q(P_0;P_0)| \leq\ & \underbrace{|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2) - \widehat{q}(\mathcal{D}_0;\mathcal{D}_2)|}_{\text{Lemma 4.1: Training contamination error}} + \underbrace{|\widehat{q}(\mathcal{D}_0;\mathcal{D}_2) - \widehat{q}(P_0;\mathcal{D}_2)|}_{\text{Lemma 4.2: Training-stage sampling error}} \\ & + \underbrace{|\widehat{q}(P_0;\mathcal{D}_2) - q(P_0;P_2)|}_{\text{Lemma 4.3: Calibration-stage sampling error}} + \underbrace{|q(P_0;P_2) - q(P_0;P_0)|}_{\text{Lemma 4.4: Calibration contamination error}}.\end{aligned}$$

4.1.1 Training-Side Error

The first term in the preceding decomposition captures how training contamination changes the calibration quantile through the fitted predictor. We control this term uniformly over the contamination distribution Q_1 by bounding the predictor's sensitivity to multiple training-sample replacements. Specifically, motivated by uniform algorithmic stability (Bousquet & Elisseeff, 2002; Liu et al., 2017; Bassily et al., 2020), we use the following multi-replacement version of the one-replacement condition studied by Pournaderi & Xiang (2024).

Definition 4.1 (Uniform replacement stability)**.** *A regression learning algorithm μ is said to be L-uniformly replacement stable if, for every pair of training datasets $\mathcal{D}$ and $\mathcal{D}'$ of size n that differ in at most k observations,*

$$\|\mu(\cdot;\mathcal{D}) - \mu(\cdot;\mathcal{D}')\|_\infty \leq Lk/n,$$

where $L < \infty$ is independent of n, k, $\mathcal{D}$, and $\mathcal{D}'$.

Next, we give an explicit stability constant for ℓ_2-regularized linear quantile regression.

Example 4.1 (ℓ_2-regularized linear quantile regression)**.** *Suppose that* $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_2 \le B_x$. *For* $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, *define the* ℓ_2*-regularized linear quantile regression estimator by*

$$\hat{\omega}(\mathcal{D}) := \operatorname{argmin}_{\omega\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \omega^\top \mathbf{X}_i) + \frac{\lambda}{2}\|\omega\|_2^2 \right\},$$

where $\rho_\tau(u) = u\big(\tau - \mathbf{1}\{u < 0\}\big)$ *is the pinball loss,* $\tau \in (0,1)$, *and* $\lambda > 0$. *Then the predictor* $\mu(\mathbf{x}; \mathcal{D}) := \mathbf{x}^\top \hat{\omega}(\mathcal{D})$ *is* L*-uniformly replacement stable with* $L = 2B_x^2 \max\{\tau, 1-\tau\}/\lambda$.

Uniform replacement stability yields the following bound on the calibration-quantile perturbation:

Lemma 4.1 (Training-contamination quantile perturbation)**.** *If* μ *is* L*-uniformly replacement stable, then, for all sufficiently large* n_2, *the following holds with probability at least* $1 - n_1^{-2}$, *uniformly over* Q_1*:*

$$|\widehat{q}(\mathcal{D}_1; \mathcal{D}_2) - \widehat{q}(\mathcal{D}_0; \mathcal{D}_2)| \le L\left(\delta_1 + \sqrt{\frac{\log n_1}{n_1}}\right).$$

Sampling stability bounds the calibration-quantile discrepancy between the fitted clean predictor and its population counterpart.

Lemma 4.2 (Training-stage sampling quantile error)**.** *Under Assumption 3.1, for all sufficiently large* n_1, n_2, *with probability at least* $1 - \rho_{n_1}(P_0)$,

$$|\widehat{q}(\mathcal{D}_0; \mathcal{D}_2) - \widehat{q}(P_0; \mathcal{D}_2)| \le \eta_{n_1}(P_0).$$

4.1.2 Calibration-Side Error

The calibration-stage error comprises empirical quantile fluctuation under P_2 and the contamination-induced shift of the population quantile.

Lemma 4.3 (Finite-sample calibration quantile error)**.** *Under Assumption 3.2, for every* $\gamma \in (0,1)$ *and all sufficiently large* n_2, *the following holds with probability at least* $1 - \gamma$*:*

$$|\widehat{q}(P_0; \mathcal{D}_2) - q(P_0; P_2)| \le \frac{1}{(1-\delta_2)r_0}\left(\sqrt{\frac{\log(2/\gamma)}{2n_2}} + \frac{2}{n_2}\right).$$

Figure 1 illustrates how upper-tail calibration contamination shifts the empirical score quantile on the `Abalone` dataset. The following lemma bounds the corresponding population-level shift.

Lemma 4.4 (Quantile bias under calibration contamination)**.** *Under Assumption 3.2 and* $\delta_2 < \alpha$, *the population* $(1-\alpha)$*-quantile of the oracle score* R_{P_0} *satisfies*

$$|q(P_0; P_2) - q(P_0; P_0)| \le \frac{\delta_2(1-\alpha)}{(1-\delta_2)r_0}.$$

The factor r_0^{-1} captures local quantile sensitivity: the same probability-level perturbation can produce a larger quantile shift when the clean score density near the target quantile is smaller.

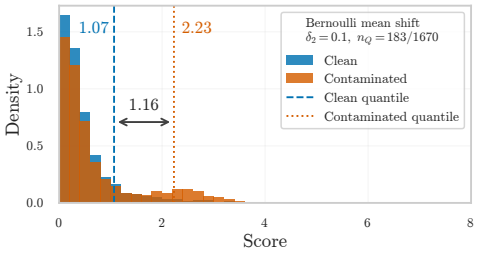


Figure 1: Clean and contaminated absolute-score distributions on `Abalone`.

4.1.3 Joint Interval-Length Deviation Bound

Combining the preceding training- and calibration-stage bounds yields the following finite-sample guarantee under joint contamination.

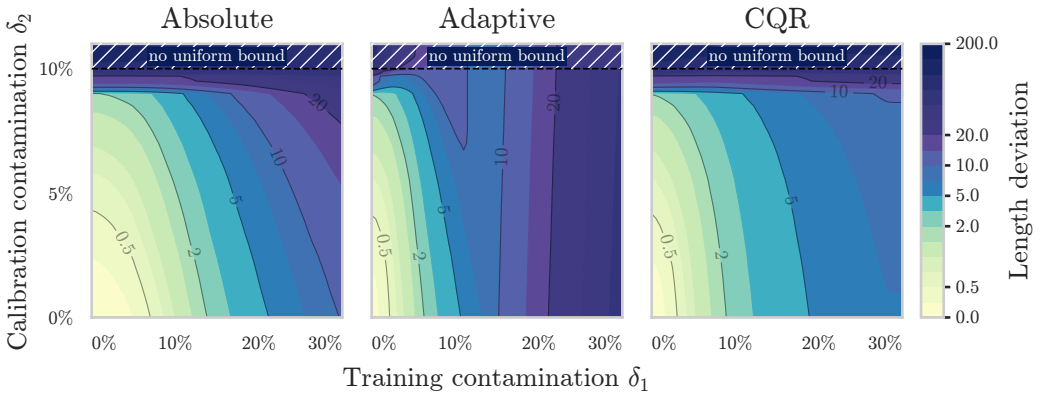


Figure 2: Joint training and calibration contamination for three conformal prediction nonconformity scores: Absolute, Adaptive, and CQR. Both stages use $Q_1 = Q_2 = \text{Dirac}((-1,-1,-1,-1,+1),-80)$.

Theorem 4.1 (Interval-length deviation under joint contamination)**.** *Under Assumptions 3.1 and 3.2, suppose also that μ is L-uniformly replacement stable as in Definition 4.1. Fix $\delta_1 \in [0,1)$ and $\delta_2 \in [0,\alpha)$. For every $\gamma \in (0,1)$ and all sufficiently large n_1, n_2, the following holds with probability at least $1 - \rho_{n_1}(P_0) - n_1^{-2} - \gamma$, uniformly over the contamination distributions Q_1 and Q_2:*

$$2\left|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2) - q(P_0;P_0)\right| \leq 2\eta_{n_1}(P_0) + 2L\left(\delta_1 + \sqrt{\frac{\log n_1}{n_1}}\right) + \frac{2}{(1-\delta_2)r_0}\left(\sqrt{\frac{\log(2/\gamma)}{2n_2}} + \frac{2}{n_2} + \delta_2(1-\alpha)\right).$$

At fixed contamination levels, the sampling terms vanish as $n_1, n_2 \to \infty$, while the contamination-dependent terms remain in the bound.

Remark 4.1 (Extensions to alternative nonconformity scores)**.** *The training–calibration decomposition extends to conformalized quantile regression (CQR) and adaptive residual scores under score-specific stability and clean-score regularity conditions. For CQR, the interval-length deviation additionally accounts for perturbations of the fitted lower and upper quantile regressors. For adaptive scores, a uniform additive score perturbation need not hold under arbitrary calibration contamination, so the training-side comparison is handled multiplicatively under a positive lower bound on the fitted scale. Appendix B.6 gives the corresponding assumptions, finite-sample bounds, and proofs.*

4.1.4 Set-level Discrepancy

Beyond interval length, the discrepancy between prediction intervals also depends on their center displacement. We therefore measure the set-level discrepancy between the fitted interval and the clean oracle interval by the Lebesgue measure of their symmetric difference. Define

$$C_{\text{clean}}(\mathbf{x}) := [\mu(\mathbf{x};P_0) - q(P_0;P_0), \mu(\mathbf{x};P_0) + q(P_0;P_0)], \quad C_{\text{cont}}(\mathbf{x}) := C_\alpha(\mathbf{x};\mathcal{D}_1,\mathcal{D}_2).$$

Let Leb denote the Lebesgue measure on $\mathbb{R}$ and $\triangle$ denote the symmetric difference. Comparing interval endpoints gives

$$\sup_{\mathbf{x}\in\mathcal{X}} \text{Leb}(C_{\text{cont}}(\mathbf{x}) \triangle C_{\text{clean}}(\mathbf{x})) \leq 2\left\|\mu(\cdot;\mathcal{D}_1) - \mu(\cdot;P_0)\right\|_\infty + 2\left|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2) - q(P_0;P_0)\right|.$$

Assumption 3.1 and Proposition B.1 give the following bound, uniformly over Q_1:

$$\mathbb{P}\left\{\left\|\mu(\cdot;\mathcal{D}_1) - \mu(\cdot;P_0)\right\|_\infty > \eta_{n_1}(P_0) + L\left(\delta_1 + \sqrt{\frac{\log n_1}{n_1}}\right)\right\} \leq \rho_{n_1}(P_0) + n_1^{-2}.$$

The center bound uses the same training events as Theorem 4.1. Combining the two bounds therefore yields the following corollary without an additional failure probability.

Corollary 4.1 (Set-level discrepancy)**.** *Under the assumptions of Theorem 4.1, for every $\gamma \in (0,1)$ and all sufficiently large n_1, n_2, with probability at least $1 - \rho_{n_1}(P_0) - n_1^{-2} - \gamma$,*

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathrm{Leb}(C_{\mathrm{cont}}(\mathbf{x}) \triangle C_{\mathrm{clean}}(\mathbf{x})) \leq 4\eta_{n_1}(P_0) + 4L\left(\delta_1 + \sqrt{\frac{\log n_1}{n_1}}\right) + \frac{2}{(1-\delta_2)r_0}\left(\sqrt{\frac{\log(2/\gamma)}{2n_2}} + \frac{2}{n_2} + \delta_2(1-\alpha)\right).$$

4.2 A Joint Minimax Lower Bound

To assess the sharpness of the component rates, we consider the location submodel

$$\mathbf{X} = 0, \qquad Y = \theta + \varepsilon, \qquad \theta \in \Theta := [-h_0/2, h_0/2],$$

where $h_0 > 0$ and the continuously differentiable noise density g satisfy Assumption B.1. Let $q_\star$ be the $(1-\alpha)$-quantile of $|\varepsilon|$, so that the clean oracle interval has length $2q_\star$. The local information scale $L_\star$ controls the distinguishability of nearby location parameters, while $r_\star$ characterizes the score density near the target quantile. The proof combines two-point indistinguishability arguments (LeCam, 1973) for the training terms with order-statistic and coupling arguments for the calibration terms.

Theorem 4.2 (Informal joint minimax lower rate)**.** *For the regular location submodel above, consider standard absolute-residual split conformal with measurable, possibly randomized training predictors $\widehat{\theta}(\mathcal{D}_1) \in \Theta$. For a fixed submodel and all sufficiently large n_1, n_2, the local minimax interval-length deviation from $2q_\star$ is*

$$\Omega\left(L_\star\left(n_1^{-1/2} + \delta_1\right) + \frac{n_2^{-1/2} + \delta_2}{r_\star}\right).$$

This lower bound holds with probability at least some constant $c_0 > 0$.

If $r_\star$ and r_0 are comparable up to fixed constant factors and the additional fixed-confidence training conditions in Appendix B.9 hold, this lower rate matches an attainable upper rate on the same fixed submodel. The comparison establishes local rate sharpness within the absolute-residual split-conformal architecture and the stated algorithm class. A formal version, including the precise assumptions and minimax quantifiers, is given in Theorem B.1.

4.3 Calibration-Specific Analysis

4.3.1 Exact Worst-Case Inflation, Contraction, and Breakdown

The preceding analysis provides an upper bound on the population interval-length deviation induced by calibration contamination. We now sharpen this analysis by characterizing the exact worst case over the contamination distribution Q_2, and by identifying contaminating distributions that attain the maximal inflation and contraction. Define

$$\mathfrak{D}_\alpha(\delta_2) := \sup_{Q_2} 2\left|q(P_0; P_2) - q(P_0; P_0)\right|.$$

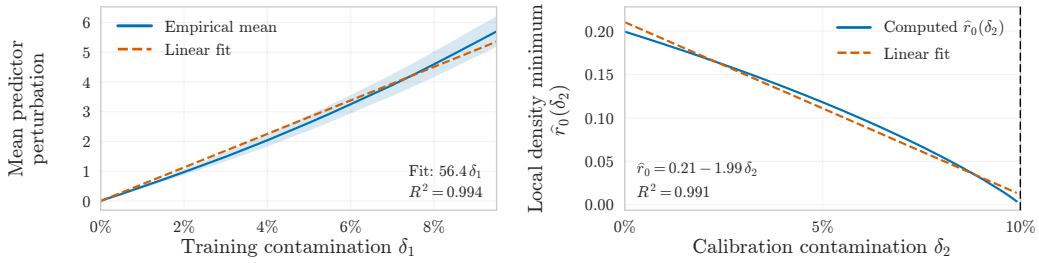
Proposition 4.1 (Exact worst-case calibration deviation)**.** *Suppose $F_{P_0,P_0}(t) < 1$ for every finite t and $\alpha \in (0, 1/2)$. Then, for $\delta_2 \in [0,1)$,*

$$\mathfrak{D}_\alpha(\delta_2) = \begin{cases} 2\max\left\{q_+(\delta_2) - q(P_0;P_0), q(P_0;P_0) - q_-(\delta_2)\right\}, & 0 \leq \delta_2 < \alpha, \\ +\infty, & \alpha \leq \delta_2 < 1. \end{cases}$$

For $\delta_2 < \alpha$, both the maximal inflation and maximal contraction are attained by point masses. For any $\mathbf{x}_0 \in \mathcal{X}$ and $M \geq 0$, let

$$Q_M := \mathrm{Dirac}(\mathbf{x}_0, \mu(\mathbf{x}_0; P_0) + M),$$

which places all contaminated score mass at M. Then $M = 0$ attains maximal contraction, whereas any $M > q_+(\delta_2)$ attains maximal inflation. For $\delta_2 \geq \alpha$, letting $M \to \infty$ yields arbitrarily large inflation. Hence, the breakdown is driven by interval inflation: contraction remains bounded, whereas maximal inflation diverges as $\delta_2 \uparrow \alpha$. Thus, α is the exact breakdown threshold.



(a) Predictor perturbation.

(b) Local density minimum $\widehat{r}_0$.

Figure 3: Training and calibration sensitivity. (a) Predictor perturbation along a fixed contamination path. (b) Local score-density minimum $\widehat{r}_0(\delta_2)$ over the relevant quantile interval.

4.3.2 Nominal Miscoverage-Level Correction

We now turn from interval-length robustness to coverage under the clean distribution P_0. When the calibration sample follows the contaminated distribution P_2, we seek a nominal miscoverage level α' that guarantees coverage at least $1-\alpha$ under P_0. Let $\alpha \in (0,1)$ be the target miscoverage level under P_0, and let $\alpha' \in (0,1)$ be the nominal miscoverage level used by split conformal prediction with calibration data from P_2. Let $\mathrm{Coverage}_P(\alpha')$ denote the marginal coverage of the resulting prediction interval when the test observation follows P. Barber et al. (2023) give a multiplicative correction of $\alpha'_{\mathrm{tight}} := \alpha(1-\delta_2)$, which guarantees

$$\mathrm{Coverage}_{P_0}(\alpha'_{\mathrm{tight}}) \geq 1-\alpha.$$

We further show that this correction is asymptotically sharp.

Proposition 4.2 (Asymptotic sharpness of the mixture-based correction)**.** *For every $\alpha' \in \big(\alpha(1-\delta_2), 1\big)$, there exist a clean distribution P_0, a contamination distribution Q_2, and an absolute-residual split conformal procedure such that, for all sufficiently large n_2,*

$$\mathrm{Coverage}_{P_2}(\alpha') \geq 1-\alpha', \qquad \mathrm{Coverage}_{P_0}(\alpha') < 1-\alpha.$$

5 Experiments

We evaluate interval-length deviation under training and calibration contamination on synthetic and real regression data, relative to the clean population oracle and paired clean split-conformal intervals, respectively. The synthetic experiments isolate the two contamination stages and examine the behavior as δ_2 approaches α, while the real-data experiments assess calibration contamination under structured distributional changes. Experimental setups and supplementary results are provided in the appendices.

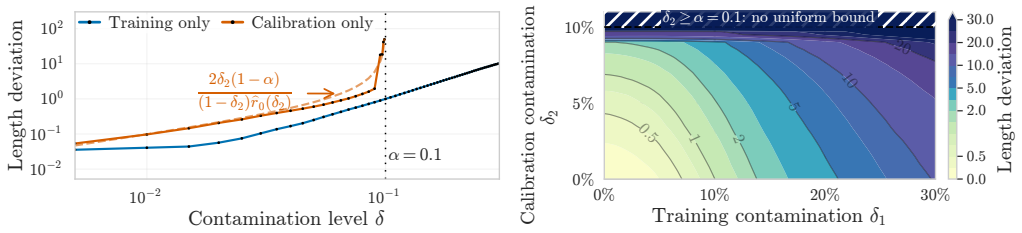
5.1 Synthetic Experiments

We generate $X = (X_1, \ldots, X_5)$ with $X_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([-1,1])$ and $Y = f(X) + \varepsilon$, where

$$f(x) = \sin(\pi x_1) + x_2^2 - \frac{1}{2}x_3$$

and $\varepsilon \sim N(0,1)$. We fit ℓ_2-regularized median regression using the 36 fixed features, with $n_1 = n_2 = 5000$, $n_{\mathrm{test}} = 10000$, and $\alpha = 0.1$.

Figure 3 illustrates the two sensitivity mechanisms underlying the interval-length deviation. Figure 3a shows that the fitted-predictor perturbation grows approximately linearly with δ_1 under the chosen contamination setting. The fitted slope therefore reflects the observed training-side sensitivity in this setting, rather than estimating the uniform stability constant L. Figure 3b shows that the computed minimum of the clean score density over the relevant quantile interval, denoted by $\widehat{r}_0(\delta_2)$, decreases as δ_2 increases, indicating greater calibration sensitivity through the inverse-density factor. Together, these patterns are consistent with the roles of L and r_0^{-1} in Theorem 4.1.



(a) Training-only vs. calibration-only.

(b) Joint contamination over (δ_1, δ_2).

Figure 4: Mean interval-length deviation under point-mass contamination: $Q_1 = Q_2 = \text{Dirac}(x^\star, -80)$, with $x^\star = (-1, -1, -1, -1, +1)$.

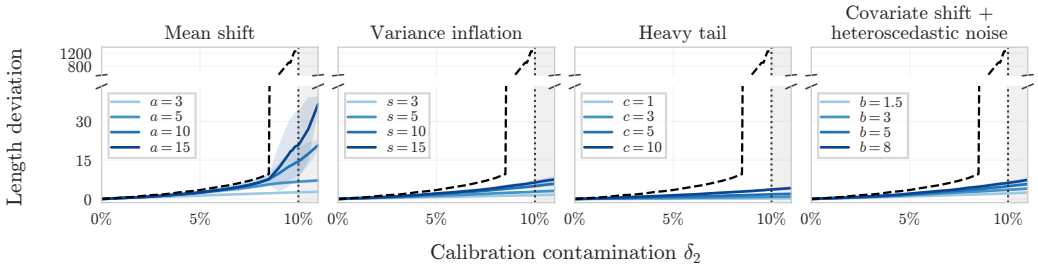


Figure 5: Calibration-only interval-length deviation on `Abalone` under four structured contamination families using the absolute-residual score. The dashed curve shows the finite high-score Dirac reference with score 200 for $\delta_2 < \alpha$.

Figure 4a separates the effects of training and calibration contamination. The training-only deviation increases steadily with δ_1, whereas the calibration-only deviation rises sharply as δ_2 approaches α, consistent with the breakdown behavior in Proposition 4.1. Figure 4b shows that these effects compound under joint contamination: increasing δ_1 raises the deviation across calibration levels, while the dependence on δ_2 becomes increasingly steep near α. These results are consistent with the distinct training- and calibration-side contributions in our theoretical analysis.

5.2 Real-Data Experiments

We evaluate calibration contamination on four real-world regression datasets: `Abalone`, `cadata`, `CPU Small`, and `YearPredictionMSD` (Chang & Lin, 2011). At $\alpha = 0.1$, we consider four structured contamination families: mean shift, variance inflation, heavy-tailed noise, and covariate shift with heteroscedastic noise.

Figure 5 presents the `Abalone` results in the main text; results for the remaining datasets are reported in Appendix D. Across the four contamination families, the interval-length deviation generally increases with both δ_2 and contamination severity. Several settings exhibit a sharp increase as δ_2 approaches α, consistent with the increasing calibration sensitivity predicted by our analysis. The structured contaminations nevertheless remain below the finite-score Dirac curve over most of the pre-breakdown regime, with substantial variation across contamination families.

6 Conclusion and Future Work

We analyze how training and calibration contamination alter split-conformal interval length relative to the clean population oracle. Under Huber contamination, we establish a finite-sample upper bound uniform over contaminating distributions and a local minimax lower bound matching an attainable fixed-confidence upper rate, up to constant factors, on a regular asymmetric location submodel. For calibration contamination, we further characterize the exact worst-case population interval-length deviation. Future work could study contamination effects for conformal procedures that explicitly optimize interval length or prediction-set volume.

AI USE STATEMENT

We used generative AI tools to assist with language editing, terminology improvement, literature searches, and reference checks. For the mathematical analyses, these tools assisted with developing and checking proofs, reviewing assumptions and derivations, and clarifying notation. For the experiments, they assisted with reviewing the experimental design, implementing and debugging code, running simulations, analyzing results, and preparing figures. The authors take full responsibility for the final manuscript, including its mathematical claims, code, and experimental results.

ETHICS STATEMENT

Conformal prediction can help users assess the reliability of regression predictions, particularly in high-stakes applications. Real-world datasets may contain contaminated observations, which can undermine the validity and informativeness of prediction intervals. We study how contamination in the training and calibration samples affects the coverage and interval length of split conformal prediction. Our analysis aims to clarify the reliability, efficiency, and limitations of split conformal prediction under contamination.

The theoretical guarantees rely on the assumptions stated in this paper, and practical use in high-stakes settings requires application-specific validation. Our experiments use synthetic data and publicly available regression benchmarks, and no new data were collected from human participants.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document the experimental procedures and implementation details. Appendix C.1 describes the common experimental pipeline, prediction-set construction, evaluation metrics, and aggregation across repetitions. Appendices C and D detail the synthetic and real-data experiments, including data generation or preprocessing, data splits, model training, contamination mechanisms, and hyperparameter settings. Appendix C.4 specifies the parameter grids, diagnostic computations, random seeds, and software versions. The fixed-fraction replacement experiments are described in Appendix E. We provide the source code in the supplementary material.

REFERENCES

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. doi: 10.1214/23-AOS2276.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Sacha Braun, Liviu Aolaritei, Michael I. Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *arXiv preprint arXiv:2503.19068*, 2025.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. URL `https://dl.acm.org/doi/10.1145/1961189.1961199`.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Jase Clarkson, Wenkai Xu, Mihai Cucuringu, Yvik Swan, and Gesine Reinert. Split conformal prediction under data contamination, 2024.

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

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N. Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25(328):1–66, 2024.

Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets. In *International Conference on Machine Learning*, pp. 18495–18527. PMLR, 2025.

Matteo Gasparin and Aaditya Ramdas. Improving the statistical efficiency of cross-conformal prediction. In *International Conference on Machine Learning*, 2025.

Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, 2016. doi: 10.1016/j.ijforecast.2016.02.001.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

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

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563, 2024.

Batiste Le Bars and Pierre Humbert. On volume minimization in conformal regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3030–3053. PMLR, 2025. URL https://proceedings.mlr.press/v267/bars25a.html.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pp. 38–53, 1973.

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.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Yangyi Li, Aobo Chen, Wei Qian, Chenxu Zhao, Divya Lidder, and Mengdi Huai. Data poisoning attacks against conformal prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27563–27574. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24l.html.

Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2159–2167. PMLR, 2017.

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.

Pierre Pinson. Wind energy: Forecasting challenges for its operational management. *Statistical Science*, 28(4):564–585, 2013. doi: 10.1214/13-STS445.

Mehrdad Pournaderi and Yu Xiang. Training-conditional coverage bounds for uniformly stable learning algorithms. *arXiv preprint arXiv:2404.13731*, 2024. Accepted to the 2024 IEEE ISIT Workshop on Information-Theoretic Methods for Trustworthy Machine Learning.

Matteo Sesia and Emmanuel J. Candes. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020. doi: 10.1002/sta4.261.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *The Annals of Statistics*, pp. 1566–1590, 2009.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *Symposium on conformal and probabilistic prediction with applications*, pp. 23–39. Springer, 2016.

Yuxin Yang, Qiang Li, Runyang Feng, Liren Shan, and Binghui Wang. Ensemble conformal predictor (encp): A new conformal predictor with robustness guarantees against data poisoning attacks. In *2026 IEEE Symposium on Security and Privacy (SP)*, pp. 156–174. IEEE, 2026.

Yunzhen Yao, Lie He, and Michael Gastpar. Non-asymptotic analysis of efficiency in conformalized regression. In *International Conference on Learning Representations*, volume 2026, pp. 42546–42583, 2026.

Soroush H Zargarbashi, Mohammad Sadegh Akhondzadeh, and Aleksandar Bojchevski. Robust yet efficient conformal prediction sets. *arXiv preprint arXiv:2407.09165*, 2024.

Matteo Zecchin, Sangwoo Park, Osvaldo Simeone, and Fredrik Hellström. Generalization and informativeness of conformal prediction. In *International Symposium on Information Theory*, pp. 244–249. IEEE, 2024.

APPENDIX CONTENTS

A List of Notations **14**

B Proofs **15**

B.1 Proofs of Lemmas for Theorem 4.1 15

B.2 Proof of Theorem 4.1 18

B.3 Proof of Proposition 4.1 18

B.4 Derivation of the Mixture-Based Correction 19

B.5 Proof of Proposition 4.2 20

B.6 Extensions to CQR and Adaptive Residual Scores 21

B.7 Derivation for Example 4.1 26

B.8 Formal Joint Minimax Lower Bound and Proof 26

B.9 Fixed-Confidence Upper Bound and Local Rate Matching 33

C Experimental details **34**

C.1 Common protocol 34

C.2 Synthetic setup 36

C.3 Contamination designs 36

C.4 Grids, diagnostics, and reproducibility 37

C.5 Further synthetic results 39

D Experiments on real data **42**

D.1 Data and model fitting 42

D.2 Training residual scale and contamination 42

D.3 Calibration-level correction 43

D.4 Calibration contamination and interval length 43

D.5 Score distributions 47

E Bernoulli and Fixed-Fraction Replacement **50**

A LIST OF NOTATIONS

Symbol	Description
Data, distributions, and contamination	
$(\mathbf{X}_i, Y_i)$, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$	Observation pairs and a generic feature value in the d-dimensional feature space.
$\mathcal{D}_0, \mathcal{D}_1, \mathcal{D}_2$; n_1, n_2	Coupled clean training sample, contaminated training sample, calibration sample, and their sample sizes.
P	A generic distribution on the observation space; subscripts identify its role.
P_0	Clean distribution.
Q, δ, P_δ	Contamination distribution, contamination level, and Huber mixture $P_\delta = (1-\delta)P_0 + \delta Q$.
P_j, Q_j, δ_j, $j \in \{1, 2\}$	Training-side ($j = 1$) and calibration-side ($j = 2$) contaminated laws, contaminating laws, and contamination levels, with $P_j = (1-\delta_j)P_0 + \delta_j Q_j$.
Split conformal construction and clean target	
A, B	Generic training source A (a population law or realized dataset) and evaluation distribution B.
$\mu(\cdot; \mathcal{D}_n)$, $\mu(\cdot; P)$	Fitted predictor and its deterministic population target under training law P.
$\|g\|_\infty$	Supremum norm $\sup_{\mathbf{x}\in\mathcal{X}} \|g(\mathbf{x})\|$.
$\eta_n(P)$, $\rho_n(P)$	Sampling-stability error and failure-probability sequences in Assumption 3.1.
K	Number of contaminated training observations.
α	Target miscoverage level under the clean distribution P_0.
γ	User-specified failure probability in finite-sample deviation bounds.
$R_A(\mathbf{x}, y)$, $R_{A,i}$	Absolute-residual score and its value on calibration observation i.
$F_{A,B}$, $\widehat{F}_{A,\mathcal{D}_2}$	Population and calibration empirical CDFs of the absolute-residual score.
$q(A; B)$, $\widehat{q}(A; \mathcal{D}_2)$	Population $(1-\alpha)$-quantile and conformal empirical quantile.
k_{n_2}	Conformal order-statistic rank $\lceil (n_2+1)(1-\alpha) \rceil$.
$C_\alpha(\cdot; \mathcal{D}_1, \mathcal{D}_2)$	Split-conformal prediction interval; its length is $2\widehat{q}(\mathcal{D}_1; \mathcal{D}_2)$, and the clean oracle length is $2q(P_0; P_0)$.
$q_-(\delta_2)$, $q_+(\delta_2)$, f_{P_0,P_0}, r_0	Local quantile endpoints, clean oracle-score density, and its positive lower bound in Assumption 3.2.
$\mathfrak{D}^+_\alpha(\delta_2)$, $\mathfrak{D}^-_\alpha(\delta_2)$, $\mathfrak{D}_\alpha(\delta_2)$	Worst-case interval-length inflation, contraction, and absolute deviation under calibration contamination.
L, k	Absolute-score replacement-stability constant and maximum number of replaced training observations.
B_x, ρ_τ, τ, λ, ω, $\hat{\omega}(\mathcal{D}_1)$	Feature-norm bound, pinball loss, quantile level, regularization parameter, coefficient vector, and fitted regularized quantile-regression estimator in Example 4.1.
Alternative nonconformity scores	
S	Score type in $\{\mathrm{Absolute}, \mathrm{CQR}, \mathrm{Adaptive}\}$.
$\ell(\cdot; A)$, $u(\cdot; A)$	Lower and upper CQR regressors associated with training source A.
$\mu(\cdot; A)$, $\sigma(\cdot; A)$	Location and scale regressors for the adaptive score; μ is the same location learner used by the absolute score.
$R^S(A; \mathbf{x}, y)$, $R^S_i(A)$	Score-S nonconformity function and its calibration values.
$F^S_{P_0,P}$, $q_S(P_0; P)$, $\widehat{q}_S(A; \mathcal{D}_2)$	Clean-oracle score CDF, population quantile, and conformal empirical quantile for score S.
$\eta^{\mathrm{CQR}}_{n_1}(P_0)$, $\eta^{\mathrm{Adaptive}}_{g,n_1}(P_0)$	Clean-sampling errors for the CQR regressors and adaptive component $g \in \{\mu, \sigma\}$.
L_{CQR}, L^{Adaptive}_g	Componentwise replacement-stability constants for the CQR regressors and adaptive component $g \in \{\mu, \sigma\}$.

Symbol	Description
$\varepsilon_{g,n_1}^{\mathrm{Adaptive}}$	Clean-anchored adaptive-component perturbation $\eta_{g,n_1}^{\mathrm{Adaptive}}(P_0) + L_g^{\mathrm{Adaptive}}(\delta_1 + \sqrt{\log(n_1)/n_1})$.
$[z]_+$	Positive part $\max\{z, 0\}$.
$\varepsilon_{\mathrm{train},n_1}^S$, $\rho_{n_1}^S(P_0)$	Training contribution to the score-quantile deviation and clean-sampling failure probability for score S.
$\varepsilon_{\mathrm{cal},n_2}^S(\gamma)$	Calibration contamination-and-sampling contribution to the score-quantile deviation.
$\varepsilon_{q,n_1,n_2}^S(\gamma)$	Total score-quantile deviation bound for score S.
r_0^S	Local positive density lower bound for the clean distribution of score S.
$\underline{\sigma}$, $\overline{\sigma}$, $\kappa_{\sigma,n_1}^{\mathrm{Adaptive}}$, $\beta_{\mu,n_1}^{\mathrm{Adaptive}}$	Global positive lower scale bound, clean-scale upper bound, relative scale perturbation $\varepsilon_{\sigma,n_1}^{\mathrm{Adaptive}}/\underline{\sigma}$, and normalized location perturbation $\varepsilon_{\mu,n_1}^{\mathrm{Adaptive}}/\underline{\sigma}$.
Joint minimax lower bound	
Θ, P_θ, θ	Parameter space, clean submodel law, and its location parameter in the minimax lower bound.
$q_\star$, $L_\star$, h_0, s_0, $r_\star$	Clean radius, information scale, local radii, and score-density scale of the least-favorable location submodel.
$\widehat{\theta}$, $\widehat{\theta}_1$, $\widehat{q}(u; \mathcal{D}_2)$	Estimator, its realized output $\widehat{\theta}(\mathcal{D}_1)$, and the calibration quantile for the constant predictor $u \in \Theta$.
$\mathcal{E}_{n_1}(\Theta)$	Class of measurable, possibly randomized, Θ-valued estimators in the lower bound.
$\mathbb{P}_{\theta,Q_1,Q_2}$	Joint probability law used in the minimax lower-bound experiments.
$\mathfrak{p}_{n_1,n_2}^{\mathrm{split}}(t)$	Minimax tail-probability profile for interval-length deviation.
a_{n_1}, b_{n_2}	Training-side and calibration-side minimax difficulty scales.
$x \wedge y$	Minimum $\min\{x, y\}$.
Set discrepancy and coverage correction	
$C_{\mathrm{clean}}, C_{\mathrm{cont}}$	Clean oracle and contaminated split-conformal intervals in the set-discrepancy comparison.
Leb, $\triangle$	Lebesgue measure on $\mathbb{R}$ and symmetric difference.
α', α'_{tight}, α'_{naive}	Nominal miscoverage level and its mixture-based and naive corrections.
$\mathrm{Coverage}_P(\alpha')$	Marginal coverage at nominal miscoverage α' under test law $P \in \{P_0, P_2, Q_2\}$.
$O_{\mathbb{P}}(\cdot)$	Big-O in probability.

B PROOFS

B.1 PROOFS OF LEMMAS FOR THEOREM 4.1

Lemma B.1 (Order-statistic perturbation). *Let $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$ be two collections of real numbers. If*

$$\max_{1 \le i \le m} |a_i - b_i| \le c,$$

then, for every $j \in \{1, \ldots, m\}$,

$$|a_{(j)} - b_{(j)}| \le c,$$

where $a_{(j)}$ and $b_{(j)}$ denote the corresponding j-th order statistics.

Proof. At least j indices satisfy $b_i \le b_{(j)}$. For every such index, $a_i \le b_i + c \le b_{(j)} + c$. Hence at least j elements of the first collection are no greater than $b_{(j)} + c$, which implies $a_{(j)} \le b_{(j)} + c$. Interchanging the two collections gives $b_{(j)} \le a_{(j)} + c$, and the result follows. □

Proposition B.1 (Training-contamination perturbation)**.** *If μ is L-uniformly replacement stable as in Definition 4.1, then, for $n_1 \geq 2$, with probability at least $1 - n_1^{-2}$,*

$$\|\mu(\cdot; \mathcal{D}_1) - \mu(\cdot; \mathcal{D}_0)\|_\infty \leq L \left(\delta_1 + \sqrt{\frac{\log n_1}{n_1}} \right), \tag{2}$$

uniformly over the contamination distribution Q_1.

Proof. For $i = 1, \ldots, n_1$, independently draw

$$Z_i^{(0)} \sim P_0, \qquad Z_i^{(Q)} \sim Q_1, \qquad U_i \sim \text{Bernoulli}(\delta_1),$$

and set

$$Z_i^{(1)} := \begin{cases} Z_i^{(0)}, & U_i = 0, \\ Z_i^{(Q)}, & U_i = 1. \end{cases}$$

Then

$$\mathcal{D}_0 := (Z_i^{(0)})_{i=1}^{n_1} \sim P_0^{n_1}, \qquad \mathcal{D}_1 := (Z_i^{(1)})_{i=1}^{n_1} \sim P_1^{n_1}.$$

The coupled datasets $\mathcal{D}_1$ and $\mathcal{D}_0$ differ in at most $K := \sum_{i=1}^{n_1} U_i$ observations, where $K \sim \text{Binomial}(n_1, \delta_1)$. The uniform replacement-stability bound in Definition 4.1 therefore gives

$$\|\mu(\cdot; \mathcal{D}_1) - \mu(\cdot; \mathcal{D}_0)\|_\infty \leq LK/n_1 \tag{3}$$

almost surely, proving (3). Hoeffding's inequality yields

$$\mathbb{P}\left\{ K/n_1 > \delta_1 + \sqrt{\frac{\log n_1}{n_1}} \right\} \leq \exp(-2 \log n_1) = n_1^{-2}.$$

Combining the two displays proves (2), uniformly over Q_1. □

Lemma 4.1 (Training-contamination quantile perturbation)**.** *If μ is L-uniformly replacement stable, then, for all sufficiently large n_2, the following holds with probability at least $1 - n_1^{-2}$, uniformly over Q_1:*

$$|\widehat{q}(\mathcal{D}_1; \mathcal{D}_2) - \widehat{q}(\mathcal{D}_0; \mathcal{D}_2)| \leq L \left(\delta_1 + \sqrt{\frac{\log n_1}{n_1}} \right). \tag{4}$$

Proof of Lemma 4.1. On the event in (2), the reverse triangle inequality gives, for every calibration observation,

$$|R_{\mathcal{D}_1,i} - R_{\mathcal{D}_0,i}| \leq |\mu(\mathbf{X}_i; \mathcal{D}_1) - \mu(\mathbf{X}_i; \mathcal{D}_0)| \leq L \left(\delta_1 + \sqrt{\frac{\log n_1}{n_1}} \right).$$

Applying Lemma B.1 at rank k_{n_2} proves (4). The event has probability at least $1 - n_1^{-2}$ uniformly over Q_1 by Proposition B.1. □

Lemma 4.2 (Training-stage sampling quantile error)**.** *Under Assumption 3.1, for all sufficiently large n_1, n_2, with probability at least $1 - \rho_{n_1}(P_0)$,*

$$|\widehat{q}(\mathcal{D}_0; \mathcal{D}_2) - \widehat{q}(P_0; \mathcal{D}_2)| \leq \eta_{n_1}(P_0). \tag{5}$$

Proof of Lemma 4.2. By Assumption 3.1, with probability at least $1 - \rho_{n_1}(P_0)$,

$$\|\mu(\cdot; \mathcal{D}_0) - \mu(\cdot; P_0)\|_\infty \leq \eta_{n_1}(P_0).$$

On this event, for every calibration observation,

$$|R_{\mathcal{D}_0,i} - R_{P_0,i}| \leq |\mu(\mathbf{X}_i; \mathcal{D}_0) - \mu(\mathbf{X}_i; P_0)| \leq \eta_{n_1}(P_0).$$

Lemma B.1 at rank k_{n_2} proves (5). □

Lemma 4.4 (Quantile bias under calibration contamination)**.** *Under Assumption 3.2 and $\delta_2 < \alpha$, the population $(1-\alpha)$-quantile of the oracle score R_{P_0} satisfies*

$$|q(P_0; P_2) - q(P_0; P_0)| \le \frac{\delta_2(1-\alpha)}{(1-\delta_2)r_0}. \tag{6}$$

For reference, the fixed-source mixture identity used in the proof is

$$F_{A,P_2}(t) = (1-\delta_2)F_{A,P_0}(t) + \delta_2 F_{A,Q_2}(t). \tag{7}$$

Proof of Lemma 4.4. Since $0 \le F_{P_0,Q_2}(t) \le 1$, the mixture identity (7) gives

$$(1-\delta_2)F_{P_0,P_0}(t) \le F_{P_0,P_2}(t) \le (1-\delta_2)F_{P_0,P_0}(t) + \delta_2.$$

Taking generalized inverses at level $1-\alpha$ yields

$$F^{-1}_{P_0,P_0}\left(\frac{1-\alpha-\delta_2}{1-\delta_2}\right) \le q(P_0; P_2) \le F^{-1}_{P_0,P_0}\left(\frac{1-\alpha}{1-\delta_2}\right). \tag{8}$$

The density lower bound makes $F^{-1}_{P_0,P_0}$ $1/r_0$-Lipschitz on this quantile interval. The probability-level distances from $1-\alpha$ to the lower and upper endpoints in (8) are

$$\frac{\alpha\delta_2}{1-\delta_2} \quad \text{and} \quad \frac{(1-\alpha)\delta_2}{1-\delta_2},$$

respectively. Since $\alpha < 1/2$, both are bounded by $(1-\alpha)\delta_2/(1-\delta_2)$, which proves (6). □

Lemma 4.3 (Finite-sample calibration quantile error)**.** *Under Assumption 3.2, for every $\gamma \in (0,1)$ and all sufficiently large n_2, the following holds with probability at least $1-\gamma$:*

$$|\widehat{q}(P_0; \mathcal{D}_2) - q(P_0; P_2)| \le \frac{1}{(1-\delta_2)r_0}\left(\sqrt{\frac{\log(2/\gamma)}{2n_2}} + \frac{2}{n_2}\right). \tag{9}$$

Proof of Lemma 4.3. For any $s < t$ in $\mathcal{I}$, the mixture identity, monotonicity of F_{P_0,Q_2}, and the density lower bound imply

$$F_{P_0,P_2}(t) - F_{P_0,P_2}(s) \ge (1-\delta_2)r_0(t-s). \tag{10}$$

Consequently, the generalized inverse of F_{P_0,P_2} is $1/((1-\delta_2)r_0)$-Lipschitz whenever the relevant quantiles lie in $\mathcal{I}$.

Let

$$\widehat{F}_{P_0,\mathcal{D}_2}(t) := \frac{1}{n_2}\sum_{i=1}^{n_2} \mathbf{1}\{R_{P_0}(\mathbf{X}_i, Y_i) \le t\}$$

be the empirical CDF of the oracle calibration scores, and set

$$\varepsilon_{n_2}(\gamma) := \sqrt{\frac{\log(2/\gamma)}{2n_2}}, \qquad u_{n_2} := \frac{k_{n_2}}{n_2}.$$

The Dvoretzky–Kiefer–Wolfowitz inequality gives, with probability at least $1-\gamma$,

$$\sup_t \left|\widehat{F}_{P_0,\mathcal{D}_2}(t) - F_{P_0,P_2}(t)\right| \le \varepsilon_{n_2}(\gamma).$$

On this event,

$$F^{-1}_{P_0,P_2}\left(u_{n_2} - \varepsilon_{n_2}(\gamma)\right) \le \widehat{q}(P_0; \mathcal{D}_2) \le F^{-1}_{P_0,P_2}\left(u_{n_2} + \varepsilon_{n_2}(\gamma)\right).$$

Since $k_{n_2} = \lceil (n_2+1)(1-\alpha) \rceil$ and $k_{n_2} \le n_2$ for all sufficiently large n_2,

$$0 \le u_{n_2} - (1-\alpha) \le \frac{2}{n_2}.$$

Because $\mathcal{I}$ is open and contains the compact interval $[q_-(\delta_2), q_+(\delta_2)]$, choose $d_{\mathcal{I}} > 0$ such that

$$[q_-(\delta_2) - d_{\mathcal{I}}, q_+(\delta_2) + d_{\mathcal{I}}] \subset \mathcal{I}.$$

The sandwich in Lemma 4.4 and (10) show that the inverse images of $u_{n_2} \pm \varepsilon_{n_2}(\gamma)$ remain in this enlarged interval once

$$\varepsilon_{n_2}(\gamma) + 2/n_2 < (1-\delta_2)r_0 d_{\mathcal{I}}.$$

Thus the relevant quantiles lie in $\mathcal{I}$ for all sufficiently large n_2, and the inverse-Lipschitz bound following from (10) proves (9).

□

B.2 Proof of Theorem 4.1

Theorem 4.1 (Interval-length deviation under joint contamination). *Under Assumptions 3.1 and 3.2, suppose also that μ is L-uniformly replacement stable as in Definition 4.1. Fix $\delta_1 \in [0,1)$ and $\delta_2 \in [0,\alpha)$. For every $\gamma \in (0,1)$ and all sufficiently large n_1, n_2, the following holds with probability at least $1-\rho_{n_1}(P_0)-n_1^{-2}-\gamma$, uniformly over the contamination distributions Q_1 and Q_2:*

$$2\left|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2)-q(P_0;P_0)\right| \le 2\eta_{n_1}(P_0)+2L\left(\delta_1+\sqrt{\frac{\log n_1}{n_1}}\right)+\frac{2}{(1-\delta_2)r_0}\left(\sqrt{\frac{\log(2/\gamma)}{2n_2}}+\frac{2}{n_2}+\delta_2(1-\alpha)\right). \tag{11}$$

For reference, the quantile-error decomposition used in the proof is

$$\left|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2)-q(P_0;P_0)\right| \le \underbrace{\left|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2)-\widehat{q}(\mathcal{D}_0;\mathcal{D}_2)\right|}_{\text{Lemma 4.1: Training contamination error}} + \underbrace{\left|\widehat{q}(\mathcal{D}_0;\mathcal{D}_2)-\widehat{q}(P_0;\mathcal{D}_2)\right|}_{\text{Lemma 4.2: Training-stage sampling error}} + \underbrace{\left|\widehat{q}(P_0;\mathcal{D}_2)-q(P_0;P_2)\right|}_{\text{Lemma 4.3: Calibration-stage sampling error}} + \underbrace{\left|q(P_0;P_2)-q(P_0;P_0)\right|}_{\text{Lemma 4.4: Calibration contamination error}}. \tag{12}$$

Proof. Fix $\gamma \in (0,1)$. On the intersection of the events in Lemmas 4.1, 4.2, and 4.3, combining (12) with Lemma 4.4 gives

$$\left|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2)-q(P_0;P_0)\right| \le \eta_{n_1}(P_0)+L\left(\delta_1+\sqrt{\frac{\log n_1}{n_1}}\right)+\frac{1}{(1-\delta_2)r_0}\left(\sqrt{\frac{\log(2/\gamma)}{2n_2}}+\frac{2}{n_2}+\delta_2(1-\alpha)\right). \tag{13}$$

The intersection has probability at least $1-\rho_{n_1}(P_0)-n_1^{-2}-\gamma$. Multiplying (13) by two yields, on the same event,

$$\begin{aligned}2\left|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2)-q(P_0;P_0)\right| &\le 2\eta_{n_1}(P_0)+2L\left(\delta_1+\sqrt{\frac{\log n_1}{n_1}}\right)+\frac{2}{(1-\delta_2)r_0}\left(\sqrt{\frac{\log(2/\gamma)}{2n_2}}+\frac{2}{n_2}\right)\\ &\quad+\frac{2\delta_2(1-\alpha)}{(1-\delta_2)r_0}.\end{aligned}$$

This is (11).

□

Asymptotic consequence. For completeness, the finite-sample decomposition also yields

$$2\left|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2)-q(P_0;P_0)\right| = O_{\mathbb{P}}\left(\eta_{n_1}(P_0)+L\left(\delta_1+n_1^{-1/2}\right)+\frac{1}{(1-\delta_2)r_0\sqrt{n_2}}+\frac{\delta_2(1-\alpha)}{(1-\delta_2)r_0}\right). \tag{14}$$

Indeed, use the almost-sure bound (3) directly instead of its $1-n_1^{-2}$ envelope. Since $K/n_1 = \delta_1 + O_{\mathbb{P}}(n_1^{-1/2})$ and $\rho_{n_1}(P_0) = o(1)$, the two training terms in (12) are jointly $O_{\mathbb{P}}(\eta_{n_1}(P_0) + L(\delta_1+n_1^{-1/2}))$. The DKW inequality gives an $O_{\mathbb{P}}(n_2^{-1/2}/((1-\delta_2)r_0))$ calibration-sampling term, and Lemma 4.4 supplies the remaining deterministic term. Multiplying by two proves (14).

B.3 Proof of Proposition 4.1

Proposition 4.1 (Exact worst-case calibration deviation). *Suppose $F_{P_0,P_0}(t) < 1$ for every finite t and $\alpha \in (0,1/2)$. Then, for $\delta_2 \in [0,1)$,*

$$\mathfrak{D}_\alpha(\delta_2) = \begin{cases} 2\max\left\{q_+(\delta_2)-q(P_0;P_0),\, q(P_0;P_0)-q_-(\delta_2)\right\}, & 0 \le \delta_2 < \alpha,\\ +\infty, & \alpha \le \delta_2 < 1. \end{cases} \tag{15}$$

Proof. Write $F_0 := F_{P_0,P_0}$ and $q_0 := F_0^{-1}(1-\alpha)$, and define

$$\Delta_{\delta_2}(Q_2) := q(P_0; (1-\delta_2)P_0 + \delta_2 Q_2) - q_0.$$

For the proof, introduce the one-sided worst-case deviations

$$\mathfrak{D}_\alpha^+(\delta_2) := 2\sup_{Q_2} \Delta_{\delta_2}(Q_2), \qquad \mathfrak{D}_\alpha^-(\delta_2) := 2\sup_{Q_2}\{-\Delta_{\delta_2}(Q_2)\}.$$

Then $\mathfrak{D}_\alpha(\delta_2) = 2\sup_{Q_2} |\Delta_{\delta_2}(Q_2)|$. Fix any $\mathbf{x}_0 \in \mathcal{X}$ and $M \geq 0$, and let Q_M be the point mass at $(\mathbf{x}_0, \mu(\mathbf{x}_0; P_0) + M)$, whose oracle score equals M almost surely.

First suppose that $0 \leq \delta_2 < \alpha$. For every Q_2, the quantile sandwich (8) gives

$$q_-(\delta_2) \leq q(P_0; P_2) \leq q_+(\delta_2).$$

Consequently,

$$\mathfrak{D}_\alpha^+(\delta_2) \leq 2\big(q_+(\delta_2) - q_0\big), \qquad \mathfrak{D}_\alpha^-(\delta_2) \leq 2\big(q_0 - q_-(\delta_2)\big).$$

For any $M > q_+(\delta_2)$, the contamination-score CDF under Q_M is zero below M. Hence, below M, the mixture CDF equals $(1-\delta_2)F_0$, and its $(1-\alpha)$-quantile is $q_+(\delta_2)$. Thus, Q_M attains the upper bound for $\mathfrak{D}_\alpha^+(\delta_2)$. Under Q_0, all contaminated scores equal zero, so for $t \geq 0$ the mixture CDF is

$$\delta_2 + (1-\delta_2)F_0(t).$$

Since $1-\alpha > \delta_2$, its $(1-\alpha)$-quantile is $q_-(\delta_2)$, and Q_0 attains the upper bound for $\mathfrak{D}_\alpha^-(\delta_2)$. Therefore,

$$\mathfrak{D}_\alpha^+(\delta_2) = 2\big(q_+(\delta_2) - q_0\big), \qquad \mathfrak{D}_\alpha^-(\delta_2) = 2\big(q_0 - q_-(\delta_2)\big). \tag{16}$$

Finally,

$$\sup_{Q_2} |a(Q_2)| = \max\left\{\sup_{Q_2} a(Q_2), \sup_{Q_2}[-a(Q_2)]\right\},$$

applied to the contaminated-minus-clean quantile difference, gives (15).

Now suppose that $\delta_2 \geq \alpha$ and take M arbitrarily large. For every $t < M$,

$$F_{P_0,P_2}(t) = (1-\delta_2)F_0(t) < 1-\alpha.$$

When $\delta_2 = \alpha$, strictness follows from the unbounded-upper-support assumption; when $\delta_2 > \alpha$, it follows from $1-\delta_2 < 1-\alpha$. For all sufficiently large M, $F_{P_0,P_2}(M) \geq 1-\alpha$, so the $(1-\alpha)$-quantile equals M. Letting $M \to \infty$ proves that both $\mathfrak{D}_\alpha^+(\delta_2)$ and $\mathfrak{D}_\alpha(\delta_2)$ are infinite. □

B.4 Derivation of the Mixture-Based Correction

Corollary B.1 (Mixture-based nominal-level correction)**.** *Define* $\alpha'_{\text{tight}} := \alpha(1-\delta_2)$. *If*

$$\text{Coverage}_{P_2}(\alpha'_{\text{tight}}) \geq 1 - \alpha'_{\text{tight}},$$

then

$$\text{Coverage}_{P_0}(\alpha'_{\text{tight}}) \geq 1-\alpha.$$

Proof. Conditional on $\mathcal{D}_1$ and $\mathcal{D}_2$, the mixture identity for the new test observation holds for the coverage event. Averaging over the two datasets gives

$$\text{Coverage}_{P_2}(\alpha'_{\text{tight}}) = (1-\delta_2)\,\text{Coverage}_{P_0}(\alpha'_{\text{tight}}) + \delta_2\,\text{Coverage}_{Q_2}(\alpha'_{\text{tight}}).$$

Since $\text{Coverage}_{Q_2}(\alpha'_{\text{tight}}) \leq 1$, the assumed P_2-coverage guarantee implies

$$\text{Coverage}_{P_0}(\alpha'_{\text{tight}}) \geq \frac{1 - \alpha'_{\text{tight}} - \delta_2}{1-\delta_2} = 1-\alpha.$$

□

B.5 Proof of Proposition 4.2

Proposition 4.2 (Asymptotic sharpness of the mixture-based correction)**.** *For every $\alpha' \in \big(\alpha(1-\delta_2), 1\big)$, there exist a clean distribution P_0, a contamination distribution Q_2, and an absolute-residual split conformal procedure such that, for all sufficiently large n_2,*

$$\mathrm{Coverage}_{P_2}(\alpha') \geq 1-\alpha', \qquad \mathrm{Coverage}_{P_0}(\alpha') < 1-\alpha.$$

Proof. Fix $\bar{\alpha} \in (\alpha(1-\delta_2), 1)$. Let the feature space contain only the origin, so that $\mathbf{X} = 0$ almost surely, and define the clean distribution by

$$Y \mid \mathbf{X} = 0 \sim \mathcal{N}(0,1) \qquad \text{under } P_0.$$

Thus $\mathbb{E}_{P_0}[Y \mid \mathbf{X} = 0] = 0$, and the clean noise is independent of $\mathbf{X}$, has a symmetric density, and is nonincreasing on $[0, \infty)$. Let the contamination distribution satisfy

$$Q_2(\mathbf{X} = 0, Y = 0) = 1.$$

Fit a no-intercept linear regression estimator $\mu(\mathbf{x}; \mathcal{D}_1) = \hat{\beta}^\top \mathbf{x}$ on the training sample. Since every feature vector is zero, the fitted predictor satisfies $\mu(0; \mathcal{D}_1) = 0$ for every training sample. Hence the calibration and test scores are

$$R = |Y - \mu(\mathbf{X}; \mathcal{D}_1)| = |Y|.$$

Let G_0 and G_2 denote the CDFs of R under P_0 and $P_2 = (1-\delta_2)P_0 + \delta_2 Q_2$, respectively. For every $t \geq 0$,

$$G_2(t) = \delta_2 + (1-\delta_2)G_0(t). \tag{17}$$

Let

$$q_{\bar{\alpha}} := G_2^{-1}(1-\bar{\alpha}).$$

Because G_0 is continuous and strictly increasing on $[0, \infty)$, (17) gives

$$G_0(q_{\bar{\alpha}}) = \max\left\{0, 1 - \frac{\bar{\alpha}}{1-\delta_2}\right\} < 1-\alpha,$$

where the strict inequality follows from $\bar{\alpha} > \alpha(1-\delta_2)$.

Now let the calibration size be $n_2 = m$, set

$$k_m := \lceil (m+1)(1-\bar{\alpha}) \rceil,$$

and let $\hat{q}_m$ be the k_m-th order statistic of the calibration scores. For every $\varepsilon > 0$, the definition of $q_{\bar{\alpha}}$ and the strict increase of G_0 give

$$G_2(q_{\bar{\alpha}} - \varepsilon) < 1 - \bar{\alpha} < G_2(q_{\bar{\alpha}} + \varepsilon).$$

The same inequalities hold when $q_{\bar{\alpha}} = 0$, despite the atom of G_2 at zero. Applying the law of large numbers to the proportions of calibration scores below the two endpoints, and using $k_m/m \to 1-\bar{\alpha}$, yields

$$\hat{q}_m \longrightarrow q_{\bar{\alpha}} \quad \text{in probability.} \tag{18}$$

The resulting interval is $C_{\bar{\alpha}} = [-\hat{q}_m, \hat{q}_m]$. Since the calibration sample is i.i.d. from P_2, the usual finite-sample split conformal guarantee gives, for all sufficiently large m,

$$\mathrm{Coverage}_{P_2}(\bar{\alpha}) \geq 1-\bar{\alpha}. \tag{19}$$

Moreover, $Y = 0$ almost surely under Q_2, and hence $\mathrm{Coverage}_{Q_2}(\bar{\alpha}) = 1$ for every m. On the other hand, independence of the calibration and clean test samples gives

$$\mathrm{Coverage}_{P_0}(\bar{\alpha}) = \mathbb{E}\left[G_0(\hat{q}_m)\right].$$

By (18) and continuity of G_0, we have $G_0(\hat{q}_m) \to G_0(q_{\bar{\alpha}})$ in probability. Because this sequence is bounded, its expectations converge as well. Therefore,

$$\mathrm{Coverage}_{P_0}(\bar{\alpha}) \longrightarrow G_0(q_{\bar{\alpha}}) < 1-\alpha.$$

Therefore, the clean coverage is below $1-\alpha$ for all sufficiently large m, while (19) continues to hold. This proves Proposition 4.2.

□

B.6 Extensions to CQR and Adaptive Residual Scores

This subsection develops the alternative-score extensions outlined in Remark 4.1. We introduce the score-specific conditions and establish finite-sample bounds for score quantiles and interval lengths, uniformly over the training and calibration contamination distributions. Absolute residuals are retained as a baseline to make the connection with Theorem 4.1 explicit.

B.6.1 Score Definitions and Stability Conditions

For $A \in \{\mathcal{D}_1, \mathcal{D}_0, P_0\}$, define

$$\begin{aligned} R^{\mathrm{Absolute}}(A; \mathbf{x}, y) &:= |y - \mu(\mathbf{x}; A)|, \\ R^{\mathrm{CQR}}(A; \mathbf{x}, y) &:= \max\{\ell(\mathbf{x}; A) - y,\, y - u(\mathbf{x}; A)\}, \\ R^{\mathrm{Adaptive}}(A; \mathbf{x}, y) &:= \frac{|y - \mu(\mathbf{x}; A)|}{\sigma(\mathbf{x}; A)}. \end{aligned}$$

Here $\ell(\cdot; A)$ and $u(\cdot; A)$ are the lower and upper CQR regressors, and $\sigma(\cdot; A) > 0$ is the adaptive scale regressor. The choices $A = \mathcal{D}_1$ and $A = \mathcal{D}_0$ denote fits on the coupled contaminated and clean samples, whereas $A = P_0$ denotes the clean population target.

For $S \in \{\mathrm{Absolute}, \mathrm{CQR}, \mathrm{Adaptive}\}$ and $P \in \{P_0, P_2, Q_2\}$, define

$$F^S_{P_0,P}(t) := \mathbb{P}_{(\mathbf{X},Y)\sim P}\{R^S(P_0; \mathbf{X}, Y) \le t\}, \qquad q_S(P_0; P) := (F^S_{P_0,P})^{-1}(1-\alpha).$$

For $(\mathbf{X}_i, Y_i) \in \mathcal{D}_2$, write $R^S_i(A) := R^S(A; \mathbf{X}_i, Y_i)$. For all sufficiently large n_2 such that $k_{n_2} = \lceil (n_2+1)(1-\alpha) \rceil \le n_2$, let

$$\widehat{q}_S(A; \mathcal{D}_2) := R^S_{(k_{n_2})}(A).$$

The corresponding CQR and adaptive prediction intervals are

$$\begin{aligned} C^{\mathrm{CQR}}_\alpha(\mathbf{x}; A, \mathcal{D}_2) &:= \left[\ell(\mathbf{x}; A) - \widehat{q}_{\mathrm{CQR}}(A; \mathcal{D}_2),\, u(\mathbf{x}; A) + \widehat{q}_{\mathrm{CQR}}(A; \mathcal{D}_2)\right], \\ C^{\mathrm{Adaptive}}_\alpha(\mathbf{x}; A, \mathcal{D}_2) &:= \left[\mu(\mathbf{x}; A) - \sigma(\mathbf{x}; A)\widehat{q}_{\mathrm{Adaptive}}(A; \mathcal{D}_2),\right. \\ &\qquad \left. \mu(\mathbf{x}; A) + \sigma(\mathbf{x}; A)\widehat{q}_{\mathrm{Adaptive}}(A; \mathcal{D}_2)\right]. \end{aligned}$$

An interval whose lower endpoint exceeds its upper endpoint is interpreted as the empty set. Thus the CQR interval has length $[u(\mathbf{x}; A) - \ell(\mathbf{x}; A) + 2\widehat{q}_{\mathrm{CQR}}(A; \mathcal{D}_2)]_+$, where $[z]_+ := \max\{z, 0\}$, whereas the adaptive interval has length $2\sigma(\mathbf{x}; A)\widehat{q}_{\mathrm{Adaptive}}(A; \mathcal{D}_2)$. The fitted intervals use $A = \mathcal{D}_1$. Their clean oracle counterparts use $A = P_0$ and replace each empirical quantile $\widehat{q}_S(A; \mathcal{D}_2)$ by $q_S(P_0; P_0)$.

For the absolute score, this score-indexed notation agrees with the notation introduced in the preliminaries:

$$\begin{aligned} R^{\mathrm{Absolute}}(A; \mathbf{x}, y) &= R_A(\mathbf{x}, y), \qquad F^{\mathrm{Absolute}}_{P_0,P} = F_{P_0,P}, \\ q_{\mathrm{Absolute}}(P_0; P) &= q(P_0; P), \qquad \widehat{q}_{\mathrm{Absolute}}(A; \mathcal{D}_2) = \widehat{q}(A; \mathcal{D}_2). \end{aligned}$$

The adaptive score intentionally uses the same location learner $\mu(\cdot; A)$ as the absolute score and differs through its scale normalization by $\sigma(\cdot; A)$.

Define the clean-sampling events

$$\begin{aligned} \mathcal{E}^{\mathrm{Absolute}}_{0,n_1} &:= \left\{\|\mu(\cdot; \mathcal{D}_0) - \mu(\cdot; P_0)\|_\infty \le \eta_{n_1}(P_0)\right\}, \\ \mathcal{E}^{\mathrm{CQR}}_{0,n_1} &:= \left\{\max_{g\in\{\ell,u\}} \|g(\cdot; \mathcal{D}_0) - g(\cdot; P_0)\|_\infty \le \eta^{\mathrm{CQR}}_{n_1}(P_0)\right\}, \\ \mathcal{E}^{\mathrm{Adaptive}}_{0,n_1} &:= \bigcap_{g\in\{\mu,\sigma\}} \left\{\|g(\cdot; \mathcal{D}_0) - g(\cdot; P_0)\|_\infty \le \eta^{\mathrm{Adaptive}}_{g,n_1}(P_0)\right\}. \end{aligned}$$

Assume

$$\mathbb{P}(\mathcal{E}^S_{0,n_1}) \ge 1 - \rho^S_{n_1}(P_0), \qquad S \in \{\mathrm{Absolute}, \mathrm{CQR}, \mathrm{Adaptive}\}.$$

For the absolute score, use the conventions $\eta^{\mathrm{Absolute}}_{n_1}(P_0) := \eta_{n_1}(P_0)$ and $\rho^{\mathrm{Absolute}}_{n_1}(P_0) := \rho_{n_1}(P_0)$. Because the adaptive score uses the same location learner, also set $\eta^{\mathrm{Adaptive}}_{\mu,n_1}(P_0) := \eta_{n_1}(P_0)$ and $L^{\mathrm{Adaptive}}_\mu := L$. For the absolute score, suppose that μ is L-uniformly replacement

stable as in Definition 4.1. For the CQR and adaptive scores, assume componentwise replacement stability: whenever $\mathcal{D}, \mathcal{D}'$ differ in at most k observations,

$$\max_{g\in\{\ell,u\}} \|g(\cdot;\mathcal{D}) - g(\cdot;\mathcal{D}')\|_\infty \le L_{\mathrm{CQR}} k/n_1,$$
$$\|g(\cdot;\mathcal{D}) - g(\cdot;\mathcal{D}')\|_\infty \le L_g^{\mathrm{Adaptive}} k/n_1, \qquad g \in \{\mu,\sigma\}. \tag{20}$$

Set $h_{n_1} := \delta_1 + \sqrt{\log(n_1)/n_1}$ and define

$$\varepsilon^{\mathrm{Absolute}}_{\mathrm{train},n_1} := \eta_{n_1}(P_0) + L h_{n_1},$$
$$\varepsilon^{\mathrm{CQR}}_{\mathrm{train},n_1} := \eta^{\mathrm{CQR}}_{n_1}(P_0) + L_{\mathrm{CQR}} h_{n_1},$$
$$\varepsilon^{\mathrm{Adaptive}}_{\mu,n_1} := \eta_{n_1}(P_0) + L h_{n_1},$$
$$\varepsilon^{\mathrm{Adaptive}}_{\sigma,n_1} := \eta^{\mathrm{Adaptive}}_{\sigma,n_1}(P_0) + L^{\mathrm{Adaptive}}_\sigma h_{n_1}.$$

Define the contaminated-training events

$$\mathcal{E}^{\mathrm{Absolute}}_{n_1} := \left\{\|\mu(\cdot;\mathcal{D}_1) - \mu(\cdot;P_0)\|_\infty \le \varepsilon^{\mathrm{Absolute}}_{\mathrm{train},n_1}\right\},$$
$$\mathcal{E}^{\mathrm{CQR}}_{n_1} := \left\{\max_{g\in\{\ell,u\}} \|g(\cdot;\mathcal{D}_1) - g(\cdot;P_0)\|_\infty \le \varepsilon^{\mathrm{CQR}}_{\mathrm{train},n_1}\right\},$$
$$\mathcal{E}^{\mathrm{Adaptive}}_{n_1} := \bigcap_{g\in\{\mu,\sigma\}} \left\{\|g(\cdot;\mathcal{D}_1) - g(\cdot;P_0)\|_\infty \le \varepsilon^{\mathrm{Adaptive}}_{g,n_1}\right\}.$$

Under the Huber coupling used in Proposition B.1, the coupled samples $\mathcal{D}_1$ and $\mathcal{D}_0$ differ in at most $K \sim \mathrm{Binomial}(n_1, \delta_1)$ observations. On $\{K/n_1 \le h_{n_1}\}$, the componentwise bounds in (20), combined with the corresponding clean-sampling event $\mathcal{E}^S_{0,n_1}$, imply $\mathcal{E}^S_{n_1}$. Hoeffding's inequality and a union bound therefore give, uniformly over Q_1,

$$\mathbb{P}(\mathcal{E}^S_{n_1}) \ge 1 - \rho^S_{n_1}(P_0) - n_1^{-2}. \tag{21}$$

Assume that all clean-sampling radii and failure probabilities above are $o(1)$.

For the adaptive score, assume that, on $\mathcal{E}^{\mathrm{Adaptive}}_{n_1}$, the fitted and clean population scales obey the global lower bound

$$\inf_{\mathbf{x}\in\mathcal{X}} \min\{\sigma(\mathbf{x};\mathcal{D}_1), \sigma(\mathbf{x};P_0)\} \ge \underline{\sigma} > 0.$$

This condition can, for example, be enforced by flooring the output of the scale learner at $\underline{\sigma}$. Because it holds on all of $\mathcal{X}$, it does not restrict the support of Q_2. For the uniform interval-length conclusion only, additionally define and assume finite the global clean-scale upper bound

$$\overline{\sigma}(P_0) := \sup_{\mathbf{x}\in\mathcal{X}} \sigma(\mathbf{x};P_0) < \infty, \qquad \overline{\sigma} := \overline{\sigma}(P_0).$$

No bound on the clean or contaminating residuals is imposed. The adaptive argument below controls the relevant order statistic directly rather than the largest pointwise score perturbation.

For the local-deviation bounds below, suppose $\alpha \in (0, 1/2)$ and $0 \le \delta_2 < \alpha$, and set

$$p_- := \frac{1-\alpha-\delta_2}{1-\delta_2}, \qquad p_+ := \frac{1-\alpha}{1-\delta_2}.$$

For each score S, assume that $F^S_{P_0,P_0}$ is absolutely continuous on an open interval containing

$$\left[(F^S_{P_0,P_0})^{-1}(p_-), (F^S_{P_0,P_0})^{-1}(p_+)\right]$$

and that its density satisfies $f^S_{P_0,P_0}(t) \ge r^S_0 > 0$ throughout this interval, with the convention $r^{\mathrm{Absolute}}_0 := r_0$.

B.6.2 Score-Quantile Deviation Bounds

For the adaptive score, define

$$\kappa^{\mathrm{Adaptive}}_{\sigma,n_1} := \frac{\varepsilon^{\mathrm{Adaptive}}_{\sigma,n_1}}{\underline{\sigma}}, \qquad \beta^{\mathrm{Adaptive}}_{\mu,n_1} := \frac{\varepsilon^{\mathrm{Adaptive}}_{\mu,n_1}}{\underline{\sigma}},$$
$$\varepsilon^{\mathrm{Adaptive}}_{\mathrm{train},n_1} := \beta^{\mathrm{Adaptive}}_{\mu,n_1} + \kappa^{\mathrm{Adaptive}}_{\sigma,n_1} q_{\mathrm{Adaptive}}(P_0;P_0).$$

For the proofs below, define the score-specific calibration errors by

$$\varepsilon^S_{\mathrm{cal},n_2}(\gamma) := \frac{\delta_2(1-\alpha) + \sqrt{\log(2/\gamma)/(2n_2)} + 2/n_2}{(1-\delta_2)r_0^S}, \qquad S \in \{\mathrm{Absolute}, \mathrm{CQR}, \mathrm{Adaptive}\}.$$

Define the total score-quantile errors by

$$\varepsilon^S_{q,n_1,n_2}(\gamma) := \begin{cases} \varepsilon^S_{\mathrm{train},n_1} + \varepsilon^S_{\mathrm{cal},n_2}(\gamma), & S \in \{\mathrm{Absolute}, \mathrm{CQR}\}, \\ \varepsilon^{\mathrm{Adaptive}}_{\mathrm{train},n_1} + \left(1 + \kappa^{\mathrm{Adaptive}}_{\sigma,n_1}\right)\varepsilon^{\mathrm{Adaptive}}_{\mathrm{cal},n_2}(\gamma), & S = \mathrm{Adaptive}. \end{cases} \tag{22}$$

Lemma B.2 (Adaptive empirical-quantile perturbation). *On $\mathcal{E}^{\mathrm{Adaptive}}_{n_1}$, for every calibration sample $\mathcal{D}_2$,*

$$|\widehat{q}_{\mathrm{Adaptive}}(\mathcal{D}_1;\mathcal{D}_2) - \widehat{q}_{\mathrm{Adaptive}}(P_0;\mathcal{D}_2)| \le \beta^{\mathrm{Adaptive}}_{\mu,n_1} + \kappa^{\mathrm{Adaptive}}_{\sigma,n_1}\widehat{q}_{\mathrm{Adaptive}}(P_0;\mathcal{D}_2). \tag{23}$$

In particular, the bound depends on the relevant empirical quantile rather than on the largest calibration residual.

Proof. Fix a calibration observation $(\mathbf{x}, y)$ and write

$$\begin{aligned} r_{\mathrm{fit}} &:= |y - \mu(\mathbf{x};\mathcal{D}_1)|, & r_{\mathrm{clean}} &:= |y - \mu(\mathbf{x};P_0)|, \\ \sigma_{\mathrm{fit}} &:= \sigma(\mathbf{x};\mathcal{D}_1), & \sigma_{\mathrm{clean}} &:= \sigma(\mathbf{x};P_0), \\ \varepsilon_\mu &:= \varepsilon^{\mathrm{Adaptive}}_{\mu,n_1}, & \varepsilon_\sigma &:= \varepsilon^{\mathrm{Adaptive}}_{\sigma,n_1}. \end{aligned}$$

On $\mathcal{E}^{\mathrm{Adaptive}}_{n_1}$,

$$|r_{\mathrm{fit}} - r_{\mathrm{clean}}| \le \varepsilon_\mu, \quad |\sigma_{\mathrm{fit}} - \sigma_{\mathrm{clean}}| \le \varepsilon_\sigma, \quad \sigma_{\mathrm{fit}}, \sigma_{\mathrm{clean}} \ge \underline{\sigma}.$$

Set $\kappa := \varepsilon_\sigma/\underline{\sigma}$ and $\beta := \varepsilon_\mu/\underline{\sigma}$. Since $\sigma_{\mathrm{clean}} \le \sigma_{\mathrm{fit}} + \varepsilon_\sigma$,

$$\begin{aligned} R^{\mathrm{Adaptive}}(\mathcal{D}_1;\mathbf{x},y) &= \frac{r_{\mathrm{fit}}}{\sigma_{\mathrm{fit}}} \\ &\le \frac{r_{\mathrm{clean}} + \varepsilon_\mu}{\sigma_{\mathrm{fit}}} \\ &\le (1+\kappa)R^{\mathrm{Adaptive}}(P_0;\mathbf{x},y) + \beta. \end{aligned}$$

Interchanging the fitted and clean quantities gives the reverse comparison

$$R^{\mathrm{Adaptive}}(P_0;\mathbf{x},y) \le (1+\kappa)R^{\mathrm{Adaptive}}(\mathcal{D}_1;\mathbf{x},y) + \beta.$$

Both comparisons hold for every $(\mathbf{x}, y)$, without bounding the residual. Monotonicity of order statistics therefore yields, with $\widehat{q}_{\mathrm{fit}} := \widehat{q}_{\mathrm{Adaptive}}(\mathcal{D}_1;\mathcal{D}_2)$ and $\widehat{q}_{\mathrm{clean}} := \widehat{q}_{\mathrm{Adaptive}}(P_0;\mathcal{D}_2)$,

$$\widehat{q}_{\mathrm{fit}} \le (1+\kappa)\widehat{q}_{\mathrm{clean}} + \beta, \qquad \widehat{q}_{\mathrm{clean}} \le (1+\kappa)\widehat{q}_{\mathrm{fit}} + \beta.$$

If $\widehat{q}_{\mathrm{fit}} \ge \widehat{q}_{\mathrm{clean}}$, the first inequality gives a difference of at most $\kappa\widehat{q}_{\mathrm{clean}} + \beta$. Otherwise, the second gives

$$\widehat{q}_{\mathrm{clean}} - \widehat{q}_{\mathrm{fit}} \le \frac{\kappa\widehat{q}_{\mathrm{clean}} + \beta}{1+\kappa} \le \kappa\widehat{q}_{\mathrm{clean}} + \beta.$$

This proves (23). □

Proposition B.2 (Unified score-quantile deviation bounds). *For every $\gamma \in (0,1)$ and all sufficiently large n_2, each $S \in \{\mathrm{Absolute}, \mathrm{CQR}, \mathrm{Adaptive}\}$ satisfies, with probability at least $1 - \rho^S_{n_1}(P_0) - n_1^{-2} - \gamma$,*

$$\begin{aligned} &|\widehat{q}_S(\mathcal{D}_1;\mathcal{D}_2) - q_S(P_0;P_0)| \\ &\quad \le \varepsilon^S_{q,n_1,n_2}(\gamma). \end{aligned} \tag{24}$$

The probability statement is uniform over the contamination distributions Q_1, Q_2. Consequently, if $\rho^S_{n_1}(P_0) = o(1)$, then, for $S \in \{\mathrm{Absolute}, \mathrm{CQR}\}$,

$$|\widehat{q}_S(\mathcal{D}_1;\mathcal{D}_2) - q_S(P_0;P_0)| = O_{\mathbb{P}}\left(\varepsilon^S_{\mathrm{train},n_1} + \frac{n_2^{-1/2} + \delta_2(1-\alpha)}{(1-\delta_2)r_0^S}\right).$$

For the adaptive score, the corresponding statement is

$$\begin{aligned}|\widehat{q}_{\mathrm{Adaptive}}(\mathcal{D}_1;\mathcal{D}_2) - q_{\mathrm{Adaptive}}(P_0;P_0)| \\ = O_{\mathbb{P}}\Big(\beta_{\mu,n_1}^{\mathrm{Adaptive}} + \kappa_{\sigma,n_1}^{\mathrm{Adaptive}} q_{\mathrm{Adaptive}}(P_0;P_0) \\ + \big(1+\kappa_{\sigma,n_1}^{\mathrm{Adaptive}}\big)\frac{n_2^{-1/2}+\delta_2(1-\alpha)}{(1-\delta_2)r_0^{\mathrm{Adaptive}}}\Big).\end{aligned}$$

Proof. We first bound the training-induced score perturbation. For the absolute score, the reverse triangle inequality gives, on $\mathcal{E}_{n_1}^{\mathrm{Absolute}}$,

$$\begin{aligned}\big|R_i^{\mathrm{Absolute}}(\mathcal{D}_1) - R_i^{\mathrm{Absolute}}(P_0)\big| &\le |\mu(\mathbf{X}_i;\mathcal{D}_1) - \mu(\mathbf{X}_i;P_0)| \\ &\le \varepsilon_{\mathrm{train},n_1}^{\mathrm{Absolute}}.\end{aligned}$$

For CQR, the triangle inequality first gives, on $\mathcal{E}_{n_1}^{\mathrm{CQR}}$,

$$\max_{g\in\{\ell,u\}} \|g(\cdot;\mathcal{D}_1) - g(\cdot;P_0)\|_\infty \le \varepsilon_{\mathrm{train},n_1}^{\mathrm{CQR}}.$$

The maximum of two coordinates is one-Lipschitz under the maximum norm, so

$$\begin{aligned}&\Big|R_i^{\mathrm{CQR}}(\mathcal{D}_1) - R_i^{\mathrm{CQR}}(P_0)\Big| \\ &\le \max\{|\ell(\mathbf{X}_i;\mathcal{D}_1) - \ell(\mathbf{X}_i;P_0)|, |u(\mathbf{X}_i;\mathcal{D}_1) - u(\mathbf{X}_i;P_0)|\} \\ &\le \varepsilon_{\mathrm{train},n_1}^{\mathrm{CQR}}.\end{aligned}$$

Thus, for $S \in \{\mathrm{Absolute}, \mathrm{CQR}\}$, Lemma B.1 implies

$$|\widehat{q}_S(\mathcal{D}_1;\mathcal{D}_2) - \widehat{q}_S(P_0;\mathcal{D}_2)| \le \varepsilon_{\mathrm{train},n_1}^{S} \tag{25}$$

on $\mathcal{E}_{n_1}^S$.

For the calibration contribution, fixing the score at its clean population value gives

$$F_{P_0,P_2}^S(t) = (1-\delta_2)F_{P_0,P_0}^S(t) + \delta_2 F_{P_0,Q_2}^S(t).$$

The quantile-sandwich argument of Lemma 4.4 yields

$$|q_S(P_0;P_2) - q_S(P_0;P_0)| \le \frac{\delta_2(1-\alpha)}{(1-\delta_2)r_0^S}. \tag{26}$$

The Dvoretzky–Kiefer–Wolfowitz inequality and the conformal rank correction further give, with probability at least $1-\gamma$,

$$|\widehat{q}_S(P_0;\mathcal{D}_2) - q_S(P_0;P_2)| \le \frac{\sqrt{\log(2/\gamma)/(2n_2)} + 2/n_2}{(1-\delta_2)r_0^S}. \tag{27}$$

For $S \in \{\mathrm{Absolute}, \mathrm{CQR}\}$, combining (25) with the two calibration bounds by the triangle inequality gives the first branch of (22).

For the adaptive score, abbreviate

$$\widehat{q}_0 := \widehat{q}_{\mathrm{Adaptive}}(P_0;\mathcal{D}_2), \qquad q_0 := q_{\mathrm{Adaptive}}(P_0;P_0),$$

and write $\kappa_\sigma := \kappa_{\sigma,n_1}^{\mathrm{Adaptive}}$ and $\beta_\mu := \beta_{\mu,n_1}^{\mathrm{Adaptive}}$. On the calibration event given by (26) and (27),

$$|\widehat{q}_0 - q_0| \le \varepsilon_{\mathrm{cal},n_2}^{\mathrm{Adaptive}}(\gamma), \qquad \widehat{q}_0 \le q_0 + \varepsilon_{\mathrm{cal},n_2}^{\mathrm{Adaptive}}(\gamma).$$

Lemma B.2 and the triangle inequality now give

$$\begin{aligned}&|\widehat{q}_{\mathrm{Adaptive}}(\mathcal{D}_1;\mathcal{D}_2) - q_0| \\ &\le \beta_\mu + \kappa_\sigma \widehat{q}_0 + \varepsilon_{\mathrm{cal},n_2}^{\mathrm{Adaptive}}(\gamma) \\ &\le \beta_\mu + \kappa_\sigma q_0 + (1+\kappa_\sigma)\varepsilon_{\mathrm{cal},n_2}^{\mathrm{Adaptive}}(\gamma),\end{aligned}$$

which is the adaptive branch of (22). The training event is uniform over Q_1 by (21), while the mixture sandwich and DKW inequality are uniform over Q_2. Taking a union bound proves (24). The $O_{\mathbb{P}}$ statements follow because $n_2^{-1} = O(n_2^{-1/2})$. □

B.6.3 Interval-Length Deviation Bounds

The score-quantile bounds yield the following deviations of the fitted interval lengths from their clean oracle counterparts.

Proposition B.3 (Alternative-score interval-length deviation bounds)**.** *Under the score-specific conditions of this subsection, for every $\gamma \in (0,1)$ and all sufficiently large n_1, n_2, the following bounds hold with probability at least $1 - \rho^S_{n_1}(P_0) - n_1^{-2} - \gamma$ for the corresponding score S:*

$$\sup_{\mathbf{x}\in\mathcal{X}} |2\widehat{q}_{\mathrm{Absolute}}(\mathcal{D}_1;\mathcal{D}_2) - 2q_{\mathrm{Absolute}}(P_0;P_0)| \le 2\varepsilon^{\mathrm{Absolute}}_{q,n_1,n_2}(\gamma),$$

$$\sup_{\mathbf{x}\in\mathcal{X}} \Big| [u(\mathbf{x};\mathcal{D}_1) - \ell(\mathbf{x};\mathcal{D}_1) + 2\widehat{q}_{\mathrm{CQR}}(\mathcal{D}_1;\mathcal{D}_2)]_+ - [u(\mathbf{x};P_0) - \ell(\mathbf{x};P_0) + 2q_{\mathrm{CQR}}(P_0;P_0)]_+ \Big| \le 4\varepsilon^{\mathrm{CQR}}_{\mathrm{train},n_1} + 2\varepsilon^{\mathrm{CQR}}_{\mathrm{cal},n_2}(\gamma),$$

$$\sup_{\mathbf{x}\in\mathcal{X}} \Big| 2\sigma(\mathbf{x};\mathcal{D}_1)\widehat{q}_{\mathrm{Adaptive}}(\mathcal{D}_1;\mathcal{D}_2) - 2\sigma(\mathbf{x};P_0)q_{\mathrm{Adaptive}}(P_0;P_0) \Big| \le 2\big(\overline{\sigma} + \varepsilon^{\mathrm{Adaptive}}_{\sigma,n_1}\big)\varepsilon^{\mathrm{Adaptive}}_{q,n_1,n_2}(\gamma) + 2q_{\mathrm{Adaptive}}(P_0;P_0)\varepsilon^{\mathrm{Adaptive}}_{\sigma,n_1}.$$

The absolute-score bound recovers Theorem 4.1. The CQR bound also accounts for perturbations of the fitted lower and upper regressors, while the adaptive bound includes both quantile and scale perturbations.

Proof of Proposition B.3. Fix a score S and intersect its training event $\mathcal{E}^S_{n_1}$ with the calibration event in Proposition B.2. On this event, write

$$E^S_q := \varepsilon^S_{q,n_1,n_2}(\gamma), \qquad |\widehat{q}_S(\mathcal{D}_1;\mathcal{D}_2) - q_S(P_0;P_0)| \le E^S_q.$$

For the absolute score,

$$\sup_{\mathbf{x}\in\mathcal{X}} |2\widehat{q}_{\mathrm{Absolute}}(\mathcal{D}_1;\mathcal{D}_2) - 2q_{\mathrm{Absolute}}(P_0;P_0)| \le 2E^{\mathrm{Absolute}}_q.$$

For CQR, set $b_A(\mathbf{x}) := u(\mathbf{x};A) - \ell(\mathbf{x};A)$. The definition of $\mathcal{E}^{\mathrm{CQR}}_{n_1}$ and the triangle inequality give

$$\sup_{\mathbf{x}\in\mathcal{X}} |b_{\mathcal{D}_1}(\mathbf{x}) - b_{P_0}(\mathbf{x})| \le 2\varepsilon^{\mathrm{CQR}}_{\mathrm{train},n_1}.$$

Because $z \mapsto [z]_+$ is one-Lipschitz,

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}} \Big| [u(\mathbf{x};\mathcal{D}_1) - \ell(\mathbf{x};\mathcal{D}_1) + 2\widehat{q}_{\mathrm{CQR}}(\mathcal{D}_1;\mathcal{D}_2)]_+ \\
&\qquad\qquad - [u(\mathbf{x};P_0) - \ell(\mathbf{x};P_0) + 2q_{\mathrm{CQR}}(P_0;P_0)]_+ \Big| \\
&\le 2\varepsilon^{\mathrm{CQR}}_{\mathrm{train},n_1} + 2E^{\mathrm{CQR}}_q \\
&= 4\varepsilon^{\mathrm{CQR}}_{\mathrm{train},n_1} + 2\varepsilon^{\mathrm{CQR}}_{\mathrm{cal},n_2}(\gamma).
\end{aligned}$$

For the adaptive score, abbreviate $\varepsilon_\sigma := \varepsilon^{\mathrm{Adaptive}}_{\sigma,n_1}$. On $\mathcal{E}^{\mathrm{Adaptive}}_{n_1}$,

$$\sup_{\mathbf{x}\in\mathcal{X}} |\sigma(\mathbf{x};\mathcal{D}_1) - \sigma(\mathbf{x};P_0)| \le \varepsilon_\sigma, \qquad \sup_{\mathbf{x}\in\mathcal{X}} \sigma(\mathbf{x};\mathcal{D}_1) \le \overline{\sigma} + \varepsilon_\sigma.$$

Since adaptive scores and hence $q_{\mathrm{Adaptive}}(P_0;P_0)$ are nonnegative, uniformly over $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&|2\sigma(\mathbf{x};\mathcal{D}_1)\widehat{q}_{\mathrm{Adaptive}}(\mathcal{D}_1;\mathcal{D}_2) - 2\sigma(\mathbf{x};P_0)q_{\mathrm{Adaptive}}(P_0;P_0)| \\
&\le 2\sigma(\mathbf{x};\mathcal{D}_1)E^{\mathrm{Adaptive}}_q + 2q_{\mathrm{Adaptive}}(P_0;P_0)|\sigma(\mathbf{x};\mathcal{D}_1) - \sigma(\mathbf{x};P_0)| \\
&\le 2(\overline{\sigma} + \varepsilon_\sigma)E^{\mathrm{Adaptive}}_q + 2q_{\mathrm{Adaptive}}(P_0;P_0)\varepsilon_\sigma.
\end{aligned}$$

Taking the supremum proves the adaptive row. The common event has probability at least $1 - \rho^S_{n_1}(P_0) - n_1^{-2} - \gamma$ by Proposition B.2. □

B.7 Derivation for Example 4.1

Example 4.1 (ℓ_2-regularized linear quantile regression)**.** *Suppose that* $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_2 \le B_x$. *For* $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, *define the* ℓ_2*-regularized linear quantile regression estimator by*

$$\hat{\omega}(\mathcal{D}) := \operatorname{argmin}_{\omega\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - \omega^\top \mathbf{X}_i) + \frac{\lambda}{2}\|\omega\|_2^2 \right\}, \tag{28}$$

where $\rho_\tau(u) = u\big(\tau - \mathbf{1}\{u < 0\}\big)$ *is the pinball loss,* $\tau \in (0,1)$, *and* $\lambda > 0$. *Then the predictor* $\mu(\mathbf{x}; \mathcal{D}) := \mathbf{x}^\top \hat{\omega}(\mathcal{D})$ *is L-uniformly replacement stable with* $L = 2B_x^2 \max\{\tau, 1-\tau\}/\lambda$.

Proof. Let $\mathcal{D}_1'$ be a training dataset of size n_1 that differs from $\mathcal{D}_1$ in at most k observations, and write $\hat{\omega}_1 := \hat{\omega}(\mathcal{D}_1)$ and $\hat{\omega}_1' := \hat{\omega}(\mathcal{D}_1')$. For each $\mathcal{A} \in \{\mathcal{D}_1, \mathcal{D}_1'\}$, write $J_\mathcal{A}$ for the objective in (28) formed from $\mathcal{A}$. This objective is λ-strongly convex, so the minimizer is unique. For a single observation $(\mathbf{X}_i, Y_i)$ and any ω, every subgradient of $\rho_\tau(Y_i - \omega^\top \mathbf{X}_i)$ with respect to ω has Euclidean norm at most

$$G := B_x \max\{\tau, 1-\tau\}.$$

This bound does not require bounded responses.

Set

$$I := \{i : (\mathbf{X}_i, Y_i) \neq (\mathbf{X}_i', Y_i')\}, \qquad |I| \le k.$$

Because $0 \in \partial J_{\mathcal{D}_1'}(\hat{\omega}_1')$, choose subgradients $\xi_i' \in \partial_\omega \rho_\tau(Y_i' - (\hat{\omega}_1')^\top \mathbf{X}_i')$ such that

$$\frac{1}{n_1}\sum_{i=1}^{n_1} \xi_i' + \lambda \hat{\omega}_1' = 0.$$

For every $i \notin I$, the two observations are identical, and we choose the corresponding subgradient for $J_{\mathcal{D}_1}$ to be $\xi_i = \xi_i'$. For $i \in I$, choose any $\xi_i \in \partial_\omega \rho_\tau(Y_i - (\hat{\omega}_1')^\top \mathbf{X}_i)$. Then

$$g := \frac{1}{n_1}\sum_{i=1}^{n_1} \xi_i + \lambda \hat{\omega}_1' \in \partial J_{\mathcal{D}_1}(\hat{\omega}_1')$$

satisfies

$$\|g\|_2 = \left\| \frac{1}{n_1} \sum_{i\in I} (\xi_i - \xi_i') \right\|_2 \le \frac{2Gk}{n_1}.$$

Since $0 \in \partial J_{\mathcal{D}_1}(\hat{\omega}_1)$ and $J_{\mathcal{D}_1}$ is λ-strongly convex, its subdifferential is λ-strongly monotone:

$$\lambda \|\hat{\omega}_1' - \hat{\omega}_1\|_2^2 \le \langle g, \hat{\omega}_1' - \hat{\omega}_1 \rangle \le \|g\|_2 \, \|\hat{\omega}_1' - \hat{\omega}_1\|_2.$$

If $\hat{\omega}_1' = \hat{\omega}_1$, there is nothing to prove. Otherwise, division by $\|\hat{\omega}_1' - \hat{\omega}_1\|_2$ gives

$$\|\hat{\omega}_1' - \hat{\omega}_1\|_2 \le \frac{2G}{\lambda}\frac{k}{n_1}.$$

Therefore,

$$\sup_{\|\mathbf{X}\|_2 \le B_x} \big|\mathbf{X}^\top(\hat{\omega}_1' - \hat{\omega}_1)\big| \le B_x \|\hat{\omega}_1' - \hat{\omega}_1\|_2 \le \frac{2B_x G}{\lambda}\frac{k}{n_1},$$

which is the claimed bound. □

B.8 Formal Joint Minimax Lower Bound and Proof

Lower-bound model. The following assumptions are used only for the lower bound; they are not consequences of sampling or replacement stability. To distinguish the density scale of this least-favorable submodel from the upper-bound constant r_0, we denote it by $r_\star$. The two quantities are related only through the constant-factor comparison used in the informal rate discussion in Section 4.2.

Assumption B.1 (Regular location submodel)**.** *There exist constants $L_\star, h_0, s_0, r_\star, \kappa > 0$ and $C_{\mathrm{KL}}, C_{\mathrm{TV}}, C_f < \infty$ with the following properties. Let $\Theta = [-h_0/2, h_0/2]$. The model contains the intercept-only location family*

$$\mathbf{X} = 0, \qquad Y = \theta + \varepsilon, \qquad \theta \in \Theta,$$

where ε has a continuously differentiable density g satisfying the conditions below; denote the law of $(\mathbf{X}, Y)$ by P_θ. On this submodel, the clean population predictor is $\mu(0; P_\theta) = \theta$. Here θ is the population target defining the lower-bound submodel; the proof does not require it to coincide with the conditional mean of Y under the asymmetric noise distribution used below. Write

$$F_h(t) := \mathbb{P}(|\varepsilon - h| \le t), \qquad q_g(h) := F_h^{-1}(1-\alpha), \qquad q_\star := q_g(0).$$

Thus $q(P_\theta; P_\theta) = q_\star$ for every $\theta \in \Theta$. For $t > 0$, let

$$f_h(t) := F_h'(t) = g(h+t) + g(h-t)$$

denote the density of the absolute-residual score $|\varepsilon - h|$. For every $\theta, \theta' \in \Theta$,

$$\mathrm{KL}(P_\theta, P_{\theta'}) \le C_{\mathrm{KL}} \frac{(\theta - \theta')^2}{L_\star^2}, \tag{29}$$

$$\mathrm{TV}(P_\theta, P_{\theta'}) \le C_{\mathrm{TV}} \frac{|\theta - \theta'|}{L_\star}. \tag{30}$$

For every $|h| \le h_0$, the score density f_h satisfies

$$r_\star \le f_h(t) \le C_f r_\star \quad \textit{whenever} \quad |t - q_g(h)| \le 2s_0, \tag{31}$$

and the population score quantile is locally identifiable:

$$|q_g(h) - q_\star| \ge \kappa |h|. \tag{32}$$

Finally, there exist $z_-, z_+ \in \mathbb{R}$ such that, for all $|h| \le h_0$,

$$|z_- - h| \le q_g(h) - 2s_0, \qquad |z_+ - h| \ge q_g(h) + 2s_0. \tag{33}$$

The assumption is non-vacuous. For example, fix α and start from a $\mathcal{N}(m, 1)$ density with $m \neq 0$. At the $(1-\alpha)$-quantile $q_{\mathrm{base}} > 0$ of its absolute value, the densities at q_{base} and $-q_{\mathrm{base}}$ are unequal. Adding a sufficiently small Gaussian component with a different mean and variance preserves this inequality at the perturbed distribution's own absolute-value quantile and produces an asymmetric two-Gaussian mixture $g_\star$. For $\sigma > 0$, the scale family $g_\sigma(x) = \sigma^{-1} g_\star(x/\sigma)$ supplies examples for the assumption: taking $g = g_\sigma$, the local KL and TV bounds hold with $L_\star$ comparable to σ up to constant factors. On compact shift intervals, for some finite constant C_g,

$$\mathrm{KL}(P_\theta, P_{\theta'}) \le C_g \frac{(\theta - \theta')^2}{\sigma^2}, \qquad \mathrm{TV}(P_\theta, P_{\theta'}) \le \frac{\|g_\star'\|_1 |\theta - \theta'|}{2\sigma}.$$

Here $\|g_\star'\|_1 := \int_{\mathbb{R}} |g_\star'(x)| \, \mathrm{d}x$. Moreover, implicit differentiation of $F_h(q_g(h)) = 1 - \alpha$ gives

$$q_g'(0) = \frac{g(-q_\star) - g(q_\star)}{g(-q_\star) + g(q_\star)} \neq 0.$$

Continuity yields (31) and (32) after choosing h_0, s_0 small enough. Taking $h_0 + 2s_0 < \inf_{|h| \le h_0} q_g(h)$ permits $z_- = 0$, and any sufficiently large $|z_+|$ supplies the high-score point.

Formal criterion and theorem. Let $\mathcal{E}_{n_1}(\Theta)$ be the class of measurable, possibly randomized estimators $\widehat{\theta} : \mathcal{D}_1 \mapsto \Theta$, and write $\widehat{\theta}_1 := \widehat{\theta}(\mathcal{D}_1)$. For $u \in \Theta$, let $\widehat{q}(u; \mathcal{D}_2)$ denote the conformal calibration quantile based on the scores $|Y_i - u|$. Fix $\delta_1, \delta_2 \in [0, 1)$. The contamination distributions Q_1, Q_2 range over all probability measures on the observation space, and the training and calibration samples are independent with laws

$$\mathcal{D}_1 \sim \big((1-\delta_1)P_\theta + \delta_1 Q_1\big)^{n_1}, \qquad \mathcal{D}_2 \sim \big((1-\delta_2)P_\theta + \delta_2 Q_2\big)^{n_2}.$$

$\mathbb{P}_{\theta, Q_1, Q_2}$ denotes probability over this joint law and any independent internal randomization of the estimator. Define the minimax tail-probability profile

$$\mathfrak{p}^{\mathrm{split}}_{n_1, n_2}(t) := \inf_{\widehat{\theta} \in \mathcal{E}_{n_1}(\Theta)} \sup_{\theta \in \Theta} \sup_{Q_1, Q_2} \mathbb{P}_{\theta, Q_1, Q_2} \left(\left| 2\widehat{q}(\widehat{\theta}_1; \mathcal{D}_2) - 2q_\star \right| \ge t \right).$$

Also define the training- and calibration-side difficulty scales

$$a_{n_1} := \left\{ L_\star\big(n_1^{-1/2} + \delta_1\big) \right\} \wedge h_0, \qquad b_{n_2} := \left\{ \frac{n_2^{-1/2} + \delta_2}{r_\star} \right\} \wedge s_0, \tag{34}$$

where $x \wedge y := \min\{x, y\}$. The truncations restrict the construction to the local regime of Assumption B.1.

Theorem B.1 (Joint minimax lower bound for interval-length deviation)**.** *Under Assumption B.1, there exist constants $c, c_0 > 0$ and $n_0 < \infty$, depending only on α and the fixed lower-bound submodel constants, such that, for all $n_1, n_2 \geq n_0$,*

$$\mathfrak{p}^{\mathrm{split}}_{n_1,n_2}\left(c\big(a_{n_1} + b_{n_2}\big)\right) \geq c_0. \tag{35}$$

When the truncations are inactive, the theorem gives the lower-bound scale

$$L_\star\big(n_1^{-1/2} + \delta_1\big) + \frac{n_2^{-1/2} + \delta_2}{r_\star}.$$

The theorem's constants may depend on the fixed submodel, so the displayed inverse dependence on $r_\star$ is not a uniform claim over a family in which $r_\star$ varies.

Criterion and scope. When $\theta = 0$, we write the corresponding clean law explicitly as $P_{\theta=0}$. A probability criterion is essential here. If $\delta_2 > 0$ and Q_2 places its mass at an arbitrarily large score M, the event that all n_2 calibration draws come from the contaminating component has probability $\delta_2^{n_2} > 0$ and makes the conformal radius arbitrarily large. Hence the supremum over Q_2 of the untruncated expected interval-length loss is infinite, whereas the fixed-probability profile is non-degenerate and comparable to the fixed-confidence upper bound.

The restriction $\widehat{\theta}_1 \in \Theta$ is part of the estimator class in the theorem. It guarantees that $|\widehat{\theta}_1 - \theta| \leq h_0$, so the local transversality and score-density conditions apply uniformly throughout the proof. Without an additional global condition on q_g or a result showing that projection onto Θ cannot increase the interval-length loss, the theorem does not extend to arbitrary real-valued constant predictors.

The four terms in (34) arise from clean training-sample indistinguishability, overlap of training Huber neighborhoods, fluctuation of the calibration order statistic, and separation of low- and high-score calibration contaminations. The supremum may select a different least-favorable distribution for each component; the theorem does not claim one distribution simultaneously realizes all four mechanisms. It is also restricted to the standard split architecture. Finally, $L_\star$ is a local information scale, whereas L in Definition 4.1 is an algorithmic stability constant. Thus the theorem establishes the structural orders, not universal optimality of the algorithm-dependent constant L. Likewise, $r_\star$ is the local score-density scale of the lower-bound submodel and need not equal the upper-bound constant r_0 outside a common model class.

A rate comparison must use the same confidence level on both sides. Fix a failure probability $\beta < c_0$, where c_0 is the lower-bound probability in Theorem B.1. On the same local model, require

$$\eta_{n_1}(P_\theta; \beta/3) = O_\beta(L_\star n_1^{-1/2}) \quad \text{uniformly over } \theta \in \Theta, \qquad L \asymp L_\star, \quad r_0 \asymp r_\star.$$

Here $\eta_{n_1}(P_\theta; \beta/3)$ is the fixed-confidence training radius defined in Appendix B.9; it is distinct from the radius paired with a vanishing failure probability in Assumption 3.1. The density interval must have a uniform margin, and the local truncations in the lower bound must be inactive. The same appendix proves the corresponding upper bound and gives a learner meeting these conditions on the fixed location submodel. Without these conditions, the lower bound identifies four structural components rather than a universal match of the algorithm-dependent upper bound.

We first record the elementary overlap fact used for the two contamination constructions.

Lemma B.3 (Intersection of two Huber neighborhoods)**.** *Let P and P' be two probability measures and let $\delta \in (0, 1)$. If*

$$\mathrm{TV}(P, P') \leq \frac{\delta}{1 - \delta}, \tag{36}$$

then there exist probability measures Q and Q' such that

$$(1 - \delta)P + \delta Q = (1 - \delta)P' + \delta Q'.$$

Proof. Let p and p' be densities of P and P' with respect to a common dominating measure. Since

$$\int \max\{p, p'\} = 1 + \mathrm{TV}(P, P'),$$

condition (36) implies

$$(1-\delta)\int \max\{p, p'\} \le 1.$$

Choose any probability density r, set

$$\zeta := 1 - (1-\delta)\int \max\{p, p'\} \ge 0,$$

and define

$$m := (1-\delta)\max\{p, p'\} + \zeta r.$$

Then m is a probability density and $m \ge (1-\delta)p, (1-\delta)p'$ pointwise. Therefore

$$q := \frac{m-(1-\delta)p}{\delta}, \qquad q' := \frac{m-(1-\delta)p'}{\delta}$$

are probability densities and satisfy

$$(1-\delta)p + \delta q = m = (1-\delta)p' + \delta q'.$$

□

Lemma B.4 (Uniform anti-concentration of the conformal rank)**.** *Let $p = 1-\alpha$ and $k_n = \lceil (n+1)p \rceil$. For every n such that $k_n \le n$, let*

$$B_n \sim \mathrm{Beta}(k_n, n+1-k_n).$$

Denote the density of B_n by f_{B_n}. For fixed $\alpha \in (0,1)$, there are constants $c_B, C_B > 0$ and $n_B < \infty$ such that, for every $n \ge n_B$,

$$\sup_{x\in[0,1]} f_{B_n}(x) \le C_B\sqrt{n}, \tag{37}$$

$$\inf_{u\in[0,1]} \mathbb{P}\left(|B_n - u| \ge \frac{c_B}{\sqrt{n}}\right) \ge \frac{1}{2}. \tag{38}$$

Moreover, for every fixed $\epsilon > 0$,

$$\mathbb{P}\left(|B_n - p| \le \epsilon\right) \ge \frac{3}{4} \tag{39}$$

for all sufficiently large n.

Proof. Increase n_B, if necessary, so that $k_n \le n$ for $n \ge n_B$. Because $k_n/(n+1) \to p \in (0,1)$, both beta shape parameters are bounded below by a fixed positive multiple of n. Evaluating the beta density f_{B_n} at its mode and applying Stirling's inequalities gives (37). Hence, uniformly in $u \in [0,1]$,

$$\mathbb{P}\left(|B_n - u| < \frac{c_B}{\sqrt{n}}\right) \le 2C_B c_B.$$

Taking $c_B \le (4C_B)^{-1}$ proves (38). Finally,

$$\mathbb{E}B_n = \frac{k_n}{n+1} = p + O(n^{-1}), \qquad \mathrm{Var}(B_n) \le \frac{1}{4(n+2)}.$$

Equation (39) follows from Chebyshev's inequality after increasing n_B. □

Proof of Theorem B.1. Fix an arbitrary $\widehat{\theta} \in \mathcal{E}_{n_1}(\Theta)$, set $\widehat{\theta}_1 := \widehat{\theta}(\mathcal{D}_1)$, and write

$$H := \widehat{\theta}_1 - \theta, \qquad \widehat{q}_H := \widehat{q}(\widehat{\theta}_1; \mathcal{D}_2).$$

Throughout the proof, $c > 0$ denotes a generic constant that may decrease from one display to the next and depends only on α and the constants in Assumption B.1. Every constant carrying a descriptive subscript is fixed when introduced. For a randomized estimator, adjoin its independent random

seed to both experiments below; this leaves every KL and TV comparison unchanged. Because both θ and $\widehat{\theta}_1$ belong to Θ, we always have $|H| \le h_0$. We prove training and calibration lower bounds separately and then combine them. It is useful to write

$$u_{n_2} := \left(\frac{1}{r_\star\sqrt{n_2}}\right) \wedge s_0, \qquad v_{n_2} := \left(\frac{\delta_2}{r_\star}\right) \wedge s_0.$$

Elementary truncation inequalities give

$$\max\{u_{n_2}, v_{n_2}\} \le b_{n_2} \le u_{n_2} + v_{n_2} \le 2\max\{u_{n_2}, v_{n_2}\}. \tag{40}$$

Step 1: clean training-sample difficulty. Choose a sufficiently small constant $c_1 > 0$ and let

$$d_1 := c_1\left(\frac{L_\star}{\sqrt{n_1}} \wedge h_0\right)$$

and take $\theta_- = -d_1/2$ and $\theta_+ = d_1/2$. For this part choose $Q_{1,\pm} = Q_{2,\pm} = P_{\theta_\pm}$, so the observed training laws are the clean laws $P_{\theta_\pm}^{n_1}$. By (29),

$$\mathrm{KL}\left(P_{\theta_-}^{n_1}, P_{\theta_+}^{n_1}\right) \le n_1 C_{\mathrm{KL}} \frac{d_1^2}{L_\star^2}.$$

Choosing c_1 small enough and applying Pinsker's inequality gives

$$\mathrm{TV}\left(P_{\theta_-}^{n_1}, P_{\theta_+}^{n_1}\right) \le \frac{1}{4}.$$

For any $z \in \Theta$, transversality and the triangle inequality imply

$$\begin{aligned} &|q_g(z-\theta_-) - q_\star| + |q_g(z-\theta_+) - q_\star| \\ &\quad \ge \kappa\left(|z-\theta_-| + |z-\theta_+|\right) \ge \kappa d_1. \end{aligned} \tag{41}$$

For every z, at least one of the two losses in (41) is at least $\kappa d_1/2$. Integrating this statement with respect to the common part of the two product measures, whose mass is $1 - \mathrm{TV}(P_{\theta_-}^{n_1}, P_{\theta_+}^{n_1})$, yields

$$\max_{\vartheta\in\{\theta_-,\theta_+\}} \mathbb{P}_{\vartheta,P_\vartheta,P_\vartheta}\left(\left|q_g(\widehat{\theta}_1 - \vartheta) - q_\star\right| \ge \frac{\kappa d_1}{2}\right) \ge \frac{3}{8}. \tag{42}$$

Step 2: training-contamination difficulty. Choose a sufficiently small constant $c_2 > 0$ and let

$$d_2 := c_2\left(L_\star\delta_1 \wedge h_0\right)$$

and again take $\theta_\pm = \pm d_2/2$. If $\delta_1 = 0$, skip this construction. Suppose henceforth in this step that $\delta_1 > 0$. By (30), after reducing c_2 if necessary,

$$\mathrm{TV}(P_{\theta_-}, P_{\theta_+}) \le C_{\mathrm{TV}}\frac{d_2}{L_\star} \le \frac{\delta_1}{1-\delta_1}.$$

Lemma B.3 supplies $Q_{1,-}, Q_{1,+}$ for which the two observed training distributions are identical:

$$(1-\delta_1)P_{\theta_-} + \delta_1 Q_{1,-} = (1-\delta_1)P_{\theta_+} + \delta_1 Q_{1,+}.$$

For $\vartheta \in \{\theta_-, \theta_+\}$, write $Q_{1,\vartheta} := Q_{1,-}$ when $\vartheta = \theta_-$ and $Q_{1,\vartheta} := Q_{1,+}$ when $\vartheta = \theta_+$. Also set $Q_{2,\pm} = P_{\theta_\pm}$, so the corresponding calibration laws are clean. Consequently, $\widehat{\theta}_1$ has the same law under the two hypotheses. The pointwise separation (41), now with d_2 in place of d_1, gives

$$\max_{\vartheta\in\{\theta_-,\theta_+\}} \mathbb{P}_{\vartheta,Q_{1,\vartheta},P_\vartheta}\left(\left|q_g(\widehat{\theta}_1 - \vartheta) - q_\star\right| \ge \frac{\kappa d_2}{2}\right) \ge \frac{1}{2}. \tag{43}$$

Since $\max\{d_1, d_2\} \ge ca_{n_1}$, use the clean construction alone when $\delta_1 = 0$ and the harder of the two constructions otherwise. Equations (42) and (43) then imply, as applicable, that there exist θ, Q_1 and a constant $c_{\mathrm{tr}} > 0$ such that

$$\mathbb{P}_{\theta,Q_1,P_\theta}\left(|q_g(H) - q_\star| \ge ca_{n_1}\right) \ge c_{\mathrm{tr}} \tag{44}$$

We next pass from the population radius to the empirical conformal radius. In the preceding construction $Q_2 = P_\theta$, so the observed calibration sample is clean. Under the regular conditional law given $H = h$, let $\widehat{q}_h$ denote the k_{n_2}-th order statistic of n_2 independent scores with CDF F_h, and put $B_{n_2,h} := F_h(\widehat{q}_h)$. Continuity of F_h gives

$$B_{n_2,h} \sim \mathrm{Beta}(k_{n_2}, n_2 + 1 - k_{n_2}),$$

uniformly in $|h| \le h_0$. In particular,

$$\mathbb{E}B_{n_2,h} = \frac{k_{n_2}}{n_2+1} = 1 - \alpha + O(n_2^{-1}), \qquad \mathrm{Var}(B_{n_2,h}) \le \frac{1}{4(n_2+2)}.$$

Chebyshev's inequality therefore supplies a constant $C_0 < \infty$ such that, for all sufficiently large n_2,

$$\sup_{|h|\le h_0} \mathbb{P}\left(|B_{n_2,h} - (1-\alpha)| > \frac{C_0}{\sqrt{n_2}} \;\middle|\; H = h \right) \le \frac{c_{\mathrm{tr}}}{2}.$$

After increasing n_2 further so that $C_0/(r_\star\sqrt{n_2}) \le 2s_0$, monotonicity of F_h and the local lower density bound in (31) imply

$$\sup_{|h|\le h_0} \mathbb{P}\left(|\widehat{q}_h - q_g(h)| > \frac{C_0}{r_\star\sqrt{n_2}} \;\middle|\; H = h \right) \le \frac{c_{\mathrm{tr}}}{2} \tag{45}$$

for all sufficiently large n_2. Indeed, if $\widehat{q}_h$ lies farther than the displayed threshold from $q_g(h)$, the CDF difference is at least $C_0/\sqrt{n_2}$, including when $\widehat{q}_h$ lies outside the local band. Increase n_0 if necessary so that $u_{n_2} = 1/(r_\star\sqrt{n_2})$ throughout the remainder of the proof. Choose a sufficiently large constant $C_{\mathrm{dom}} > 0$. If $a_{n_1} \ge C_{\mathrm{dom}} u_{n_2}$, then (44) and (45), followed by the reverse triangle inequality, give, for some fixed $c_{\mathrm{train}} > 0$,

$$\sup_{\theta, Q_1, Q_2} \mathbb{P}_{\theta,Q_1,Q_2}\left(|\widehat{q}_H - q_\star| \ge c a_{n_1} \right) \ge c_{\mathrm{train}}. \tag{46}$$

Step 3: clean calibration-sample difficulty. Fix $\theta = 0$ and choose $Q_1 = P_{\theta=0}$ and $Q_2 = P_{\theta=0}$. Conditional on $H = h$, the calibration scores are i.i.d. with continuous CDF F_h. The probability integral transform for the k_{n_2}-th order statistic gives

$$F_h(\widehat{q}_h) \sim \mathrm{Beta}\left(k_{n_2}, n_2 + 1 - k_{n_2}\right). \tag{47}$$

Under this conditional law, write $B_{n_2} := F_h(\widehat{q}_h)$ for the beta random variable in (47). We now derive a uniform probability lower bound, rather than an expected loss. First suppose that $|q_\star - q_g(h)| \le s_0$ and put $u_h := F_h(q_\star)$. By Lemma B.4, with conditional probability at least $1/2$,

$$|B_{n_2} - u_h| \ge \frac{c_B}{\sqrt{n_2}}.$$

If $\widehat{q}_h$ lies outside $[q_g(h) - 2s_0, q_g(h) + 2s_0]$, then $|\widehat{q}_h - q_\star| \ge s_0$. Otherwise, absolute continuity and the upper density bound in (31) give

$$|B_{n_2} - u_h| = |F_h(\widehat{q}_h) - F_h(q_\star)| \le C_f r_\star |\widehat{q}_h - q_\star|.$$

Thus, in this first case,

$$\mathbb{P}\left(|\widehat{q}_h - q_\star| \ge c u_{n_2} \mid H = h\right) \ge \frac{1}{2}.$$

If $|q_\star - q_g(h)| > s_0$, use (39) with $\epsilon = r_\star s_0/4$. The lower density bound then implies that, with conditional probability at least $3/4$ for all sufficiently large n_2, $|\widehat{q}_h - q_g(h)| \le s_0/2$. Hence $|\widehat{q}_h - q_\star| \ge s_0/2$ on that event. Combining the two cases, uniformly over $|h| \le h_0$, and then integrating over H, gives, for some fixed $c_{\mathrm{cal},0} > 0$,

$$\sup_{\theta, Q_1, Q_2} \mathbb{P}_{\theta,Q_1,Q_2}\left(|\widehat{q}_H - q_\star| \ge c u_{n_2} \right) \ge c_{\mathrm{cal},0}. \tag{48}$$

Step 4: calibration-contamination difficulty. Choose $c_3 > 0$ small enough that $c_3 r_\star s_0 \le \alpha/2$ and that

$$\frac{\bar{\delta}}{1-\bar{\delta}} \le \frac{r_\star s_0}{2} \qquad \text{whenever} \qquad 0 \le \bar{\delta} \le c_3 r_\star s_0.$$

Set

$$\bar{\delta}_2 := \delta_2 \wedge (c_3 r_\star s_0).$$

When $\delta_2 > 0$, define contamination distributions so that the effective non-clean mass is $\bar{\delta}_2$: let

$$\widetilde{Q}_{2,-} := \mathrm{Dirac}(0, z_-), \qquad \widetilde{Q}_{2,+} := \mathrm{Dirac}(0, z_+),$$

and set

$$Q_{2,\pm} := \frac{\bar{\delta}_2}{\delta_2}\widetilde{Q}_{2,\pm} + \left(1 - \frac{\bar{\delta}_2}{\delta_2}\right) P_{\theta=0}.$$

Then

$$(1-\delta_2)P_{\theta=0} + \delta_2 Q_{2,\pm} = (1-\bar{\delta}_2)P_{\theta=0} + \bar{\delta}_2 \widetilde{Q}_{2,\pm}.$$

The case $\delta_2 = 0$ is already covered by (48); hence assume $\delta_2 > 0$ for the remainder of this step.

Conditional on $H = h$, set

$$q_{\mathrm{low}}(h) = F_h^{-1}\left(\frac{1-\alpha-\bar{\delta}_2}{1-\bar{\delta}_2}\right),$$
$$q_{\mathrm{high}}(h) = F_h^{-1}\left(\frac{1-\alpha}{1-\bar{\delta}_2}\right).$$

Both probability levels differ from $1-\alpha$ by at most $\bar{\delta}_2/(1-\bar{\delta}_2) \le r_\star s_0/2$. The lower density bound in (31) therefore gives, uniformly over $|h| \le h_0$,

$$\max\{|q_{\mathrm{low}}(h) - q_g(h)|, |q_{\mathrm{high}}(h) - q_g(h)|\} \le s_0/2.$$

Condition (33) now places the low atom strictly below both quantiles and the high atom strictly above both; hence $q_{\mathrm{low}}(h)$ and $q_{\mathrm{high}}(h)$ are precisely the two mixed population quantiles. Their probability levels differ by $\bar{\delta}_2/(1-\bar{\delta}_2)$. The density upper bound therefore implies, uniformly in $|h| \le h_0$,

$$q_{\mathrm{high}}(h) - q_{\mathrm{low}}(h) \ge \frac{\bar{\delta}_2}{(1-\bar{\delta}_2)C_f r_\star} \ge c\left(\frac{\delta_2}{r_\star} \wedge s_0\right).$$

For $|h| \le h_0$, define the two mixed-score CDFs by

$$G_{\pm,h}(t) := (1-\bar{\delta}_2)F_h(t) + \bar{\delta}_2 \mathbf{1}\{|z_\pm - h| \le t\}.$$

The distribution of H is the same under $Q_{2,-}$ and $Q_{2,+}$, because $\widehat{\theta}_1$ is a function only of $\mathcal{D}_1$. Let $V_1, \ldots, V_{n_2}$ be i.i.d. $\mathrm{Unif}(0,1)$ variables, independent of H, and let $V_{(k_{n_2})}$ be their k_{n_2}-th order statistic. Set

$$B_{n_2}^{\mathrm{cpl}} := V_{(k_{n_2})} \sim \mathrm{Beta}(k_{n_2}, n_2 + 1 - k_{n_2}).$$

Conditional on $H = h$, couple the two score samples as $G_{\pm,h}^{-1}(V_1), \ldots, G_{\pm,h}^{-1}(V_{n_2})$. Their conformal order statistics are

$$\widehat{q}_{h,\pm} := G_{\pm,h}^{-1}(V_{(k_{n_2})}) = G_{\pm,h}^{-1}(B_{n_2}^{\mathrm{cpl}}).$$

Finally, define $\widehat{q}_{H,\pm} := G_{\pm,H}^{-1}(B_{n_2}^{\mathrm{cpl}})$. For all sufficiently large n_2, choose $c_4 > 0$ small enough that $c_4 r_\star s_0 \le \alpha/4$ and, whenever $|b - (1-\alpha)| \le c_4 r_\star s_0$, both transformed levels

$$\frac{b-\bar{\delta}_2}{1-\bar{\delta}_2}, \qquad \frac{b}{1-\bar{\delta}_2}$$

differ from $1-\alpha$ by at most $r_\star s_0$. The central-concentration part of Lemma B.4 shows that the event

$$\mathcal{E}_B := \left\{\left|B_{n_2}^{\mathrm{cpl}} - (1-\alpha)\right| \le c_4 r_\star s_0\right\}$$

has probability at least $3/4$. On this event, the lower density bound places both inverses in $[q_g(h) - s_0, q_g(h) + s_0]$. By (33), the low atom lies strictly below this interval and the high atom lies strictly above it. Thus the generalized inverses satisfy

$$G^{-1}_{-,h}(B^{\mathrm{cpl}}_{n_2}) = F^{-1}_h\left(\frac{B^{\mathrm{cpl}}_{n_2} - \bar{\delta}_2}{1 - \bar{\delta}_2}\right),$$
$$G^{-1}_{+,h}(B^{\mathrm{cpl}}_{n_2}) = F^{-1}_h\left(\frac{B^{\mathrm{cpl}}_{n_2}}{1 - \bar{\delta}_2}\right).$$

The two probability levels differ by $\bar{\delta}_2/(1 - \bar{\delta}_2)$. The upper density bound therefore gives, on $\mathcal{E}_B$,

$$\widehat{q}_{h,+} - \widehat{q}_{h,-} \geq c v_{n_2}.$$

For every coupled realization in $\mathcal{E}_B$, at least one of the two deviations from $q_\star$ is at least $cv_{n_2}/2$. Taking probabilities under the coupling, whose marginals are the two contaminated experiments, gives

$$\mathbb{P}_{0,P_{\theta=0},Q_{2,+}}\left(|\widehat{q}_{H,+} - q_\star| \geq cv_{n_2}/2\right) + \mathbb{P}_{0,P_{\theta=0},Q_{2,-}}\left(|\widehat{q}_{H,-} - q_\star| \geq cv_{n_2}/2\right) \geq \mathbb{P}(\mathcal{E}_B) \geq \frac{3}{4}.$$

Consequently,

$$\sup_{\theta,Q_1,Q_2} \mathbb{P}_{\theta,Q_1,Q_2}\left(|\widehat{q}_H - q_\star| \geq cv_{n_2}\right) \geq \frac{3}{8}. \tag{49}$$

Combining (49) with the clean calibration bound (48) and (40) gives, for some fixed $c_{\mathrm{cal}} > 0$,

$$\sup_{\theta,Q_1,Q_2} \mathbb{P}_{\theta,Q_1,Q_2}\left(|\widehat{q}_H - q_\star| \geq cb_{n_2}\right) \geq c_{\mathrm{cal}}. \tag{50}$$

Step 5: combining training and calibration. If $a_{n_1} \geq C_{\mathrm{dom}} u_{n_2}$, then (46) and (50) show that the larger of a_{n_1} and b_{n_2} is incurred with probability bounded away from zero under one of their least-favorable experiments. Since their maximum is at least half their sum, the desired radius-deviation statement follows. If $a_{n_1} < C_{\mathrm{dom}} u_{n_2}$, then $u_{n_2} \leq b_{n_2}$ by (40), so $a_{n_1} \leq C_{\mathrm{dom}} b_{n_2}$, and (50) alone controls the sum. Therefore, for constants $c, c_0 > 0$ independent of the estimator,

$$\sup_{\theta,Q_1,Q_2} \mathbb{P}_{\theta,Q_1,Q_2}\left(|\widehat{q}_H - q_\star| \geq c(a_{n_1} + b_{n_2})\right) \geq c_0.$$

Finally,

$$|2\widehat{q}_H - 2q_\star| = 2\,|\widehat{q}_H - q_\star|\,.$$

The preceding argument holds for every $\widehat{\theta} \in \mathcal{E}_{n_1}(\Theta)$. Multiplying by two and taking the infimum over this class proves (35). □

B.9 Fixed-Confidence Upper Bound and Local Rate Matching

Sampling stability in Assumption 3.1 uses a failure probability tending to zero. A fixed-confidence comparison with the lower bound instead uses a training radius indexed by its failure probability. For a clean law P and $\gamma \in (0,1)$, let $\eta_n(P;\gamma)$ be a deterministic radius satisfying

$$\mathbb{P}_{\mathcal{D}_n \sim P^n}\{\|\mu(\cdot;\mathcal{D}_n) - \mu(\cdot;P)\|_\infty > \eta_n(P;\gamma)\} \leq \gamma.$$

This notation does not assert that the radius in Assumption 3.1 is $O(n^{-1/2})$. The distinction matters because a fixed confidence level allows a constant multiple of $n^{-1/2}$, whereas a failure probability tending to zero may require a larger radius.

Proposition B.4 (Fixed-confidence interval-length bound)**.** *Suppose Assumption 3.2 holds and μ is L-uniformly replacement stable. Fix $\beta \in (0,1)$ and contamination levels $\delta_1 \in [0,1)$, $\delta_2 \in [0,\alpha)$, with $\alpha \in (0,1/2)$. If the clean training radius $\eta_{n_1}(P_0;\beta/3)$ satisfies the preceding probability bound, then, for all sufficiently large n_2, with probability at least $1-\beta$, uniformly over Q_1, Q_2,*

$$2|\widehat{q}(\mathcal{D}_1;\mathcal{D}_2) - q(P_0;P_0)| \leq 2\eta_{n_1}(P_0;\beta/3) + 2L\left(\delta_1 + \sqrt{\frac{\log(3/\beta)}{2n_1}}\right)$$
$$+ \frac{2}{(1-\delta_2)r_0}\left(\sqrt{\frac{\log(6/\beta)}{2n_2}} + \frac{2}{n_2} + \delta_2(1-\alpha)\right). \tag{51}$$

Proof. Under the Huber coupling, $K \sim \mathrm{Binomial}(n_1, \delta_1)$. Hoeffding's inequality gives

$$\mathbb{P}\left\{ \frac{K}{n_1} > \delta_1 + \sqrt{\frac{\log(3/\beta)}{2n_1}} \right\} \le \beta/3.$$

On this event's complement and the clean training event, replacement stability bounds the predictor error by $\eta_{n_1}(P_0; \beta/3) + L(\delta_1 + \sqrt{\log(3/\beta)/(2n_1)})$. The absolute-score and order-statistic inequalities transfer this bound to the empirical quantile on the same calibration sample. Lemma 4.3 with $\gamma = \beta/3$ bounds calibration sampling error, and Lemma 4.4 bounds population calibration bias. A union bound for the three failure events and multiplication by two prove (51). □

An attaining learner on the location submodel. Consider the fixed location model in Assumption B.1, with known noise law g and a fixed scale σ comparable to $L_\star$, as in its scale-family example. We first apply a bounded transformation so that even an arbitrarily large replacement has a bounded effect on the sample average. We then invert its population mean to recover the location parameter. Define

$$\psi(y) := \tanh(y/\sigma), \qquad m(\theta) := \mathbb{E}_g \psi(\theta + \varepsilon),$$
$$c_\psi := \inf_{\theta \in \Theta} \mathbb{E}_g \operatorname{sech}^2((\theta + \varepsilon)/\sigma),$$

where $\operatorname{sech}(z) := 1/\cosh(z)$. Positivity, continuity, and compactness of Θ give $c_\psi > 0$ and $m'(\theta) \ge c_\psi/\sigma$. Let $\mathrm{Proj}_{m(\Theta)}$ denote Euclidean projection onto the interval $m(\Theta)$, and set

$$\widehat{\theta}(\mathcal{D}_n) := m^{-1}\left(\mathrm{Proj}_{m(\Theta)} \left[\frac{1}{n} \sum_{i=1}^{n} \psi(Y_i) \right] \right), \qquad \mu(\mathbf{x}; \mathcal{D}_n) := \widehat{\theta}(\mathcal{D}_n).$$

This estimator takes values in Θ. Since projection is nonexpansive and m^{-1} is σ/c_ψ-Lipschitz, replacing k arbitrary observations gives

$$|\widehat{\theta}(\mathcal{D}_n) - \widehat{\theta}(\mathcal{D}'_n)| \le \frac{2\sigma}{c_\psi} \frac{k}{n}.$$

For clean data, $\mathbb{E}_\theta \psi(Y) = m(\theta)$ and $|\psi| \le 1$. Hoeffding's inequality therefore yields, uniformly over $\theta \in \Theta$,

$$\mathbb{P}_\theta \left\{ |\widehat{\theta} - \theta| > \frac{\sigma}{c_\psi} \sqrt{\frac{2\log(2/\gamma)}{n}} \right\} \le \gamma.$$

Thus the clean population target is θ, $L = 2\sigma/c_\psi \asymp L_\star$, and $\eta_n(P_\theta; \gamma) = O_\gamma(L_\star n^{-1/2})$ uniformly over Θ.

Restrict δ_2 to a sufficiently small fixed neighborhood of zero so that the quantile range lies inside the common density interval in Assumption B.1. The clean oracle score law is the same for every θ, so r_0 can be chosen comparable to $r_\star$ with a uniform interval margin. Proposition B.4 then gives the order $L_\star(n_1^{-1/2} + \delta_1) + (n_2^{-1/2} + \delta_2)/r_\star$. When the local truncations are inactive, Theorem B.1 supplies the matching lower order for any fixed failure probability $\beta < c_0$. This establishes attainability within the stated estimator class on this fixed submodel. It does not establish the same training rate for every replacement-stable learner or for the experimental neural networks.

C Experimental details

C.1 Common protocol

We study length deviations in prediction intervals under training and calibration contamination. The training and calibration contamination probabilities are δ_1 and δ_2, respectively. Test observations remain uncontaminated. The target miscoverage level is $\alpha = 0.1$. We use this level for calibration in the following text unless comparing the corrections in Figures 7, 10, and 11.

Scores and prediction sets. We compare Absolute, Adaptive, and CQR scores:

$$s_{\mathrm{abs}}(x,y) = |y - \widehat{\mu}(x)|,$$
$$s_{\mathrm{adp}}(x,y) = |y - \widehat{\mu}(x)|/\widehat{\sigma}(x),$$
$$s_{\mathrm{cqr}}(x,y) = \max\{\widehat{q}^{\mathrm{lo}}(x) - y, y - \widehat{q}^{\mathrm{hi}}(x)\}.$$

The location model estimates the conditional median. Adaptive score uses the same location model and a separate positive scale estimate. CQR estimates the 0.05 and 0.95 conditional quantiles. We sort the two fitted CQR endpoints at each input before computing scores. These base quantile levels remain fixed in the correction experiments.

For n_{cal} calibration scores, let

$$k = \lceil (n_{\mathrm{cal}} + 1)(1-\alpha) \rceil, \qquad \widehat{q} = R_{(k)}, \qquad R_{(n_{\mathrm{cal}}+1)} = +\infty.$$

The three prediction sets are

$$\begin{aligned} \text{Absolute:} \quad & [\widehat{\mu}(x) - \widehat{q}, \widehat{\mu}(x) + \widehat{q}] \\ \text{Adaptive:} \quad & [\widehat{\mu}(x) - \widehat{q} \cdot \widehat{\sigma}(x), \widehat{\mu}(x) + \widehat{q} \cdot \widehat{\sigma}(x)] \\ \text{CQR:} \quad & [\widehat{q}^{\mathrm{lo}}(x) - \widehat{q}, \widehat{q}^{\mathrm{hi}}(x) + \widehat{q}] \end{aligned}$$

CQR scores and threshold $\hat{q}$ may be negative. If the final lower endpoint exceeds the upper endpoint, the prediction set is empty. Both its length and coverage contribution are zero.

Replacement sampling. Except for the diagnostic in Figure 3a and the fixed-fraction experiments in Appendix E, each observation is replaced with the specified contamination probability δ_j at stage j. Within a repetition, we draw one independent uniform variable per observation and replace the observation if this draw is at most δ_j at stage j. Thus, the replacement sets are nested as δ_j increases, while ensuring that each set has the required Bernoulli marginal distribution. Training and calibration use independent random streams. For structured contamination, we use the same replacement masks and candidate pairs (X_Q, Y_Q) across scores. Within each family, different severity levels also share the same masks and base noise draws.

Algorithm 1 Split conformal prediction with contamination

Input: Training, calibration, and clean test sets; contamination rules Q_1, Q_2; probabilities δ_1, δ_2; level α.

1: Replace training observations according to Q_1 and δ_1. Fit all models needed for the chosen score.
2: If required by the design, construct Q_2 from the fitted models; if not, skip this step.
3: Replace calibration observations according to Q_2 and δ_2.
4: Compute calibration scores. Set $k = \lceil (n_{\mathrm{cal}}+1)(1-\alpha) \rceil$ and $\widehat{q} = R_{(k)}$, with $R_{(n_{\mathrm{cal}}+1)} = +\infty$.
5: Form $C(x) = \{y : s(x,y) \le \widehat{q}\}$ and evaluate it on the clean test set.

Metrics and repetitions. Let $w_r^{\delta}(x)$ be the interval length in repetition r under the specified contamination setting. For the score under consideration, define the reference length by

$$w_r^{\mathrm{ref}}(x) = \begin{cases} w_{\mathrm{pop}}(x), & \text{for synthetic data,} \\ w_r^0(x), & \text{for real data,} \end{cases}$$

where $w_{\mathrm{pop}}(x)$ is the clean population-oracle length and $w_r^0(x)$ is the length of clean split conformal prediction fitted and calibrated within the same real-data repetition at $\alpha = 0.1$. We record the signed mean difference

$$D_r = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left[w_r^{\delta}(X_i) - w_r^{\mathrm{ref}}(X_i) \right].$$

Positive values indicate interval inflation; negative values indicate contraction. The same dataset-specific reference is used for the correction experiments. We also record the mean pointwise absolute difference,

$$\frac{1}{n_{\mathrm{test}}} \sum_i |w_r^{\delta}(X_i) - w_r^{\mathrm{ref}}(X_i)|.$$

We use the mean pointwise absolute difference for the contamination plots and the signed difference D_r for the calibration-level correction plots (Figures 7, 10, and 11). Coverage is defined as the fraction of clean test responses contained in their prediction sets.

For all curves and surfaces, we report means over 20 repetitions. Where shaded bands appear, they indicate the sample standard deviation across repetitions, using denominator 19. The score histograms are based on repetition $r = 0$, which was selected in advance.

C.2 Synthetic setup

The five coordinates of X are independent and uniformly distributed on $[-1, 1]$. We generate the response from

$$Y = f(X) + \sigma\varepsilon, \qquad f(x) = \sin(\pi x_1) + x_2^2 - \tfrac{1}{2}x_3, \qquad \varepsilon \sim N(0, 1).$$

In each repetition, the training, calibration, and test samples are independent and have sizes 5,000, 5,000, and 10,000, respectively. We set $\sigma = 1$, except in Figure 7, where $\sigma = 3$. Each regression model uses 36 fixed features: an intercept; the linear, quadratic, and cubic terms of each coordinate; $\sin(\pi x_j)$ and $\cos(\pi x_j)$ for each coordinate; and the ten pairwise interactions $x_i x_j$, $i < j$. For a quantile level τ, we minimize

$$\frac{1}{n_{\text{train}}} \sum_i \rho_\tau\big(Y_i - \theta^\top \phi(X_i)\big) + \frac{0.1}{2}\|\theta\|_2^2, \qquad \rho_\tau(u) = u\big(\tau - \mathbf{1}\{u < 0\}\big).$$

After fitting, all predictions are clipped to $[-20, 20]$. The optimization uses accelerated projected gradient updates on the dual, with working-set refinement, a relative primal–dual gap tolerance of 10^{-8}, and at most 200,000 iterations. The gap is checked every ten iterations.

The location model is fitted at $\tau = 0.5$. The Adaptive scale model also uses $\tau = 0.5$, using absolute training residuals as responses. Its predictions are clipped to $[0.25, 20]$. When the training data are contaminated, both models are fitted on the corresponding contaminated training sample. CQR fits separate models at $\tau = 0.05$ and 0.95.

For the synthetic experiments, the reference is the clean population oracle obtained using the same regularized learner and population score calibration; Adaptive and CQR use pointwise oracle widths evaluated at the corresponding test covariates.

C.3 Contamination designs

Joint training and calibration contamination. For Figures 4 and 6, we set

$$Q_1 = Q_2 = \text{Dirac}(x^\star, -80), \qquad x^\star = (-1, -1, -1, -1, +1).$$

The same covariate and response values are used at both stages and for all three scores. The models are refitted for each training contamination level. Consequently, the score assigned to the fixed calibration point may vary with δ_1. Figure 3a uses the same training point mass, without calibration contamination.

Structured calibration contamination. The training data remain clean in these experiments. Table 2 defines the four contamination families. In the first three families, we set $X_Q = X$. The Gaussian and Student-t draws are independent of the clean data. In the shift family, we retain the observed response and then add the shift.

Table 2: Structured contamination for the synthetic experiments ($\sigma = 1$). Here $\varepsilon_Q \sim N(0,1)$ and T_3 has a Student-t distribution with three degrees of freedom.

Family	X_Q	Y_Q	Parameter values
Mean shift	X	$Y + a$	$a \in \{3, 5, 10, 15\}$
Variance inflation	X	$f(X) + s\sigma\varepsilon_Q$	$s \in \{3, 5, 10, 15\}$
Heavy tail	X	$f(X) + c\sigma T_3/\sqrt{3}$	$c \in \{1, 3, 5, 10\}$
Covariate shift + heteroscedastic noise	$X + b\mathbf{1}_5$	$f(X_Q) + \sigma(0.5{+}1.5\|X_Q\|_\infty)\varepsilon_Q$	$b \in \{1.5, 3, 5, 8\}$

The multiplier s multiplies the noise standard deviation; the corresponding variance multiplier is therefore s^2. The random variable $T_3/\sqrt{3}$ has unit variance. The final family changes both the covariate distribution and the noise scale. For example, setting $b = 8$ shifts the support of the synthetic covariates to $[7, 9]^5$. Each family is evaluated over its own severity grid.

Finite high-score references. Figures 5, 8, and 12 use calibration point masses with score $M = 200$ as empirical references. The training data remain clean. For the synthetic data, the covariate is $x^\star$, and we choose the response for each score as

$$y_Q = \begin{cases} \widehat{\mu}_0(x^\star) - M, & \text{Absolute,} \\ \widehat{\mu}_0(x^\star) - M\widehat{\sigma}_0(x^\star), & \text{Adaptive,} \\ \widehat{q}_0^{\,\mathrm{lo}}(x^\star) - M, & \text{CQR.} \end{cases}$$

The subscript 0 denotes a fit on clean training data. For $\delta_2 < \alpha$, the Dirac reference curves show the mean length deviation under these point masses, plus one sample standard deviation. These are descriptive summaries for a specified finite contamination distribution, not theoretical upper bounds. Randomness in the replacement counts can cause a sharp increase before the contamination probability reaches 0.1.

Low-score contamination and calibration-level correction. For Absolute and Adaptive, we retain the selected calibration covariates and replace their responses with $\widehat{\mu}_0(X)$. Their scores are then zero. For CQR, let

$$B_0(x) = \widehat{q}_0^{\,\mathrm{hi}}(x) - \widehat{q}_0^{\,\mathrm{lo}}(x).$$

We numerically search for a point $x^\star_{\mathrm{CQR}}$ with large $B_0(x)$ and set

$$X_Q = x^\star_{\mathrm{CQR}}, \qquad Y_Q = \frac{\widehat{q}_0^{\,\mathrm{lo}}(x^\star_{\mathrm{CQR}}) + \widehat{q}_0^{\,\mathrm{hi}}(x^\star_{\mathrm{CQR}})}{2}.$$

The resulting score is $-B_0(x^\star_{\mathrm{CQR}})/2$. For synthetic data, the search domain is $[-1, 1]^5$. We evaluate a scrambled Sobol sequence of 2^{16} points and all 32 corners. The eight best candidates are used to initialize Powell searches, each with at most 300 iterations, a position tolerance of 10^{-8}, and an objective tolerance of 10^{-10}. We select the point with the largest fitted band length among all candidate and locally searched points.

We compare the uncorrected level with two corrections:

$$\alpha'_{\mathrm{unc}} = \alpha, \qquad \alpha'_{\mathrm{tight}} = \alpha(1 - \delta_2), \qquad \alpha'_{\mathrm{naive}} = \alpha - \delta_2.$$

Each rule replaces α only in the calibration rank. The additive rule is used only for $\delta_2 \le \alpha$. At $\delta_2 = \alpha$, it gives $\alpha'_{\mathrm{naive}} = 0$, rank $n_{\mathrm{cal}} + 1$, an infinite interval length, and coverage equal to one. For any common δ_2 for which both rules are defined, the additive rule selects a threshold that is at least as large as the one selected by the multiplicative rule.

C.4 Grids, diagnostics, and reproducibility

For the length-trend and joint-contamination experiments, δ_2 ranges from zero to 0.11 in increments of 0.005. We also include 0.096, 0.097, 0.098, and 0.099, for a total of 27 points. The joint

training grid ranges from zero to 0.30 in increments of 0.005, for a total of 61 points. The correction experiments range from zero to 0.95 with the same step size, for a total of 191 points. Figures 8 and 12 display structured means at

$$\delta_2 \in \{0, 0.015, 0.03, 0.045, 0.06, 0.075, 0.09, 0.10, 0.11\}.$$

The results for the full grid are retained.

Predictor and density diagnostics. Figure 3a uses exact replacement counts along one fixed random permutation of the training sample. The first 11 settings replace zero through ten of the $5{,}000$ training observations. The remaining contamination levels range from 0.005 to 0.095 in increments of 0.005. In each repetition, we compute

$$\eta_r(\delta_1) = \max_{1 \le i \le n_{\text{test}}} \left| \widehat{\mu}_{r,\delta_1}(X_i) - \widehat{\mu}_{r,0}(X_i) \right|$$

and fit a line through the origin to these values. The mean fitted slope across repetitions is 56.37, with a standard deviation of 5.28. This slope summarizes the sensitivity of the predictor to this point-mass contamination along the specified replacement path, as measured on the clean test sample.

All prediction-interval thresholds remain empirical order statistics, as specified in Appendix C.1. For the density diagnostic in Figure 3b and the reference curve in Figure 4a, we condition on each saved clean predictor $\widehat{\mu}_{r,0}$ and compute the distribution of $|f(X) - \widehat{\mu}_{r,0}(X) + \varepsilon|$ under the known synthetic law. Here $X \sim \text{Unif}([-1,1]^5)$ and $\varepsilon \sim N(0,1)$ are independent. We integrate over the Gaussian noise analytically and over the covariates numerically using 2^{16} scrambled Sobol points (seed 20260915). This yields the score CDF and density without kernel smoothing. For $\delta_2 < \alpha$, the two endpoint probabilities of the interval are

$$p_-(\delta_2) = \frac{1 - \alpha - \delta_2}{1 - \delta_2}, \qquad p_+(\delta_2) = \frac{1 - \alpha}{1 - \delta_2}.$$

We numerically invert this CDF at the two endpoint probabilities to locate the endpoints of the diagnostic interval. For each evaluated model, we verify that the density is decreasing throughout this interval, so its minimum is attained at the upper endpoint, and we compute that minimum. These diagnostic quantiles are not used to construct prediction intervals. Let $\widehat{r}_{0,r}(\delta_2)$ denote the computed density minimum. The computed minima change by less than 1.72×10^{-5} relative to the primary computation when using 2^{14} points or an independent scramble. Figure 3b reports the mean of this quantity across repetitions. The reference curve in Figure 4a is computed as

$$\overline{B}(\delta_2) = \frac{1}{20} \sum_{r=0}^{19} \frac{2\delta_2(1-\alpha)}{(1-\delta_2)\widehat{r}_{0,r}(\delta_2)}.$$

This diagnostic curve evaluates the population-level contamination expression with fitted-model scores and then averages over repetitions. It uses the clean score distribution of each fitted predictor instead of that of the population learner, and it omits the training and calibration sampling terms in Theorem 4.1. It is distinct from the clean population-oracle interval reference used in the synthetic length metric and is not a certified finite-sample upper bound. Comparisons near $\delta_2 = \alpha$ are therefore interpreted descriptively and remain subject to finite-sample variability. Figure 4a uses logarithmic axes and shows only nonzero contamination levels. Its calibration-only branch is restricted to $\delta_2 < \alpha$, consistent with the density diagnostic.

Score histograms. Figures 9, 13, and 14 use repetition 0 and $\delta_2 = 0.1$. The family parameters are $a = 5$, $s = 5$, $c = 10$, and $b = 8$. Scores and source labels are taken from the same contaminated calibration samples used for the length trends. Each panel uses 40 equal-width bins. Each source component has height $N_{\text{source,bin}}/(n_{\text{cal}} \times \text{bin width})$, so the stacked components sum to the mixture histogram. The vertical lines use the conformal order statistics for the clean and contaminated calibration samples.

Most panels cover the full range of the pooled scores from the clean, contamination-only, and mixture settings. For heavy tails, the initial range is based on pooled quantiles 0.002 and 0.985, extended to include both thresholds, and given a 4% margin. If the lower endpoint is nonnegative before adding the margin, the final lower limit is clipped at zero. Figure 1 overlays the clean and contaminated Absolute-score histograms for Abalone at $a = 5$, using the same fixed repetition, 40 bins, and range $[0, 8]$. Density heights are always normalized using the full sample size, including observations outside the displayed range. Across the displayed histograms, at most 0.46% of a mixture sample lies outside its window.

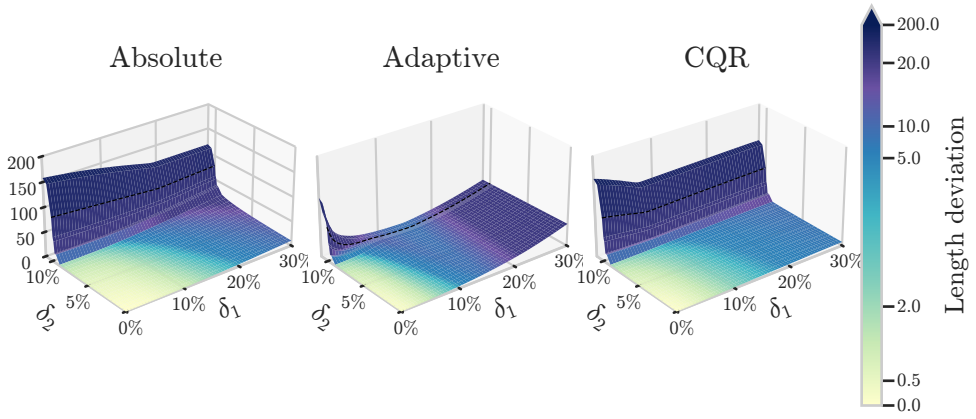


Figure 6: Joint training and calibration contamination for Absolute, Adaptive, and CQR. Both stages use $Q_1 = Q_2 = \mathrm{Dirac}((-1, -1, -1, -1, +1), -80)$. The two horizontal axes are δ_1 and δ_2; height and color show the mean length deviation over 20 repetitions at $\alpha = 0.1$. Height uses a linear 0–200 scale, and color follows the labeled nonlinear scale.

Random seeds and saved outputs. Repetitions are indexed by $r = 0, \ldots, 19$, with a data-split seed $42 + 10{,}000r$. Auxiliary streams are seeded with $42 + 10{,}000r + 100j + c$, where j and c are fixed stream identifiers recorded for each experiment. We save the split indices, fitted models, replacement masks, contaminated observations, scores, calibration ranks, and per-repetition metrics. The implementation uses Python 3.12.14, NumPy 2.5.3, SciPy 1.16.2, scikit-learn 1.7.2, and PyTorch 2.8.0 (CPU). Numerical computations are run using one process and one thread per library.

C.5 Further synthetic results

Figure 6 extends the joint contamination experiment to all three scores. Across the evaluated grid, the magnitude of the mean length deviation depends on the score and the two contamination levels. The fixed point mass produces finite widths across the evaluated grid, including for $\delta_2 \geq \alpha$. The theoretical breakdown result is stated in terms of the supremum over calibration contamination distributions.

Figure 7 shows the effect of contamination from low-score calibration points. Without correction, intervals contract and coverage on the clean test set falls. Across all three scores and 191 contamination levels, the mean coverage over 20 repetitions under the multiplicative correction ranges from 0.897 to 0.904. The additive rule produces wider intervals where defined and an infinite interval at $\delta_2 = 0.1$.

Figure 8 shows that interval inflation depends on both the score and the contamination family. Under the combined covariate and noise change, Absolute and CQR exhibit much larger maximum mean changes in interval length than Adaptive over the tested grid. Under mean shift, the largest changes are similar across scores.

Figure 9 displays the score mixtures underlying these changes. For the fixed repetition and severity levels, contamination raises the conformal threshold in every panel. The shape and magnitude of the shift vary across scores. These panels correspond to a single calibration sample from repetition 0.

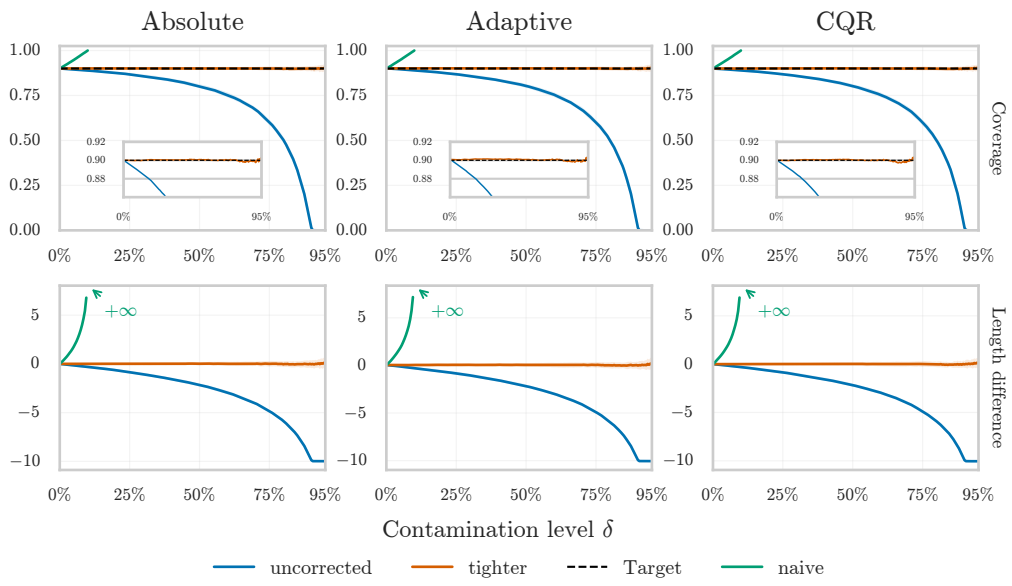


Figure 7: Calibration-level correction under low-score contamination in the synthetic model with $\sigma = 3$. Columns correspond to Absolute, Adaptive, and CQR. The top row shows clean-test coverage and the bottom row the signed length difference. Curves compare α, $\alpha(1-\delta_2)$, and $\alpha - \delta_2$, with target coverage 0.9. Results are means over 20 repetitions. The additive rule has infinite length at $\delta_2 = \alpha = 0.1$ and is undefined beyond this point.

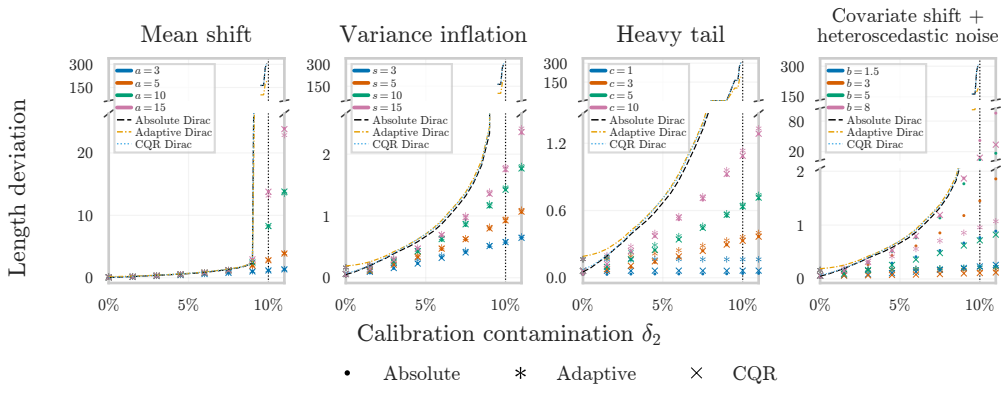


Figure 8: Synthetic calibration contamination at $\alpha = 0.1$. Columns show mean shift, variance inflation, heavy tails, and covariate shift plus heteroscedastic noise. Colors indicate severity and marker shapes distinguish scores. Points show mean length deviation over 20 repetitions. The score-200 Dirac references show mean plus one standard deviation for $\delta_2 < \alpha$. Broken vertical axes separate these large reference values from the structured results; the final column uses three segments. Tick ranges differ between panels.

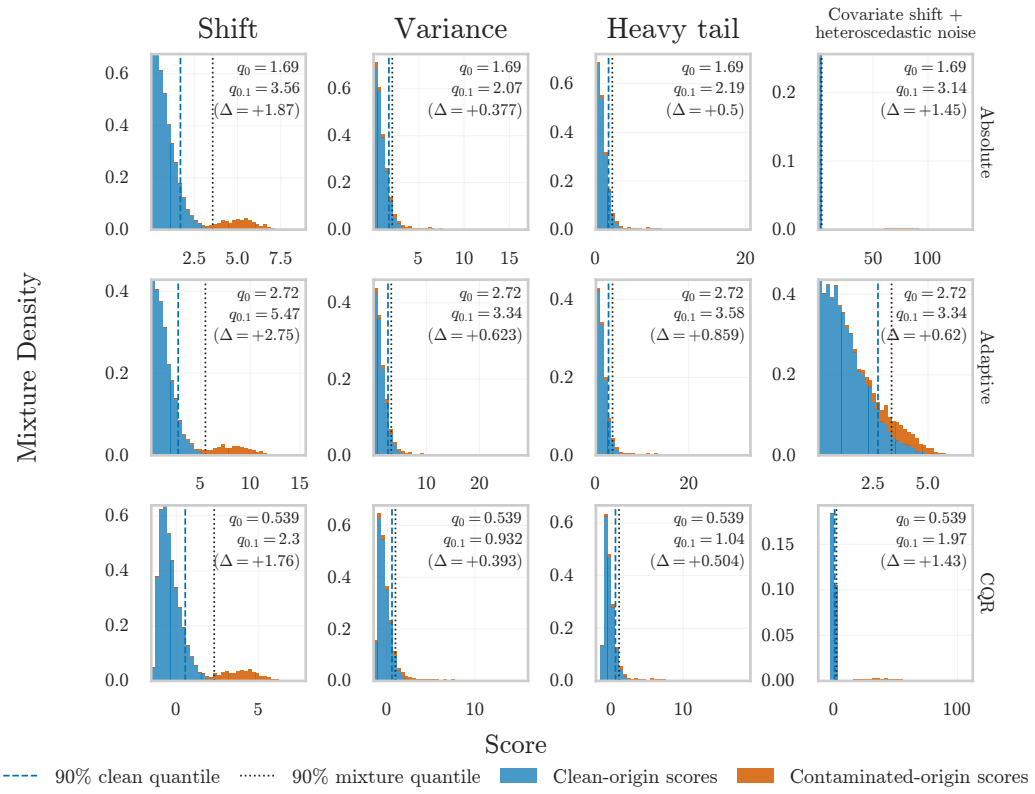


Figure 9: Synthetic score distributions at $\delta_2 = 0.1$, using the fixed repetition zero. Rows correspond to Absolute, Adaptive, and CQR; columns use $a = 5$, $s = 5$, $c = 10$, and $b = 8$. The respective replacement counts are 537, 463, 452, and 491 out of 5,000, shared across scores. Stacked bars distinguish retained clean observations from replacements. Vertical lines show the clean and contaminated conformal thresholds at $\alpha = 0.1$. Each panel reports their difference.

D EXPERIMENTS ON REAL DATA

D.1 DATA AND MODEL FITTING

We use the LIBSVM versions of `Abalone`, `California Housing (cadata)`, `CPU Small`, and `YearPredictionMSD (MSD)`. Table 3 reports the dimensions and sample sizes. For the first three datasets, a fixed `RandomState(42)` permutation assigns the first $\lfloor 0.8N \rfloor$ observations to a training pool and the remaining observations to a clean test set. Each repetition permutes the pool and selects disjoint training and calibration samples. For both Abalone and CPU Small, one pool observation is unused in each split. MSD provides predefined training and test files with $463{,}715$ and $51{,}630$ observations, respectively. Using seed 42, we first select a fixed training pool of $50{,}000$ observations and a fixed test subset of $10{,}000$ observations. Each repetition then draws $5{,}000$ training and $5{,}000$ calibration observations from this pool. Test sets remain fixed across repetitions. Variation across repetitions reflects the train–calibration split, model fitting, and contamination sampling, conditional on the fixed dataset and test set.

Table 3: Real-data dimensions and sample sizes. The training pool is fixed before the repeated train–calibration splits.

Dataset	d	Pool	Train	Calibration	Test
Abalone	8	3,341	1,670	1,670	836
California Housing	8	16,512	8,256	8,256	4,128
CPU Small	12	6,553	3,276	3,276	1,639
YearPredictionMSD	90	50,000	5,000	5,000	10,000

Features are standardized coordinate-wise using only the clean training split. Responses are centered and scaled using the mean and population standard deviation of the clean training split. These transformations are then applied to the calibration and test observations. Interval lengths are converted back to the original response units. Score histograms use the internal standardized response scale; for Adaptive, they additionally use the fitted scale division.

Each quantile model is a neural network trained with the pinball loss defined in Section C.2. Table 4 reports its configuration. The outer training split contains an internal 10% validation subset for early stopping. Absolute and Adaptive share the clean median network $\widehat{\mu}_0$. Adaptive fits a separate median network to the absolute residuals computed on the outer training split. CQR uses separate networks at levels 0.05 and 0.95.

Table 4: Neural-network configuration for all real-data experiments.

Component	Setting
Architecture	d–64–32–1, with ReLU activations
Optimizer	AdamW; learning rate 10^{-3}; weight decay 10^{-4}
Training	Batch size 256; at most 100 epochs; gradient norm clipped at 5
Early stopping	Internal validation fraction 0.1; patience 10; minimum improvement 10^{-4}; restore the best validation checkpoint

D.2 TRAINING RESIDUAL SCALE AND CONTAMINATION

All response-side contamination parameters are defined in standardized response units. After fitting $\widehat{\mu}_0$, we compute

$$e_i = |Y_i - \widehat{\mu}_0(X_i)|, \qquad \overline{e} = \frac{1}{n_{\text{train}}} \sum_i e_i,$$

$$\sigma_{\text{res}} = \max\left\{ \sqrt{\frac{1}{n_{\text{train}}} \sum_i (e_i - \overline{e})^2},\, 10^{-12} \right\}.$$

This is the standard deviation of the absolute residuals on the clean outer training split, computed with denominator n_{train}. We fix it before applying calibration contamination. The Adaptive-scale predictions are clipped to $[0.25\sigma_{\text{res}}, 20\sigma_{\text{res}}]$.

Training remains clean throughout the real-data experiments. Structured calibration contamination uses the same parameter grids as those in Table 2. The observed response is replaced by $Y + a\sigma_{\text{res}}$ for mean shift, by $\widehat{\mu}_0(X) + s\sigma_{\text{res}}\varepsilon_Q$ for variance inflation, and by $\widehat{\mu}_0(X) + c\sigma_{\text{res}}T_3/\sqrt{3}$ for heavy tails. For the composite family, we use

$$X_Q = X + b\mathbf{1}_d, \qquad Y_Q = \widehat{\mu}_0(X_Q) + \sigma_{\text{res}}(0.5 + 1.5\|X_Q\|_\infty)\varepsilon_Q.$$

The feature shift is applied after training-based standardization. The same clean median model defines these contamination centers for all three scores.

For the finite score-200 references, the anchor is the coordinate-wise signed extreme of the standardized training design:

$$x_j^\star = X_{i_j^\star, j}, \qquad i_j^\star \in \arg\max_i |X_{ij}|.$$

We set Y_Q to $\widehat{\mu}_0(x^\star) + 200$ for Absolute, $\widehat{\mu}_0(x^\star) + 200\widehat{\sigma}_0(x^\star)$ for Adaptive, and $\widehat{q}_0^{\text{hi}}(x^\star) + 200$ for CQR. These choices all give score 200 in the corresponding fitted score function.

Low-score contamination follows Section C.3. For the CQR search, the domain is the coordinate-wise minimum–maximum box of the standardized clean training features. Candidates comprise the training observations, 2^{14} scrambled Sobol points, 256 random corners, the corner with the largest absolute value in each coordinate, and its opposite corner. The best 16 candidates are refined by 300 Adam steps with learning rate 0.05, constrained to the box. We retain the best point found across the candidates and refinements.

D.3 Calibration-level correction

Figures 10 and 11 compare the three calibration-level rules under low-score contamination. The multiplicative correction maintains coverage near the target level 0.9, although differences remain across datasets and scores. Table 5 reports the range of mean coverage across the full grid and all three scores. Mean coverage on Abalone is above 0.9 throughout the grid. Some settings for the other datasets fall below 0.9. The additive correction is more conservative and yields infinite length at $\delta_2 = 0.1$.

Table 5: Range of mean clean-test coverage with $\alpha' = \alpha(1 - \delta_2)$. Each range spans all three scores and all 191 contamination levels; each mean uses 20 repetitions.

Dataset	Minimum–maximum mean coverage
Abalone	0.916–0.934
California Housing	0.888–0.899
CPU Small	0.886–0.908
YearPredictionMSD	0.887–0.897

D.4 Calibration contamination and interval length

Figure 12 compares the four structured families at $\alpha = 0.1$. For each dataset and score, the largest mean length difference across the tested severity and contamination grids occurs under mean shift. The ordering of the other families depends on the dataset and score. The Dirac curves show the mean plus one sample standard deviation, as defined in Section C.3.

We use broken vertical axes to display both the structured points and the much larger finite Dirac references. Each panel has its own set of axis ranges. California Housing uses an axis multiplier of 10^6, which is applied to both segments in each panel.

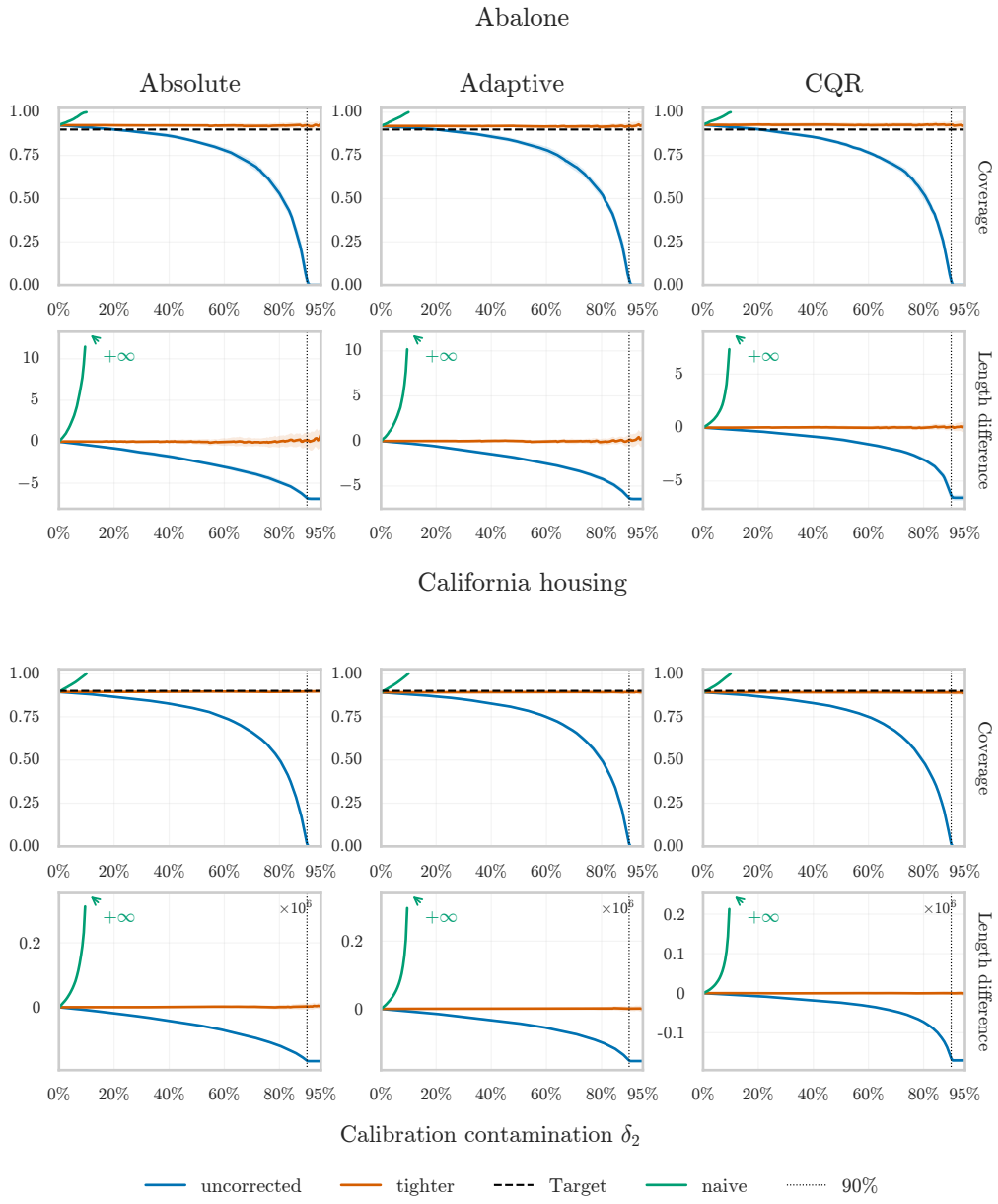


Figure 10: Calibration-level correction on Abalone (top two rows) and California Housing (bottom two rows). Columns correspond to Absolute, Adaptive, and CQR. Each dataset has a coverage row and a signed length difference row. Curves compare α, $\alpha(1-\delta_2)$, and $\alpha-\delta_2$ under low-score contamination and average 20 repetitions. The dashed horizontal line marks coverage 0.9, and the dotted vertical line marks $\delta_2 = 0.9$. The additive rule has infinite length at $\delta_2 = 0.1$. Length differences use the original response units.

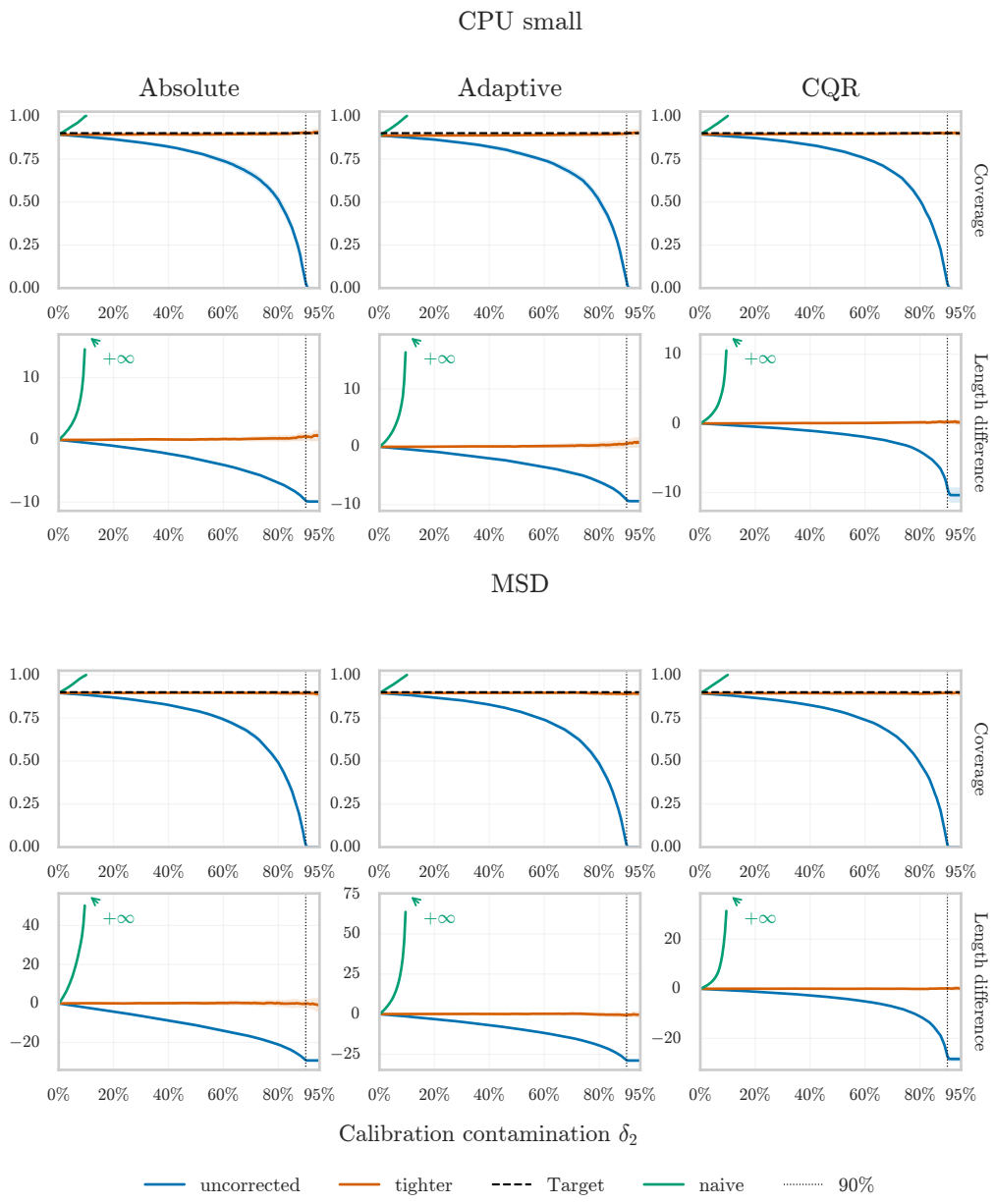


Figure 11: Calibration-level correction on CPU Small (top two rows) and YearPredictionMSD (bottom two rows), using the same conventions as Figure 10. Columns correspond to Absolute, Adaptive, and CQR. Results average 20 repetitions, and length differences are in the original response units. The additive rule has infinite length at $\delta_2 = 0.1$.

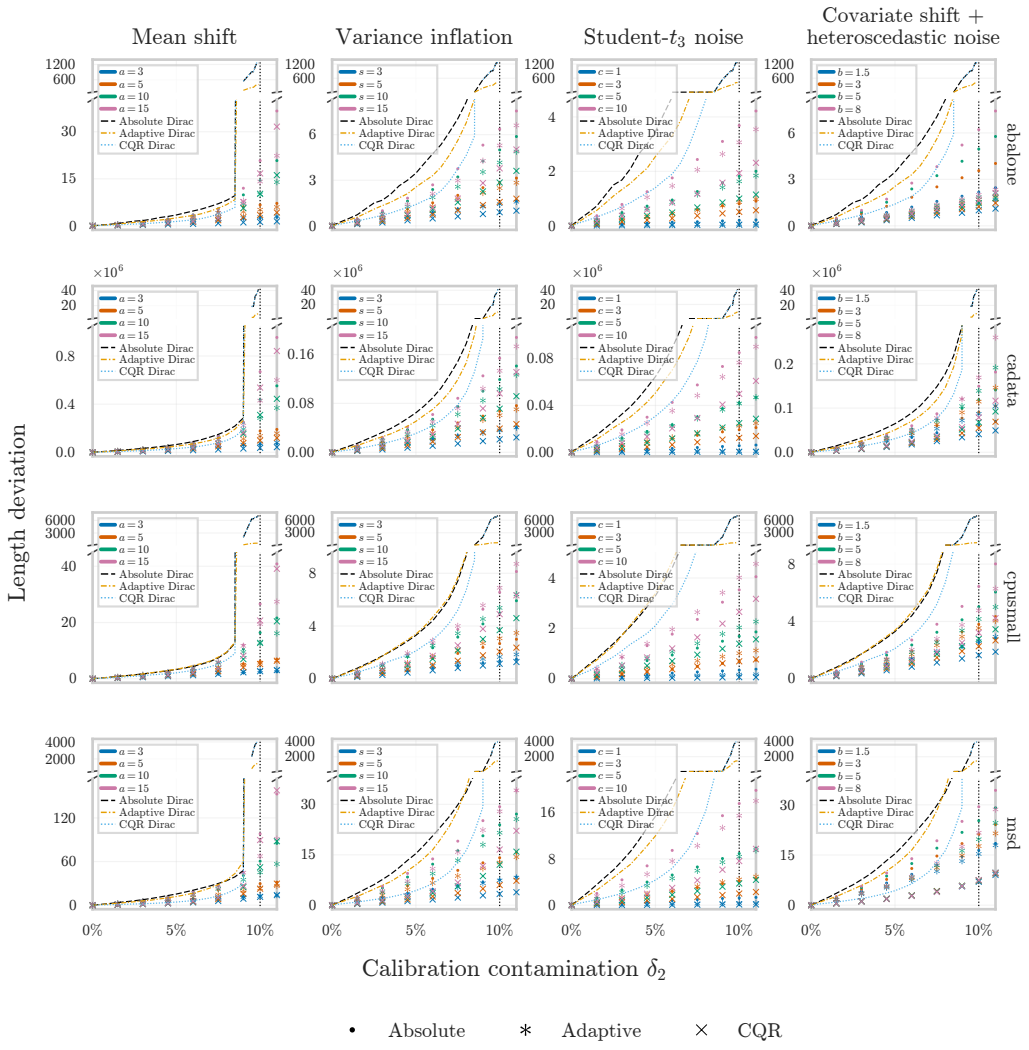


Figure 12: Real-data calibration contamination at $\alpha = 0.1$. Rows correspond to Abalone, California Housing, CPU Small, and YearPredictionMSD. Columns show mean shift, variance inflation, heavy tails, and covariate shift plus heteroscedastic noise. Colored markers show severity levels, and marker shapes distinguish scores. Points show mean length deviation over 20 repetitions, in original response units. The score-200 Dirac curves show mean plus one standard deviation only for $\delta_2 < \alpha$. Each panel uses two independently labeled vertical-axis segments; both California Housing segments share the displayed 10^6 multiplier.

D.5 SCORE DISTRIBUTIONS

Figures 13 and 14 use the fixed repetition 0, $\delta_2 = 0.1$, and the same four severity values as in the synthetic histograms. In each displayed panel, the contaminated conformal threshold is above the clean threshold. Their separation and the mixture shape vary across datasets, scores, and contamination families. The realized replacement fractions also differ because the sampling is Bernoulli. Table 6 reports the realized counts, which are shared across scores within each dataset and family. The effect on length depends on the score: for Adaptive, the threshold is multiplied by the fitted scale, while CQR adjusts the endpoints of its fitted base interval.

Table 6: Realized replacement counts for the real-data score histograms at $\delta_2 = 0.1$ in repetition zero. Columns use $a = 5$, $s = 5$, $c = 10$, and $b = 8$, respectively. "Composite" denotes covariate shift plus heteroscedastic noise.

Dataset	n_{cal}	Shift	Variance	Heavy tail	Composite
Abalone	1,670	183	161	144	161
California Housing	8,256	870	791	769	812
CPU Small	3,276	349	303	304	338
YearPredictionMSD	5,000	537	463	452	491

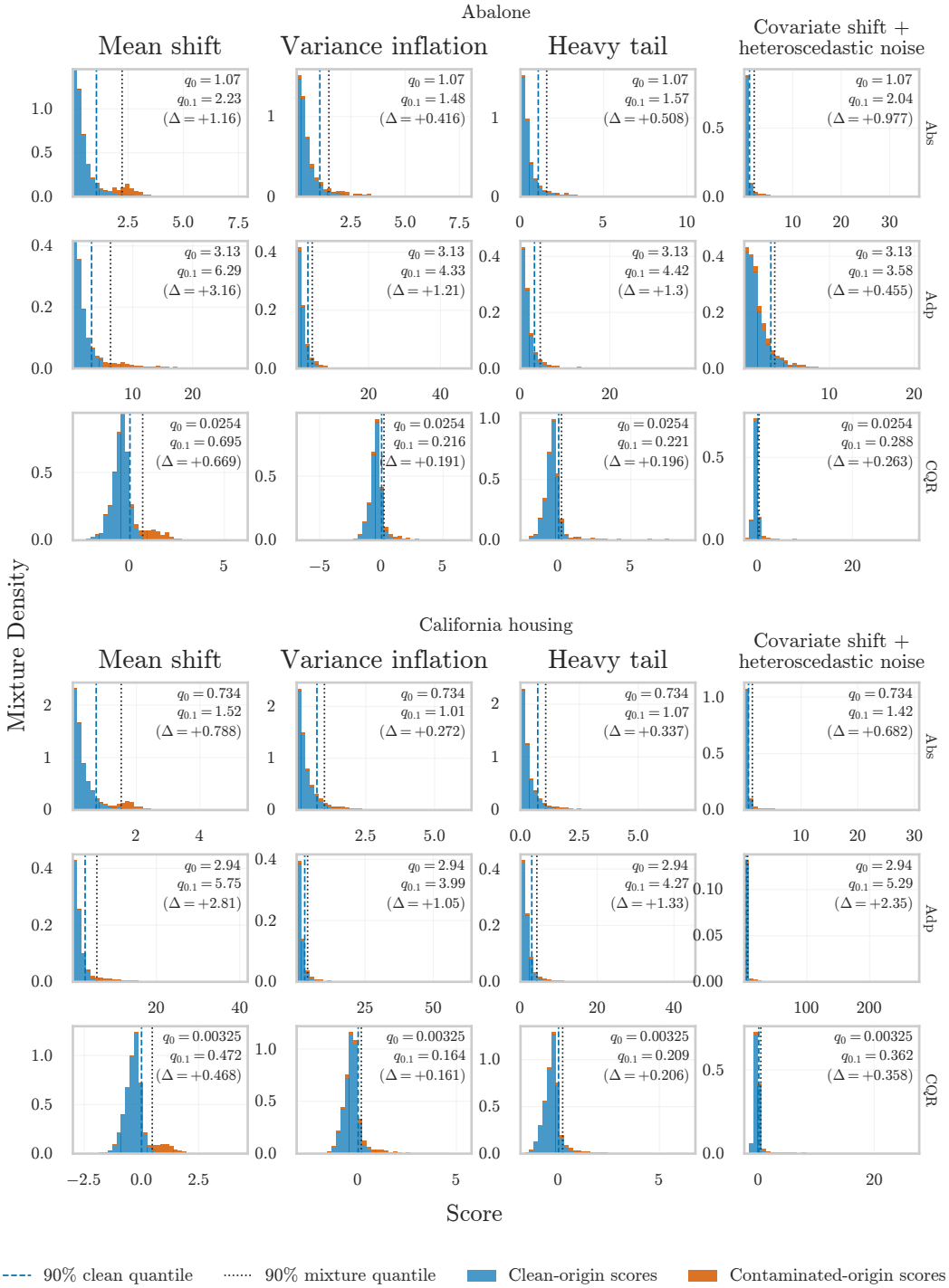


Figure 13: Score mixtures on Abalone (top) and California Housing (bottom), using repetition zero and $\delta_2 = 0.1$. Within each dataset, rows show Absolute, Adaptive, and CQR. Columns use mean shift $a = 5$, variance inflation $s = 5$, heavy tails $c = 10$, and covariate shift plus heteroscedastic noise $b = 8$. Stacked bars show the source components; vertical lines are conformal thresholds at $\alpha = 0.1$. Panel annotations report their differences.

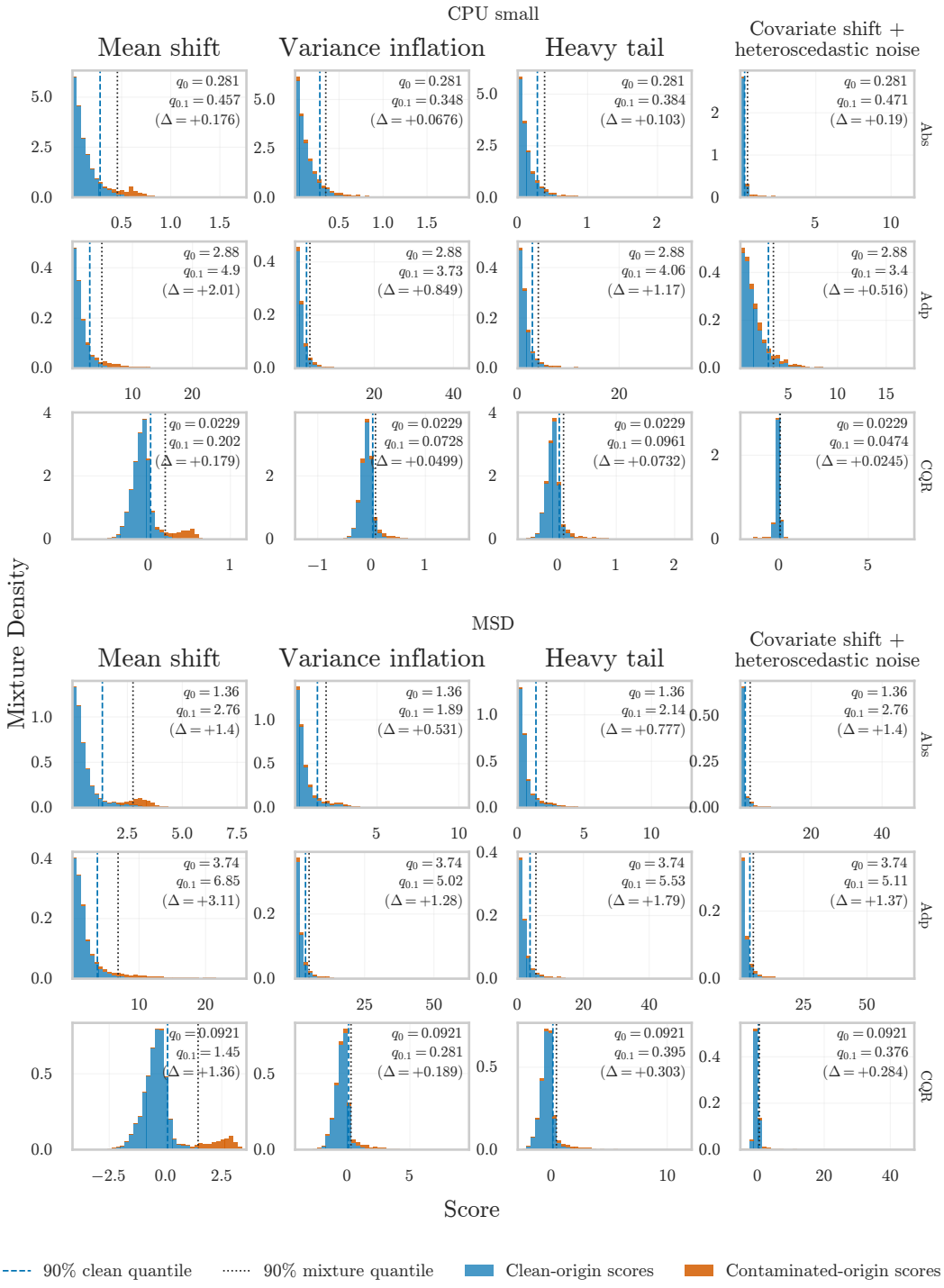


Figure 14: Score mixtures on CPU Small (top) and YearPredictionMSD (bottom), with the same ordering and conventions as Figure 13. Results use repetition zero at $\delta_2 = 0.1$.

E BERNOULLI AND FIXED-FRACTION REPLACEMENT

Conditional relationship. At each stage $j \in \{1, 2\}$, write $(n, \delta, Q) = (n_j, \delta_j, Q_j)$, where n, δ, and Q denote the sample size, replacement probability, and contamination distribution. Under Bernoulli replacement, independent indicators $U_i \sim \mathrm{Bernoulli}(\delta)$ mark the replaced observations, and $K = \sum_{i=1}^{n} U_i \sim \mathrm{Binomial}(n, \delta)$ counts the replacements. For $0 < \delta < 1$ and any $u \in \{0, 1\}^n$ with $\sum_i u_i = k$,

$$\mathbb{P}\big((U_1, \ldots, U_n) = u \mid K = k\big) = \binom{n}{k}^{-1}.$$

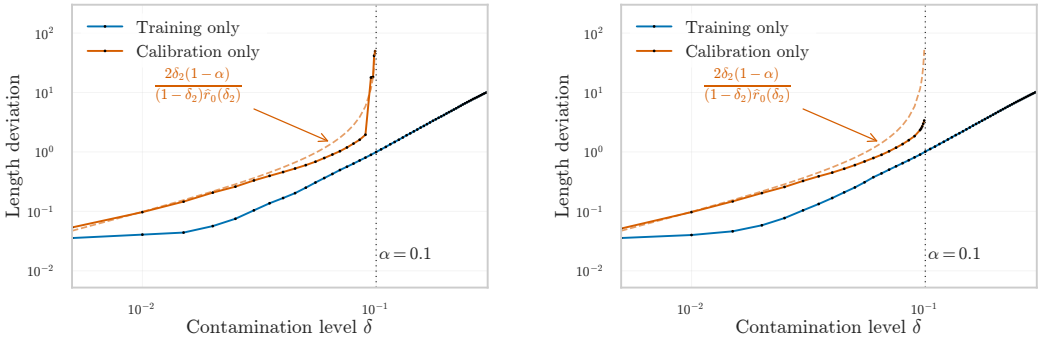
Thus, choosing exactly $k = \lfloor n\delta \rfloor$ locations uniformly and independently of the data and replacing those observations with draws from Q is precisely the Bernoulli replacement scheme conditional on $K = k$. The two stages use independent copies of this construction. At $\delta = 0$, both schemes leave the sample unchanged. Moreover, $K/n \to \delta$ in probability and $k/n \to \delta$, so the two schemes have the same limiting replacement fraction.

Interval-length bound. Let $k_j = \lfloor n_j \delta_j \rfloor$ denote the fixed replacement counts. Under the conditions of Theorem 4.1, for every $\gamma \in (0, 1)$ and all sufficiently large n_1, n_2, the fixed-fraction procedure satisfies, with probability at least $1 - \rho_{n_1}(P_0) - \gamma$,

$$2\,|\widehat{q}(\mathcal{D}_1; \mathcal{D}_2) - q(P_0; P_0)| \le 2\eta_{n_1}(P_0) + \frac{2Lk_1}{n_1} + \frac{2}{(1 - \delta_2) r_0} \left(\sqrt{\frac{\log(2/\gamma)}{2n_2}} + \frac{2}{n_2} + \delta_2(1 - \alpha) \right).$$

Indeed, replacement stability bounds the training perturbation by Lk_1/n_1 deterministically. Conditional on the selected calibration locations, the oracle scores are independent: $n_2 - k_2$ are clean draws, and k_2 are draws from the contaminating score distribution. Applying Hoeffding's inequality at the two thresholds bracketing the target quantile gives the sampling bound in Lemma 4.3 with replacement fraction k_2/n_2. Since $k_2/n_2 \le \delta_2$, the original density lower bound and population-bias bound remain valid. The clean-training sampling term is unchanged, which completes the same decomposition as in Theorem 4.1. The upper bound in Theorem 4.1 therefore remains valid, with the training count-fluctuation term removed.

Experimental comparison. We keep the original splits, contamination distributions, learners, and 20 repetitions. At each stage, the k_j smallest saved uniform draws determine the replacement locations; training-contaminated models are refitted. Figure 15 compares the two schemes in the setting of Figure 4(a). Using fixed counts removes the early rise caused by replacement-count fluctuations near $\delta_2 = \alpha$. Figure 16 complements this comparison with the four structured contamination settings in the Abalone dataset. The reference curves in these figures are empirical references, not certified finite-sample upper bounds.



(a) Bernoulli replacement.

(b) Fixed-fraction replacement.

Figure 15: Training-only and calibration-only mean length deviation relative to the clean population oracle, averaged over 20 repetitions. Both panels use the setting of Figure 4(a) and the same numerical reference for the population calibration term.

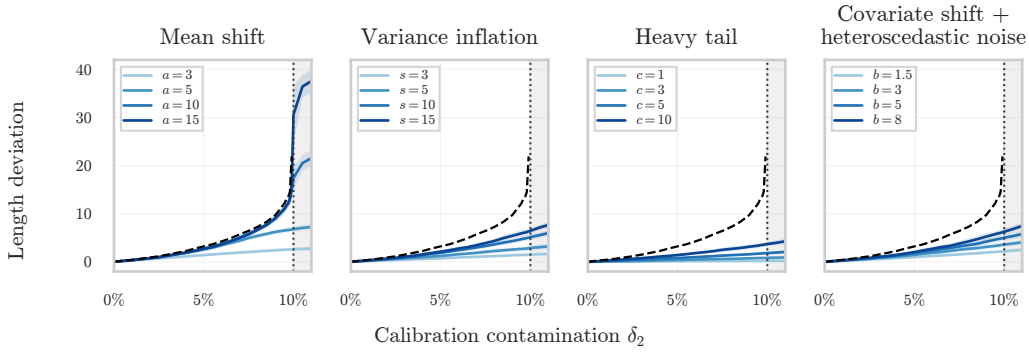


Figure 16: Fixed-fraction counterpart of Figure 5: calibration-only mean length deviation on Abalone with absolute-residual scores. Curves and shading show means and one standard deviation over 20 repetitions. The dashed curve is the finite score-200 Dirac mean plus one standard deviation for $\delta_2 < \alpha$.