Multi-Distribution Robust Conformal Prediction

Yuqi Yang and Ying Jin*

Abstract

In many fairness and distribution robustness problems, one has access to labeled data from multiple source distributions yet the test data may come from an arbitrary member or a mixture of them. We study the problem of constructing a conformal prediction set that is uniformly valid across multiple, heterogeneous distributions, in the sense that no matter which distribution the test point is from, the coverage of the prediction set is guaranteed to exceed a pre-specified level. We first propose a max-p aggregation scheme that delivers finite-sample, multi-distribution coverage given any conformity score associated with each distribution. Upon studying several efficiency optimization programs subject to uniform coverage, we prove the optimality and tightness of our aggregation scheme, and propose a general algorithm to learn conformity scores that lead to efficient prediction sets after the aggregation. We discuss how our framework relates to group-wise distributionally robust optimization, sub-population shift, fairness, and multi-source learning. In synthetic and real-data experiments, our method delivers valid worst-case coverage across multiple distributions, while greatly reducing the set size compared with naively applying max-p aggregation to single-source conformity scores. In some settings, our prediction sets are comparable in size to single-source prediction sets with popular, standard conformity scores.

1 Introduction

Reliable uncertainty quantification is critical for deploying machine learning systems in high-stakes domains [Platt et al., 1999, Gal et al., 2016, Guo et al., 2017, Lakshminarayanan et al., 2017, Kuleshov et al., 2018, Jiang et al., 2012, Kompa et al., 2021]. Conformal prediction is a powerful distribution-free framework for this purpose; given any prediction model, it offers prediction sets whose coverage guarantees hold without strong parametric assumptions on the data generating process [Vovk et al., 2005, Lei et al., 2018].

This paper studies how to maintain such reliability when models are deployed across multiple heterogeneous environments [Crammer et al., 2008, Mansour et al., 2008, Hashimoto et al., 2018, Romano et al., 2019a]. For example, a clinical risk prediction model trained on data from several hospitals must remain reliable when a new patient's record comes from one of these sites. Our goal is to construct prediction sets with valid coverage even when it is impossible to reveal where that patient came from.

Formally, we assume access to labeled data $\mathcal{D} = \cup_{k=1}^{K} \mathcal{D}^{(k)}$ from K heterogeneous sources, where each $\mathcal{D}^{(k)} = \{(X_i^{(k)}, Y_i^{(k)})\}_{i \in I_k}$ consists of i.i.d. samples from an unknown distribution $P^{(k)}$. Here $X_i^{(k)} \in \mathcal{X}$ is the features, and $Y_i^{(k)} \in \mathcal{Y}$ is the response. For a new test point (X_{n+1}, Y_{n+1}) drawn from one of these sources, we aim to build a prediction set $\hat{C}(X_{n+1}) \subseteq \mathcal{Y}$ with *uniform coverage*:

$$\min_{k \in [K]} \ \mathbb{P}_{\mathcal{D} \times P^{(k)}} \left(Y_{n+1} \in \hat{C}(X_{n+1}) \right) \geq 1 - \alpha, \tag{1}$$

where $\mathbb{P}_{\mathcal{D} \times P^{(k)}}$ denotes the joint distribution of the labeled data and a new test point $(X_{n+1}, Y_{n+1}) \sim P^{(k)}$. In words, $\hat{C}(\cdot)$ should achieve the nominal coverage level simultaneously for all possible sources.

*Department of Statistics and Data Science, University of Pennsylvania. Email: yjinstat@wharton.upenn.edu. Yuqi Yang is an undergraduate student in Mathematics and Computer Science, Hong Kong University of Science and Technology. Reproduction code for experimental results in the paper can be found in https://github.com/AragornBFRer/MDCP.

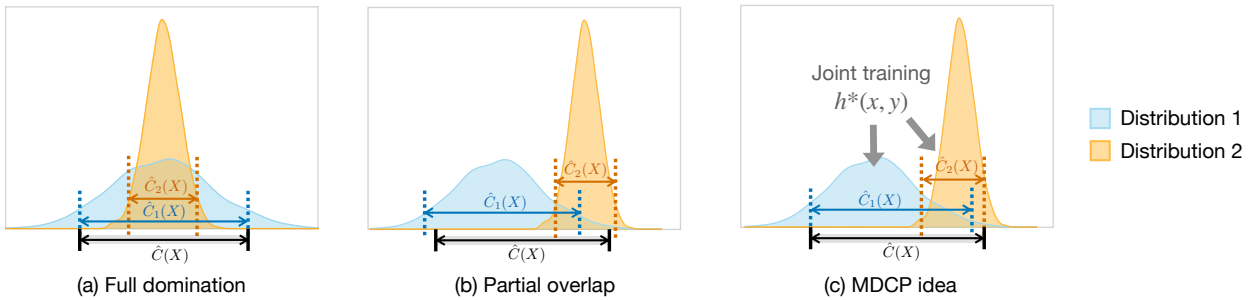


Figure 1: Prediction sets with uniform coverage need to balance the coverage across multiple distributions. (a) When one distribution *dominates* the other, a valid prediction set $\hat{C}(X)$ may coincide with the larger one $\hat{C}_1(X)$. (b) When two distributions *partially overlap*, a valid prediction set sits in between two distributions $\hat{C}(X)$ and is longer than single-distribution sets $\hat{C}_1(X)$ and $\hat{C}_2(X)$. (c) MDCP achieves uniform coverage by jointly training a conformity score and aggregating multiple prediction sets from the trained score.

Several practically important scenarios motivate such a guarantee:

- **Fairness without protected attributes.** Fair prediction across protected attributes such as race, gender, or socioeconomic status is a central goal of equitable machine learning [Madras et al., 2018, Hashimoto et al., 2018]. In this context, group-conditional coverage demands the coverage guarantee of the prediction sets to hold for all groups [Romano et al., 2019a, Jung et al., 2022, Gibbs et al., 2025]. However, existing methods require the group information (i.e., which distribution the new test point is from) to construct the prediction sets. In many sensitive scenarios, the group labels may be unavailable or ethnically protected [Gupta et al., 2018, Martinez et al., 2021, Lahoti et al., 2020], necessitating a single prediction set with coverage over all groups. Letting $P^{(k)}$ denote a sensitive group, uniform coverage (1) provides such a guarantee regardless of which group the test sample is from.
- **Subpopulation shift.** When each distribution represents a subpopulation, uniform coverage (1) protects against arbitrary subpopulation shift [Sagawa et al., 2019, Santurkar et al., 2020, Subbaswamy et al., 2021, Yang et al., 2023]. Formally, subpopulation shift assumes the training (labeled) data come from a mixture $P_{\text{train}} = \sum_{k=1}^{K} \pi_k P^{(k)}$, where each $P^{(k)}$ represents a subpopulation (e.g., hospitals, demographic groups, or regions). The test distribution is $P_{\text{test}} = \sum_{k=1}^{K} \pi'_k P^{(k)}$, with distinct mixture weights $\{\pi'_k\}_{k=1}^{K}$. Any prediction set satisfying (1) guarantees valid coverage under any such shift, since $\mathbb{E}_{P_{\text{test}}}(Y_{n+1} \in \hat{C}(X_{n+1})) = \sum_{k=1}^{K} \pi'_k \mathbb{P}_{P^{(k)}}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$ for any weights $\{\pi'_k\}_{k=1}^{K}$ that sum to 1.
- **Multi-source data.** Many scientific and engineering applications naturally aggregate heterogeneous datasets collected under different protocols or environments [Crammer et al., 2008, Mansour et al., 2008], which has attracted interests in conformal prediction as well [Lu et al., 2023, Spjuth et al., 2019, Liu et al., 2024]. Examples include hospitals with varying patient demographics, or satellite sensors operating under distinct conditions. Standard conformal prediction (calibrated on pooled data from all sites) is valid only for the mixture distribution induced by the training sample. In contrast, (1) offers guarantees to all individual sources, ensuring reliability even when the test data align with only one of them.

To achieve uniform coverage (1) with a single prediction set $\hat{C}(X_{n+1})$, one needs to balance the heterogeneous sources for reasonable efficiency (prediction set size). We demonstrate the efficiency-validity tension via two examples in panels (a-b) of Figure 1. In panel (a), distribution 1 (D1, whose "oracle prediction set is $\hat{C}_1(X)$) "dominates" distribution 2 (D2, whose "oracle" prediction set is $\hat{C}_2(X)$) with wider support and heavier tails. A uniformly valid $\hat{C}(X)$ thus has to over-cover under D2. In panel (b), the two distributions partially overlap and expand to opposite tails. A uniformly valid $\hat{C}(X)$ then must extend to both tail regions to satisfy (1) for each source, and its size often needs to strictly exceed that of either single-source set.

1.1 Preview of results

In this paper, we propose Multi-Distribution Conformal Prediction (MDCP), a general framework for constructing efficient prediction sets that achieve the uniform coverage guarantee (1) given per-source datasets.

We first propose a natural *max-p aggregation* mechanism to achieve uniform coverage: compute per-source conformal p-values using any conformity scores, then aggregate them by taking the maximum. Inverting this p-value leads to a prediction set that is the union of single-source prediction sets, which therefore guarantees finite-sample uniform coverage. While seeming conservative, such a max-p aggregation—taking the union of individually calibrated prediction sets—is necessary and nearly tight for worst-case coverage. To see why over-coverage for certain sources may be inevitable, in Figure 1(a), any prediction set with valid coverage under the "dominating" distribution must over-cover the other.

Situations like Figure 1(b), however, inspire flexible solutions: a strict subset $\hat{C}(X) \subset \hat{C}_1(X) \cup \hat{C}_2(X)$ may offer uniform coverage, which can again viewed as the union of some single-source prediction sets $\tilde{C}_1(X)$ and $\tilde{C}_2(X)$, although they are not the most "natural" single-source prediction set and whose construction needs to take the other source into account. This observation motivates us to learn *which single-source prediction sets to aggregate*, in order to build efficient aggregated sets (Figure 1c).

To this end, we analyze several optimization programs to theoretically characterize the optimal prediction sets with the minimal size/length subject to uniform coverage. These population-level programs reveal several messages on the optimality of the max-p aggregation scheme. In specific, there exists one single conformity score function $h^*: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which depends on a dual problem involving all the distributions under consideration, such that the max-p aggregation based on this score yields the optimally efficient prediction set. We then propose to learn such a score function and couple it with the max-p aggregation to produce the final MDCP prediction set. We show that our MDCP set achieves finite-sample uniform coverage (1), and asymptotically coincides with the oracle optimal prediction set under mild consistency conditions.

We implement MDCP for both classification and regression problems with general learning algorithms and practical training strategies. Finally, in extensive simulations and real applications to satellite imagery and medical service datasets, MDCP offers tight worst-case coverage while achieving improved efficiency than simple aggregation baselines, sometimes even comparable with single-source prediction sets.

We summarize our contributions as follows:

- We propose a general max-p aggregation scheme to achieve uniform coverage with single-source data.
- We establish the optimal form of prediction sets subject to uniform coverage, and show that the max-p aggregation achieves it with properly chosen conformity score functions.
- We propose an end-to-end pipeline to learn the conformity scores and construct efficient MDCP sets.
- We demonstrate the effectiveness of MDCP in extensive simulations and real data applications.

1.2 Related work

Distribution robustness and multi-source/group learning. This work is connected to a rich line of work on the robustness to distribution shift and heterogeneity across data sources. Classical domain adaptation and multi-source learning frameworks aim to generalize models across environments with distinct data-generating mechanisms [Crammer et al., 2008, Mansour et al., 2008, Ben-David et al., 2010]. In the modern ML setting, distributionally robust optimization (DRO) and group-DRO formulations seek to minimize the worst-case loss over predefined groups or uncertainty sets [Hashimoto et al., 2018, Sagawa et al., 2019, Santurkar et al., 2020, Lahoti et al., 2020, Martinez et al., 2020, Subbaswamy et al., 2021, Yang et al., 2023]. Our approach parallels this perspective but operates in the space of coverage guarantees rather than loss minimization: MDCP ensures that the coverage constraint holds for all source distributions, serving as a conformal analogue of group-DRO with exact finite-sample validity.

Group-conditional conformal prediction. Within conformal prediction, our setting is close to the group-conditional conformal prediction, sometimes framed for fairness [Romano et al., 2019a] and extended to multi-validity [Jung et al., 2022, Gibbs et al., 2025]. MDCP can be viewed as addressing a similar problem when a distribution represents a group-conditional distribution. In contrast, however, our method achieves so without observing the group label at test time, which can be particularly useful in sensitive scenarios with protected labels. As such, the techniques we develop are in sharp contrast to those methods.

Multi-source/distribution conformal prediction. Our setting is also connected to several recent work on conformal prediction from multiple data sources, where the goals vary, including learning with limited communications across sources [Lu et al., 2023], using other sources to improve efficiency in one source [Liu et al., 2024], aggregating individual prediction sets without data sharing [Spjuth et al., 2019]. In contrast, our goal is to leverage all data sources during training to construct a uniformly valid prediction set for a new test point from any source, leading to distinct techniques and guarantees.

2 Max-p conformal prediction

In this section, we introduce the general max-p aggregation scheme and establish its finite-sample uniform validity. Following the split conformal prediction framework [Vovk et al., 2005, Lei et al., 2018], we assume each data source $\mathcal{D}^{(k)}$ is randomly split to a training fold $\mathcal{D}^{(k)}_{\text{train}}$ and a calibration fold $\mathcal{D}^{(k)}_{\text{calib}}$. We assume $\{\mathcal{D}^{(k)}_{\text{train}}\}_{k=1}^K$ are used to obtain any conformity score function $s_k \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ associated with each distribution $k \in [K]$, which can be viewed as independent of the calibration folds and the test data.

Our method begins by calibrating single-source conformal p-values

$$p^{(k)}(y) := \frac{1 + \sum_{i \in \mathcal{D}^{(k)}_{\text{calib}}} \mathbb{1}\{s_k(X_i, Y_i) \le s_k(X_{n+1}, y)\}}{1 + |\mathcal{D}^{(k)}_{\text{calib}}|}.$$

From the standard conformal prediction theory, this p-value is valid for the k-th distribution, and inverting $p^{(k)}$ leads to a valid conformal prediction set for the k-th distribution:

$$\mathbb{P}_{\mathcal{D} \times P^{(K)}}\big(Y_{n+1} \in \hat{C}^{(k)}(X_{n+1})\big) \ge 1 - \alpha, \quad \text{where} \quad \hat{C}^{(k)}(X_{n+1}) = \big\{y \in \mathcal{Y} \colon p^{(k)}(y) \ge 1 - \alpha\big\}. \tag{2}$$

Our max-p aggregation scheme simply takes the maximum over the K p-values:

$$p(y) = \max_{k \in [K]} p^{(k)}(y),$$

and the prediction set is given by

$$\hat{C}(X_{n+1}) = \big\{y \in \mathcal{Y} \colon p(y) \ge \alpha\big\}. \tag{3}$$

It is straightforward to show that this prediction set is the union of single-source prediction sets in (2), and it enjoys finite-sample uniform validity. The proof of Theorem 1 is included in Appendix A.1.

Theorem 1 (Finite-sample uniform validity)**.** *Let $\{p^{(k)}(y)\}_{k=1}^K$ and $p(y)$ be defined above. Then, the aggregated set equals the union of the per-source conformal sets:*

$$\hat{C}(X_{n+1}) = \bigcup_{k=1}^K \hat{C}^{(k)}(X_{n+1}).$$

For an independent test point $(X_{n+1}, Y_{n+1}) \sim P$ with any mixture distribution $P = \sum_k \pi_k P^{(k)}$ and arbitrary weights $\sum_{k=1}^K \pi_k = 1$, $\pi_k \ge 0$, the prediction set achieves valid coverage

$$\mathbb{P}_{\mathcal{D} \times P}\big(Y_{n+1} \in \hat{C}(X_{n+1})\big) \ge 1 - \alpha,$$

which implies the uniform coverage (1) for any individual source as a special case.

Despite the generality and validity of this approach, several questions remain. The first is *tightness*. While enjoying uniform validity, a natural concern is that the aggregated set is larger than any single-source prediction set (each with valid coverage in at least one source); it is therefore unclear whether the coverage of (3) may turn out way above $1-\alpha$ whichever distribution the test point is from. The second is *efficiency*. How to choose the individual scores $\{s_k\}_{k=1}^K$ so that our prediction sets are of a reasonable size/length? Note that the individual scores determine the form of $\hat{C}^{(k)}(X_{n+1})$; if these sets do not overlap enough, their union set may be overly conservative. It is thus important to judiciously choose $\{s_k\}_{k=1}^K$, so that the resulting MDCP set concentrates on regions that matter for as many sources as possible. We answer these questions in the next section. Our results shed light on the choice of conformity scores and reveal that our aggregation scheme is both tight and optimal.

3 Optimality of max-p aggregation

In this part, we address the questions above by studying the optimality of our max-p aggregation. First, we solve several population-level optimization programs to derive the optimal form of prediction sets that achieve the smallest size/length subject to uniform coverage. Then, we show that our max-p aggregation is asymptotically equivalent to such optimal sets when the conformity scores converge to an "oracle" score.

3.1 Size optimization under uniform validity

We begin with the optimization problems of minimizing prediction set size/length subject to uniform validity (both marginal over X and conditional on X). Let $(\mathcal{X}, \mathcal{A}, \nu)$ and $(\mathcal{Y}, \mathcal{B}, \mu)$ be finite measure spaces with $\nu(\mathcal{X}) < \infty$ and $\mu(\mathcal{Y}) < \infty$, and write $\rho := \nu \otimes \mu$, where μ is the count measure for classification and the Lebesgue measure for regression. Throughout the paper, we assume that for each $k = 1, \ldots, K$, the covariate distribution $P_X^{(k)}$ admits a density $r_k(x)$ with respect to ν, and that $Y \,|\, X = x$ has density $f_k(\cdot \,|\, x)$ with respect to μ. For a measurable subset $C(X) \subseteq \mathcal{Y}$, we define $|C(X)|$ as the cardinality in classification problems when $|\mathcal{Y}| < \infty$, and the Lebesgue measure in regression problems when $\mathcal{Y} = \mathbb{R}$.[1]

Optimal prediction set under marginal validity. We consider the following optimization problem:

$$\begin{aligned} \underset{C(X)\subseteq\mathcal{Y},\ \text{measurable}}{\text{minimize}} \quad & \int_{\mathcal{X}} |C(X)|\, d\nu(x) \\ \text{subject to} \quad & \mathbb{P}^{(k)}(Y \in C(X)) \geq 1-\alpha, \quad \forall k = 1, \ldots, K. \end{aligned} \tag{4}$$

We integrate $|C(x)|$ over $\nu(\cdot)$ to ensure a scalar objective. By definition, (4) seeks the measurable prediction set with the smallest size that achieves uniform coverage. Rigorously speaking, by "measurable", we mean $\mathbb{1}\{y \in C(x)\}$ is a measurable function on $\mathcal{X} \times \mathcal{Y}$, or $C(x)$ is a measurable subset of $\mathcal{Y}$ for ν-a.s. all $x \in \mathcal{X}$.

Solving (4) amounts to a change-of-variable via the indicator function $I(x, y) := \mathbb{1}\{y \in C(x)\}$. For a clear presentation, we relax the range of $I(x, y)$ to $[0, 1]$, so that $I(x, y)$ can be viewed as the probability of $y \in C(x)$ for a randomized prediction set.[2] The optimization problem becomes

$$\begin{aligned} \underset{I(x,y)\in[0,1],\ \text{measurable}}{\text{minimize}} \quad & \iint_{\mathcal{X}\times\mathcal{Y}} I(x,y) d\rho(x,y) \\ \text{subject to} \quad & \iint_{\mathcal{X}\times\mathcal{Y}} I(x,y) r_k(x) f_k(y \,|\, x) d\rho(x,y) \geq 1-\alpha, \quad \forall k = 1, \ldots, K. \end{aligned} \tag{5}$$

[1] For clarity, we assume sufficient regularity of the underlying distributions so that the prediction sets under consideration are at least measurable with respect to the Lebesgue measure in regression problems.

[2] In settings where the distribution of $Y \,|\, X$ has point mass (e.g., in classification), random tie-breaking is needed for exact coverage in conformal prediction. In a similar spirit, relaxing the range of $I(x, y)$ to $[0, 1]$ allows us to characterize the complementary slackness without unnecessary complications.

Theorem 2 characterizes the globally optimal prediction set with smallest size subject to uniform validity, whose proof is in Appendix A.2. Here, the coverage probability should be understood as that of a randomized prediction set with probability $I(x, y) \in [0, 1]$.

Theorem 2 (Marginal optimality)**.** *Consider the marginal size-minimization problem (5). There exists a vector of nonnegative constants $\lambda^* = (\lambda_1^*, \ldots, \lambda_K^*) \in \mathbb{R}_+^K$ such that with $h_\lambda(x, y) = \sum_{k=1}^K \lambda_k \, r_k(x) \, f_k(y \,|\, x)$, one optimal solution C^* to (5) takes the following form:*

$$C^*(x) \;=\; \{\, y \in \mathcal{Y} \colon h_{\lambda^*}(x, y) > 1 \,\} \;\cup\; S(x), \qquad S(x) \subseteq \{\, y \in \mathcal{Y} \colon h_{\lambda^*}(x, y) = 1 \,\}.$$

In particular, $\lambda^* = (\lambda_1^*, \ldots, \lambda_K^*) \in \mathbb{R}_+^K$ is the optimal solution to the dual problem

$$\Phi(\lambda) = (1 - \alpha) \sum_{k=1}^K \lambda_k \;-\; \int_{\mathcal{X}} \int_{\mathcal{Y}} (h_\lambda(x, y) - 1)_+ \, d\mu(y) \, d\nu(x), \tag{6}$$

where $(h_\lambda(x, y) - 1)_+ = \max\{h_\lambda(x, y) - 1, 0\}$. Moreover, the complementary slackness holds:

(i) If $\lambda_k^* > 0$ then the k-th constraint is active, with $P^{(k)}(Y \in C^*(X)) = 1 - \alpha$;

(ii) If $\lambda_k^* = 0$ then the k-th constraint is (weakly) inactive, with $P^{(k)}(Y \in C^*(X)) \geq 1 - \alpha$;

(iii) There exists at least one $k^* \in [K]$ such that $\lambda_{k^*}^* > 0$ and $P^{(k^*)}(Y \in C^*(X)) = 1 - \alpha$.

If additionally $\mu(\{y : h_\lambda(x, y) = 1\}) = 0$ for ν-a.e. x, then C^* is unique up to $(\nu \otimes \mu)$-null sets.

Theorem 2 reveals the central role of a single score function $h_{\lambda^*}(x, y)$: the optimal solution $C^*(x)$ is determined by thresholding this score value. Whether to include the boundary set $\{y \colon h_{\lambda*}(x, y) = 1\}$ in the prediction set is subject to the user's preference. If $h_{\lambda^*}(x, y)$ does not have point mass over $\nu \otimes \mu$, the inclusion of the boundary set does not affect the average size or coverage probability. Otherwise, one need to randomize the inclusion to achieve exact $1 - \alpha$ coverage, or include it with slight over-coverage.

The complementary slackness in statement (iii) of Theorem 2 is worth noting: there always exists one source distribution under which $C^*(X)$ achieves exact $1 - \alpha$ coverage. (in the presence of point mass, such coverage needs to be understood as that of a randomized prediction set with $\mathbb{P}(y \in C^*(x)) = I^*(x, y) \in [0, 1]$).

We cautiously remark that since the objective of (4) integrates over the base measure $\nu(\cdot)$, the solution does not necessarily aim for the smallest average size for *the* test distribution in a specific problem. Arguably, it would be more natural to study a conditional optimality problem for a fixed $x \in \mathcal{X}$ subject to conditional uniform coverage, in which case the objective is inherently a scalar. In the remaining of the paper, we focus on the optimal solution to the conditional problem.

Optimal prediction set under conditional validity. Related to (1), a natural goal in practice is conditional coverage, i.e., the prediction set is uniform validity conditional on the observed feature [Foygel Barber et al., 2021, Jung et al., 2022, Gibbs et al., 2025]. The optimal solution to (4), which only aims for marginal uniform coverage, has a practical drawback: it may allocate the conditional coverage unequally across $\mathcal{X}$, with higher coverage in regions where it is easier to achieve high coverage with small prediction sets.

While conditional coverage is in general impossible to achieve in finite sample [Foygel Barber et al., 2021], it is still helpful to study the optimal form of prediction sets to design suitable conformity scores that approximate the conditional coverage in practice while maintaining marginal uniform coverage.

With the same setup as above, we study the following size optimization for any fixed $x \in \mathcal{X}$:

$$\begin{aligned} \underset{C}{\text{minimize}} \quad & |C(x)| \;=\; \int_{\mathcal{Y}} \mathbb{1}\{y \in C(x)\} \, d\mu(y) \\ \text{subject to} \quad & \int_{\mathcal{Y}} \mathbb{1}\{y \in C(x)\} \, f_k(y \mid x) \, d\mu(y) \;\geq\; 1 - \alpha, \quad k = 1, \ldots, K. \end{aligned} \tag{7}$$

The optimal objective of (7) (when marginalized over $\nu(X)$) is no smaller than that of (4), since its feasibility set is smaller (uniform conditional coverage implies uniform marginal coverage).

Theorem 3 offers the form of optimal prediction sets via the dual problem of (7). Again, to facilitate clean presentation, we relax the indicator function to any function bounded in $[0, 1]$.

Theorem 3 (X-conditional optimality)**.** *For a fixed $x \in \mathcal{X}$, there exists constants $\lambda^*(x) = (\lambda_1^*(x), \ldots, \lambda_K^*(x)) \in \mathbb{R}_+^K$ such that, with $h_\lambda(x, y) := \sum_{k=1}^K \lambda_k(x)\, f_k(y \,|\, x)$, an optimal solution to (7) takes the following form:*

$$C^*(x) \;=\; \big\{y \in \mathcal{Y} \colon h_{\lambda^*}(x, y) > 1\big\} \;\cup\; S(x), \qquad S(x) \subseteq \{y \in \mathcal{Y} \colon h_{\lambda^*}(x, y) = 1\}.$$

In particular, $\lambda^*(x) = (\lambda_1^*(x), \ldots, \lambda_K^*(x)) \in \mathbb{R}_+^K$ is the optimal solution to the dual problem

$$\Phi_x(\lambda(x)) = (1 - \alpha) \sum_k \lambda_k(x) - \int_{\mathcal{Y}} (h_\lambda(x, y) - 1)_+ \, d\mu(y), \tag{8}$$

where $(h_\lambda(x, y) - 1)_+ = \max\{h_\lambda(x, y) - 1,\, 0\}$. Moreover, complementary slackness holds for each k:

(i) If $\lambda_k^*(x) > 0$, then $P^{(k)}(Y_{n+1} \in C^*(x) \,|\, X_{n+1} = x) = 1 - \alpha$;

(ii) If $\lambda_k^*(x) = 0$, then $P^{(k)}(Y_{n+1} \in C^*(x) \,|\, X_{n+1} = x) \geq 1 - \alpha$.

(iii) There exists some $k^* \in [K]$ such that $\lambda_{k^*}^*(x) > 0$ and $P^{(k^*)}(Y_{n+1} \in C^*(x) \,|\, X_{n+1} = x) = 1 - \alpha$.

If additionally $\mu\big(\{y :\; h_{\lambda^*}(x, y) = 1\}\big) = 0$, then $C^*(x)$ is unique up to μ-null sets.

Theorem 3 reveals that the conditionally optimal prediction set also relies on thresholding a single score function at 1, which depends on the optimal dual variables $\lambda^*(x) \in \mathbb{R}_+^K$.

The boundary set $\{y \in \mathcal{Y} \colon h_{\lambda^*}(x, y) = 1\}$ is again challenging to characterize. If it has point mass (e.g., in classification), users may either randomize it to achieve exact $1 - \alpha$ coverage or include it for slight conservativeness. In the complementary slackness results, one should understand the coverage probability $P^{(k)}(Y_{n+1} \in C^*(x) \,|\, X_{n+1} = x)$ as randomizing the set with some probability $I^*(x, y) \in [0, 1]$.

3.2 Asymptotic optimality of max-p aggregation

Having studied the population-level optimal solutions, in what follows, we proceed to show that our max-p aggregation is indeed *optimal* and *tight*. Connecting Section 3.1 with max-p aggregation, our result states that, when using individual conformity scores that converge to an optimal score, the resulting prediction set converges to the optimal (while retaining finite-sample coverage due to Theorem 1).

As discussed, we focus on the (conceptually natural) conditional problem (7), and a similar result for the marginal problem follows from similar ideas. Let $\lambda^*(x) \in \mathbb{R}_+^K$ be a dual maximizer for the program (7), and let $h^*(x, y) := \sum_{k=1}^K \lambda_k^*(x) f_k(y|x)$. Recall that the oracle-optimal set $C^*(x)$ is of the form

$$C^*(x) = \{y : h^*(x, y) > 1\} \cup S^*(x), \quad \text{with } S^*(x) \subseteq T(x) := \{y : h^*(x, y) = 1\}. \tag{9}$$

The boundary set $T(x)$ is in general challenging to pinpoint: one may either randomize its inclusion to achieve exact $1 - \alpha$ coverage, or include $T(x)$ for a deterministic prediction set with slightly inflated coverage. In classification problems, such choices would affect the prediction set size since the size of $T(x)$ is non-negligible under the count measure. Therefore, to avoid over-complication, we stick to the generic form of $C^*(x)$ in (9) and isolate $S^*(x)$ in our results in what follows.

To describe the practical prediction set under max-p aggregation, we follow the procedure in Section 2. Splitting each source into training and calibration folds, we let $n_k = |\mathcal{D}_{\text{calib}}^{(k)}|$. Assuming access to estimators $\hat{f}_k^{(n)}(\cdot \,|\, \cdot)$ and $\hat{\lambda}^{(n)}(\cdot)$ obtained from $\cup_{k=1}^K \mathcal{D}_{\text{train}}^{(k)}$, we define $\hat{h}(x, y) := \sum_{k=1}^K \hat{\lambda}_k(x) \hat{f}_k(y \,|\, x)$, and use the conformity score $s_k(x, y) := -\hat{h}(x, y)$ to construct the prediction set (3). To simplify boundary conditions, we adopt

the randomized version of our general max-p aggregation which remains finite-sample valid. Specifically, for source $k \in [K]$ and a candidate label value $y \in \mathcal{Y}$ at a feature value $x \in \mathcal{X}$, we define the randomized p-value

$$p^{(k)}(x,y) := \frac{\sum_{i \in \mathcal{D}^{(k)}_{\text{calib}}} \mathbb{1}\{S_{k,i} > s_k(x,y)\} + (1 + \sum_{i \in \mathcal{D}^{(k)}_{\text{calib}}} \mathbb{1}\{S_{k,i} = s_k(x,y)\}) \cdot U_k}{n_k + 1}, \tag{10}$$

where $U_k \sim \text{Unif}([0,1])$ are i.i.d. and independent of everything else, and we write $S_{k,i} = -\hat{h}(X_i^{(k)}, Y_i^{(k)})$, and $s_k(x,y) = -\hat{h}(x,y)$. Finally, we construct our prediction set at level $\alpha \in (0,1)$ via

$$\hat{C}^{(n)}(x) := \{y : p(x,y) \geq \alpha\}, \qquad p(x,y) := \max_k \; p^{(k)}(x,y).$$

The superscript emphasizes the dependence on the sample size.

Theorem 4 provides a size gap guarantee between our aggregated set and the oracle set as $n_k \to \infty$, controlled by the tie-region size. The proof is in Appendix A.4.

Theorem 4. *Assume for each k, there exists constants $B_k > 0$ such that $\sup_{x,y} f_k(y|x) \leq B_k < \infty$, and $\sup_x \|\hat{\lambda}(x) - \lambda^*(x)\|_\infty \xrightarrow{p} 0$, and $\sup_{x,y} |\hat{f}_k(y|x) - f_k(y|x)| \xrightarrow{p} 0$ as $n_k, n \to \infty$, where $\sup_x \|\hat{\lambda}(x)\|_\infty \leq M$ (tight in probability) for a constant $M > 0$. Then we have $\sup_{x,y} |\hat{h}(x,y) - h^*(x,y)| \xrightarrow{p} 0$. Furthermore, let $T := \{(x,y) : h^*(x,y) = 1\}$ and write its measure as $\rho(T) = \int_T d\mu(y) d\nu(x)$. Then*

$$\limsup_{n \to \infty} \rho\big(\hat{C}^{(n)} \triangle \{(x,y) : h^*(x,y) \geq 1\}\big) \leq \rho(T).$$

Further, let $|C| := \int_X \mu(C(x)) d\nu(x)$, then for any optimal $C^* = \{(x,y) : h^*(x,y) > 1\} \cup S^*(x)$, we have

$$\limsup_{n \to \infty} \big|\, |\hat{C}^{(n)}| - |C^*| \,\big| \leq \rho(T).$$

Moreover, for ν-almost all fixed value of x, there exists a measurable subset $S_\infty(x) \subseteq T(x) \subseteq \mathcal{Y}$ such that there exists a subsequence $\{n^{(j)}\}$, along which

$$\rho\left(\hat{C}^{(n^{(j)})} \triangle \left(\{h^* > 1\} \cup S_\infty\right)\right) \xrightarrow{p} 0.$$

Consequently, if C^* is chosen with $S^*(x) = S_\infty(x)$, then $|\hat{C}^{(n^{(j)})}| \xrightarrow{p} |C^*|$ (even when $\rho(T) > 0$).

In words, Theorem 4 shows that, as long as our procedure in Section 2 is instantiated with individual score functions that are consistent for $-h^*(x,y)$, our prediction set under max-p aggregation is asymptotically equivalent to the oracle set $C^*(x)$ up to the boundary set $T(x)$ (whose inclusion may depend on practitioners' choice). This result has two key takeaways:

- First, it shows that max-p aggregation is *optimal* since it attains the oracle-optimal prediction set size with suitable choice of scores.
- Second, it shows the max-p aggregation is also *tight*: the oracle set C^* is shown in Theorem 3 to achieve exact $1 - \alpha$ coverage for at least one distribution, which implies that max-p aggregation provides (asymptotically) exact coverage for at least one source, even though it is larger than any of the single-source prediction set.

While we focus on the conditional problem throughout, we shall see that aiming for the conditionally-optimal prediction set typically leads to tight *marginal* worst-case coverage when evaluated over a specific test distribution in both simulations and real data experiments.

4 Practical algorithms

The above optimality results establish the conceptual foundations of implementing MDCP. In particular, a natural idea is learning to approximate the optimal scores $s_k(x,y) := -h_{\lambda^*}(x,y)$, and couple them with the max-p aggregation for finite-sample uniform validity. In this section, we develop the concrete algorithms for MDCP in classification and regression problems. At a high level, they proceed in three steps:

(i) First, we estimate per-source conditional models $\hat{f}_k(y \mid x)$ by a black-box model.

(ii) Then, we learn a covariate-dependent nonnegative weight vector $\hat{\lambda}(x) \in \mathbb{R}_+^K$ based on a dual problem.

(iii) Finally, we apply the max-p aggregation $s_k(x,y) := -\sum_k \hat{\lambda}_k(x)\hat{f}_k(y \mid x)$ to build the MDCP sets.

Our implementations focus on approximating the conditionally optimal score in Theorem 3, which relies on the conditional models $f_k(y \mid x)$ and the unknown dual variable functions $\{\lambda_k^*(x)\}_{k=1}^K$. One may implement similar ideas for the marginally optimal score in Theorem 2 by learning the scalars $\{\lambda_k\}_{k=1}^K$; however, it needs to begin with the marginal densities $f_k(x,y)$ which can be challenging to estimate.

In Section 4.1, we introduce the general dual objective that allows the estimation of $\{\lambda_k(x)\}_{k=1}^K$, and demonstrate the consistency of this approach under suitable conditions. We then present the concrete implementations for classification in Section 4.2 and for regression in Section 4.3, respectively.

4.1 Optimizing scores via an empirical dual objective

Section 3.2 motivates us to approximate the optimal solution $\lambda(\cdot)$ to the (integrated) dual problem

$$\Phi(\lambda) := (1-\alpha)\int_{\mathcal{X}} \sum_{j=1}^K \lambda_j(x) d\tilde{\nu}(x) - \int_{\mathcal{X}}\int_{\mathcal{Y}} \big(h_\lambda(x,y)-1\big)_+ d\mu(y) d\tilde{\nu}(x) \tag{11}$$

for a properly chosen distribution $\tilde{\nu}(\cdot)$, and recall that $\mu(\cdot)$ is the counting measure in classification and Lebesgue measure in regression. Theorem 5 is a simplified statement of a formal result in Appendix A.5.

Theorem 5 (Informal)**.** *For any $\tilde{\nu}(\cdot)$ that covers the support of $\mathcal{X}$ in the data, the optimal solution $\lambda^*\colon \mathcal{X} \to \mathbb{R}$ that maximizes $\Phi(\lambda)$ in (11) coincides with the dual solution $\lambda^*(x)$ given in Theorem 3.*

A convenient option is to take $\tilde{\nu}(\cdot)$ as the covariate distribution for the pooled dataset with all data sources. This leads to the empirical dual objective

$$\hat{\Phi}_{\text{marg}}(\lambda(\cdot)) := \frac{1}{N}\sum_{i=1}^N \Big[\big(1-h_\lambda(X_i,Y_i)\big)_-/\hat{p}_{\text{data}}(Y_i \mid X_i)\Big] + (1-\alpha)\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^K \lambda_j(X_i), \tag{12}$$

where $\hat{p}_{\text{data}}(y \mid x)$ estimates $p_{\text{data}}(y \mid x) = \sum_{k=1}^K w_k f_k(x,y) \,/\, \sum_{k=1}^K w_k f_k(x)$, the conditional density under the mixture of all data sources, and w_k is the fraction of the k-th source data among the pooled dataset. A natural idea is then to parameterize the function $\lambda(\cdot)$ within a function class $\mathcal{F}$, and solve the empirical risk minimization (ERM) problem based on $\hat{\Phi}_{\text{marg}}(\lambda(\cdot))$.

In the following, we present such a procedure using the method of sieves [Geman and Hwang, 1982]: we consider an increasing sequence $\Theta_1 \subset \Theta_2 \subset \cdots$ of spaces of smooth functions, and solve

$$\hat{\lambda}(\cdot) = \operatorname*{argmax}_{\lambda \in \Theta_n^K} \hat{\Phi}_{\text{marg}}(\lambda(\cdot)),$$

where $\hat{p}_{\text{data}}(\cdot)$ is a pre-trained model consistent for $p_{\text{data}}(\cdot)$. We consider two examples of sieves inspired by Yadlowsky et al. [2022], Jin et al. [2022].

Example 6 (Polynomials). *Let $Pol(J,\epsilon)$ be the space of J-th order polynomials on $[0,1]$ truncated at $\epsilon > 0$:*

$$Pol(J,\epsilon) = \Big\{x \mapsto \max\{\epsilon, \textstyle\sum_{j=0}^{J} a_j x^j\}\colon a_j \in \mathbb{R}\Big\}.$$

Then we define $\Theta_n = \Theta_{n,0}^K$, where $\Theta_{n,0} = \{x \mapsto \prod_{j=1}^d f_j(x_j)\colon f_j \in \mathrm{Pol}(J_n, 0), j = 1,\dots,d\}$ for $J_n \to \infty$.

Example 7 (Splines). *Let $0 = t_0 < \cdots < t_{J+1} = 1$ be knots that satisfy $\frac{\max_{0\le j\le J}(t_{j+1}-t_j)}{\min_{0\le j\le J(t_{j+1}-t_j)}} \le c$ for some $c > 0$. We define the space for r-th order truncated splines with J knots as*

$$Spl(r,J) = \Big\{x \mapsto \max\big\{\epsilon, \textstyle\sum_{k=0}^{r-1} a_k x^k + \sum_{j=1}^{J} b_j (x-t_j)_+^{r-1}\big\}\colon a_k, b_k \in \mathbb{R}\Big\}$$

Then we define $\Theta_n = \Theta_{n,0}^K$, where $\Theta_{n,0} = \{x \mapsto \prod_{j=1}^d f_j(x_j)\colon f_j \in \mathrm{Spl}(J_n, 0), j = 1,\dots,d\}$ for $J_n \to \infty$.

In both examples, we consider coordinate-wise function $\{f_j(x_j)\}_{j=1}^d$ for $\mathcal{X} \subseteq \mathbb{R}^d$ in a sieve series, so that $\prod_{j=1}^d f_j(x_j) \in \Theta_{n,0}$, and the optimal dual variables $\lambda^*(\cdot)\colon \mathcal{X} \to \mathbb{R}^K$ is approximated by elements in $\Theta_n = \Theta_{n,0}^K$. Here, we truncate the functions away from zero for simplicity. Note that if $\lambda_k^*(x)$ is always positive and continuous and $\mathcal{X}$ is a compact set, then there exists a positive $\epsilon > 0$ such that $\inf_{x\in\mathcal{X}} \lambda_k^*(x) \ge \epsilon$. In practice, we can set ϵ to be small enough, or let $\epsilon = \epsilon_n$ decays slowly to zero.

Next, we show that, if the true dual variables $\lambda_k^*(x)$ is sufficiently smooth in x, the solution $\hat\lambda(\cdot) \in \Theta_n$ by sieve estimation converges to $\lambda^*(\cdot)$. For $p_1 = \lceil p\rceil - 1$ and $p_2 = p - p_1$, we define

$$\Lambda_c^p = \left\{h \in C^{p_1}(\mathcal{X})\colon \sup_{\substack{x\in\mathcal{X}\\ \sum_{l=1}^d \alpha_l < p_1}} |D^\alpha h(x)| + \sup_{\substack{x\notin x'\in\mathcal{X}\\ \sum_{l=1}^d \beta_l = p_1}} \frac{|D^\beta h(x) - D^\beta h(x')|}{\|x-x'\|^{p_2}} \le c\right\}$$

To ensure non-negativeness, we also define the truncated function class $\Lambda_{c,+}^p := \{x \mapsto \max\{f(x), 0\}\colon f \in \Lambda_c^p\}$.

To be consistent with the language of ERM, we write the loss function and our estimator via

$$\hat\lambda(\cdot) = \operatorname*{argmin}_{\lambda(\cdot)\in\Theta_n} \hat{\mathbb{E}}_n\big[\hat\ell(\lambda(\cdot), X, Y)\big], \tag{13}$$
$$\text{where}\quad \hat\ell(\lambda(\cdot), x, y) = -(1 - h_\lambda(x,y))_-/\hat p_{\mathrm{data}}(y\,|\,x) - (1-\alpha)\textstyle\sum_{k=1}^K \lambda_k(x).$$

where (X_i, Y_i) are i.i.d. from the mixture distribution of the pooled data $p_{\mathrm{data}}(x,y)$. Note that this loss function relies on an estimator $\hat p_{\mathrm{data}}$. We denote the oracle minimizer and loss function as

$$\lambda^*(\cdot) = \operatorname*{argmin}_{\lambda\in\Theta=(\Lambda_{c,+}^p)^K} \mathbb{E}_{\mathrm{data}}[\ell(\lambda(\cdot), X, Y)], \tag{14}$$
$$\text{where}\quad \ell(\lambda(\cdot), x, y) = -(1 - h_\lambda(x,y))_-/p_{\mathrm{data}}(y\,|\,x) - (1-\alpha)\textstyle\sum_{k=1}^K \lambda_k(x).$$

Throughout, $\mathbb{E}_{\mathrm{data}}[\cdot]$ denotes the expectation under the pooled data distribution, and $\hat\ell(\cdot)$ is viewed as a fixed function. Recall that assuming the solution to the conditional optimality program (7) obeys $\lambda^* \in \Theta = (\Lambda_{c,+}^p)^K$, our results in Appendix A.5 ensures the above minimizer $\lambda^*(\cdot)$ in (14) coincides with the optimal solution we need; we thus use the same notations for them.

Similar to Jin et al. [2022, Theorem 1], we can show that the solution (13) that is close to the population minimizer $\bar\lambda^*$, which is further close to λ^* once the estimation error in $\hat p_{\mathrm{data}}$ is small. Our formal results build on the following two assumptions. Assumption 8 is a natural condition that the estimation error in $\hat p_{\mathrm{data}}$ translates to errors in population risk minimizer of the same order. Assumption 9 collects regularity conditions that are standard in the literature and hold for convex and smooth functions.

Assumption 8. *Assume $\|\lambda^* - \bar\lambda^*\|_{L_2} = O_P(\|\hat p_{\mathrm{data}} - p_{\mathrm{data}}\|_{L_2})$ and $\|\lambda^* - \bar\lambda^*\|_\infty = O_P(\|\hat p_{\mathrm{data}} - p_{\mathrm{data}}\|_\infty)$.*

Assumption 9. *Suppose $\mathcal{X} = \prod_{j=1}^{d} \mathcal{X}_j$ is the Catesian product of compact intervals, and $\theta^* \in \Theta = (\Lambda_c^p)^K$ for some $c > 0$. Suppose P_{data} has positive density on $\mathcal{X}$. We assume the function $\mathbb{E}_{data}[\hat{\ell}(\lambda, x, Y) \,|\, X = x]$ is η-strongly convex at $\bar{\lambda}^*(x)$ for all $x \in \mathcal{X}$. Also, $|\hat{\ell}(\theta, x, y) - \hat{\ell}(\bar{\lambda}^*, x, y)| \le \bar{\ell}(x, y)\|\theta(x) - \bar{\lambda}^*(x)\|_2$ for $\|\theta(x) - \bar{\lambda}^*(x)\|_2 < \epsilon$ for sufficiently small $\epsilon > 0$, where $\|\cdot\|_2$ is the Euclidean norm, and $\sup_{x \in \mathcal{X}} \mathbb{E}_{data}[\bar{\ell}(x, Y)^2] < M$ for some constant $M > 0$. Furthermore, there exists a constant C_1 such that $\mathbb{E}_{data}[\hat{\ell}(\theta, X, Y) - \hat{\ell}(\bar{\lambda}^*, X, Y)] \le C_1 \|\theta - \bar{\lambda}^*\|_{L_2}^2$ when $\theta \in (\Lambda_c^p)^K$ and $\|\theta - \bar{\lambda}^*\|_{L_2}$ is sufficiently small.*

Theorem 10 justifies using sieve approximation and ERM to learn the functions $\lambda^*(\cdot)$. Its proof largely follows Jin et al. [2022], and is included in Appendix A.6 for completeness.

Theorem 10. *Under Assumptions 8 and 9, We set $J_n = (\frac{\log n}{n})^{1/(2p+d)}$ for the sieve estimators in Examples 6 and 7, and suppose $\hat{\lambda} = \operatorname{argmin}_{\theta \in \Theta} \hat{\mathbb{E}}_{data}\big[\hat{\ell}(\theta, X, Y)\big]$. Then employing the function classes in the two examples, we have $\|\hat{\lambda} - \lambda^*\|_{L_2} = O_P\big((\frac{\log n}{n})^{p/(2p+d)}\big) + O_P(\|\hat{p}_{\text{data}} - p_{\text{data}}\|_{L_2})$ and $\|\hat{\lambda} - \lambda^*\|_\infty = O_P\big((\frac{\log n}{n})^{2p^2/(2p+d)^2}\big) + O_P(\|\hat{p}_{\text{data}} - p_{\text{data}}\|_\infty)$.*

Our results so far justify an empirical risk minimization approach with a plug-in estimator $\hat{p}_{\text{data}}(y \,|\, x)$ to learn the unknown components $\lambda_k^*(x)$ and approximate the optimal MDCP set. Suppose the true optimal dual variables $\lambda_k^*(x)$ in Theorem 3, as a function of x, is sufficiently smooth, and suppose we solve the ERM (13) with a suitable sieve function class based on a consistent estimator $\hat{p}_{\text{data}}(y \,|\, x)$, Theorem 10 ensures that the learned $\hat{\lambda}(\cdot)$ will converge to $\lambda^*(\cdot)$. Thus, as long as the conditional density functions $\hat{f}_k(\cdot \,|\, \cdot)$ are consistent, by Theorem 4, taking $s_k(x, y) = -\sum_{k=1}^{K} \hat{\lambda}_k(x) \hat{f}_k(y \,|\, x)$ yields an MDCP set that is asymptotically optimal. The next two subsections introduce the concrete algorithms that implement this in classification and regression problems.

4.2 Algorithm for classification

We now state the concrete MDCP algorithm for the classification setting. Recall that we split the labeled data into the training fold $\mathcal{D}_{\text{train}} = \cup_{k=1}^{K} \mathcal{D}_{\text{train}}^{(k)}$ and the calibration fold $\mathcal{D}_{\text{calib}} = \cup_{k=1}^{K} \mathcal{D}_{\text{calib}}^{(k)}$. For each source k, we fit a classifier on the training fold $\mathcal{D}_{\text{train}}^{(k)}$ for class probabilities $\hat{p}_k(y \,|\, x)$ for $y \in \mathcal{Y}$ by any off-the-shelf algorithm. Next, we learn $\lambda(x)$ via solving an empirical optimization objective. To model the covariate-dependent, nonnegative weights $\lambda_j(x)$, we use a spline function approximation. Let $\Lambda(x) \in \mathbb{R}^m$ denote the vector of spline basis functions evaluated at a covariate value x. In our implementation, $\Lambda(x)$ is taken to be a cubic B-spline basis with 3 polynomial degree and 5 interior knots placed uniformly over the range of the observed covariates $\{X_i\}$, together with an intercept term. This basis can be constructed using the `SplineTransformer` in the `scikit-learn` Python package, which returns the design matrix $\Lambda = \Lambda(X_{\text{train}}) \in \mathbb{R}^{n \times m}$, where each row $\Lambda(x_i)^\top$ corresponds to the spline features for observation x_i. For K latent components, we collect the spline coefficients into a matrix $\Theta \in \mathbb{R}^{K \times m}$, with row $\theta_j^\top$ parameterizing the j-th weight function. We then define

$$\lambda_k(x; \Theta) \;=\; \text{softplus}\big(\Lambda(x)^\top \theta_k\big), \qquad k = 1, \ldots, K, \tag{15}$$

where $\text{softplus}(t) = \log(1 + e^t)$ is applied elementwise to guarantee $\lambda_k(x) \ge 0$. Accordingly, the score function with parameter Θ is $h(x, y; \Theta) := \sum_{k=1}^{K} \lambda_k(x; \Theta) \cdot \hat{p}_k(y \,|\, x)$.

We fit the parameters $\hat{\Theta}$ by maximizing the Lagrangian-inspired empirical objective in Section 4.1. For a miscoverage level α, the objective as a function of Θ is given by

$$\hat{\mathbb{E}}_{\mathcal{D}_{\text{train}}}\left[\frac{(1 - h(X, Y; \Theta))_-}{\hat{p}_{\text{data}}(Y \,|\, X)}\right] + (1 - \alpha)\hat{\mathbb{E}}_{\mathcal{D}_{\text{train}}}\left[\sum_{k=1}^{K} \lambda_k(X; \Theta)\right], \tag{16}$$

where $\hat{\mathbb{E}}_{\mathcal{D}_{\text{train}}}[\cdot]$ denotes the empirical average across the pooled training fold, $(t)_- = \min\{t, 0\}$ denotes the negative part, and $\hat{p}_{\text{data}}(y \,|\, x)$ is an estimator for $p_{\text{data}}(y \,|\, x)$. For fast implementation, we directly

assemble the per-source models to form the mixture probability estimator $\hat{p}_{\text{data}}(y \mid x) = \sum_k \hat{w}_k \hat{p}_k(y \mid x)$ with the marginal weights $\hat{w}_k = |\mathcal{D}^{(k)}_{\text{train}}| / \sum_\ell |\mathcal{D}^{(\ell)}_{\text{train}}|$, avoiding another model fit. In practice, one may also fit $p_{\text{data}}(y \mid x)$ by simply running a classification algorithm on the pooled dataset.

Finally, given the fitted parameters $\hat{\Theta}$, we define the score function

$$s_k(X_i, Y_i) := -\sum_{k=1}^{K} \lambda_k(X_i; \hat{\Theta}) \hat{p}_k(Y_i | X_i),$$

and use them to calibrate the final prediction sets following (3) or the procedure outlined in Section 3.2. The entire procedure is summarized in Algorithm 1, which also covers regression problems below.

Algorithm 1 Multi-Distribution Conformal Prediction (MDCP)

Input: Data $\mathcal{D} = \cup_{k=1}^{K} \mathcal{D}^{(k)}$ from K sources, test input X_{n+1}, significance level α, problem `mode`.

Split the data $\mathcal{D}$ into $\mathcal{D}_{\text{train}} = \cup_{k=1}^{K} \mathcal{D}^{(k)}_{\text{train}}$ and $\mathcal{D}_{\text{calib}} = \cup_{k=1}^{K} \mathcal{D}^{(k)}_{\text{calib}}$.
`// Train per-source models`
for $k = 1$ **to** K **do**
  if `mode` = classification **then**
    Fit any classifier $\hat{p}_k(y \mid x)$ on $\mathcal{D}^{(k)}_{\text{train}}$.
  else if `mode` = Regression **then**
    Fit conditional density estimator $\hat{f}_k(y|x)$ on $\mathcal{D}^{(k)}_{\text{train}}$ via, e.g., conditional gaussian model.
  end if
end for
`// Fit Lagrange multiplier` $\lambda(\cdot)$
Solve the empirical objective (16) on $\mathcal{D}_{\text{train}}$ to obtain spline parameters $\hat{\Theta}$.
`// MDCP set on test point` x
if `mode` = classification **then**
  Set $s_k(x, y) = -\sum_{k=1}^{K} \lambda_k(x; \hat{\Theta}) \hat{p}_k(y \mid x)$ via (15).
  Compute $s_k(X_{n+1}, y)$ for all $y \in \mathcal{Y}$, and $p^{(k)}(y)$ with $\mathcal{D}^{(k)}_{\text{calib}}$ using (10) for $k = 1, \ldots, K$.
  Compute $\hat{C}(x) = \{y : p(y) \geq \alpha\}$ with $p(y) = \max_k p^{(k)}(y)$.
else if `mode` = regression **then**
  Set $s_k(x, y) = -\sum_{k=1}^{K} \lambda_k(x; \hat{\Theta}) \hat{f}_k(y \mid x)$ via (15).
  Generate y-grid and use a grid search to construct prediction set $\hat{C}(X_{n+1})$ (Appendix B.3).
end if

Output: Prediction set $\hat{C}(X_{n+1})$.

4.3 Algorithm for regression

For regression problems, the data splitting, parameterization and estimation of $\lambda(x)$ are similar to the classification case. The key difference is in fitting the conditional density $f_k(y \mid x)$. While one is free to choose any appropriate estimator, here we model $Y = \mu(X) + \sigma(X) \cdot \epsilon$ for some $\epsilon \sim N(0, 1)$. Then, we use nonparameteric methods, such as gradient boosting, to estimate $\mu(x)$ and $\sigma(x)$ using each $\mathcal{D}^{(k)}_{\text{train}}$; see Appendix B.2 for a detailed estimation procedure.

Given the estimators $\hat{\mu}(x)$ and $\hat{\sigma}(x)$, our working conditional model is $\hat{f}_k(y \mid x) \propto \exp(-(y - \hat{\mu}(x))/(2\hat{\sigma}(x)^2)$. In the single-source case, this reduces to the prediction set proposed by Lei et al. [2018]. We then follow the same parameterization and training objective as in the classification case to obtain the estimated spline parameters $\hat{\Lambda}$ and the scores $s_k(x, y) = -\sum_{k=1}^{K} \lambda_k(x; \hat{\Theta}) \hat{f}_k(y \mid x)$, which are then used to calibrate the single-source p-values (10) and the corresponding MDCP set in the same way as in Section 4.2.

Finally, we note that thresholding the learned score function $s_k(x, y) = -\sum_{k=1}^{K} \lambda_k(x; \hat{\Theta}) \hat{f}_k(y \mid x)$ does not necessarily lead to an interval MDCP set. However, as we model $\hat{f}_k(y \mid x)$ as a normal distribution, the

MDCP set must be the union of at most K intervals. This structure allows us to compute a super-set of our MDCP set via a grid search. For brevity, we defer the details and justifications to Appendix B.3.

5 Simulations

We assess the validity and efficiency of our algorithms in diverse synthetic settings in both classification and regression, and investigate how the separation among sources impact the performance.

5.1 Simulation settings

We begin by outlining the common setup in both classification and regression settings. We consider $K = 3$ sources, a feature dimension of $d = 10$, and a target miscoverage level $\alpha = 0.1$. Across all experiments, we generate $n_k = 2000$ labeled samples from each source, which are randomly split into training (75%) and calibration (25%) folds. The test samples are of the same size in each source for evaluation. Across all settings, the features are generated by $X_i^{(k)} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.2 + 0.8\,\mathbb{1}\{i = j\}$, which are then standardized per source so that each coordinate has zero mean and unit variance. With identical feature distribution, the heterogeneity across sources stem from the distinctions in the conditional distributions of the label. In each run of the experiments, we draw a signal set $\mathcal{I} \subset \{1, \ldots, d\}$ of size $|\mathcal{I}| = 4$ uniformly at random, so the labels depend on X only through $X_{\mathcal{I}}$. We examine three suites of experiments:

(a). `Linear`: In this suite, the labels are generated by a linear function of $X_{\mathcal{I}}$.

(b). `Nonlinear`: In this suite, we keep the same feature distribution and noise structure as in the linear case, but let the source-specific conditional mean functions be nonlinear in x.

(c). `Temperature`: This final suite focuses on the linear setting where a "temperature" parameter τ controls the degree of heterogeneity or separation across sources.

The specific data generating processes are given in Section 5.2 and Section 5.3 for classification and regression settings, respectively. In each setting, we evaluate three competing methods:

(i). `Baseline-src-`k: The standard conformal prediction set $\hat{C}_{\text{src-}k}$ with calibration data from source k.

(ii). `Baseline-agg`: A simple max-p aggregation of per-source prediction sets $\hat{C}_{\text{max-p}} := \cup_{k=1}^{K} \hat{C}_{\text{src-}k}$. This is the baseline without efficiency-oriented score learning.

(iii). `MDCP`: Our method in Algorithm 1.

For each configuration, we repeat the experiments for $N = 100$ times and report the mean results. For fair comparison, all the methods build on the same conditional mean and standard deviation estimators to be specified later in a similar fashion. With these choices, the single-source baseline is standard conformal prediction sets with the widely-used APS score [Romano et al., 2020] in classification and the variance-adaptive score of Lei et al. [2018] in regression problems.

5.2 Simulations in classification settings

Data generating processes. We simulate $C = 6$ classes. For source $k \in [K]$ and class $c \in [C]$, the conditional class probability is given by the multinomial model $f_k(y = c \,|\, x) = \frac{\exp\{\eta_{kc}(x)\}}{\sum_{c'=1}^{C} \exp\{\eta_{kc'}(x)\}}$ with $\eta_{kc}(x) = \lambda_k(b_{kc} + \beta_{kc}^\top x) + g(x)$. Here, with a temperature parameter $\tau \in \mathbb{R}$, the linear signal strength is given by $\lambda_k = 2.5(1 + 0.25\tau \cdot u_k)$ with $u_k \overset{\text{i.i.d.}}{\sim} \text{Unif}([-1, 1])$, and the heterogeneous intercept is independently sampled as $b_{kc} \sim \mathcal{N}(0, (0.4\tau)^2)$. The source-specific linear coefficients are given by $\beta_{kc} = \bar{\beta}_c + \tau \cdot \Delta_{kc}$ where, after a random sample of signal sets $\mathcal{I} \subseteq [d]$ with $|\mathcal{I}| = 4$, we independently sample $(\bar{\beta}_c)_j \sim \mathcal{N}(0, 1)$ and $(\Delta_{kc})_j \sim \mathcal{N}(0, 0.15^2)$ for each $j \in \mathcal{I}$ and set $(\bar{\beta}_c)_j = (\Delta_{kc})_j = 0$ for $j \notin \mathcal{I}$. Finally, the nonlinear component

$g(x)$ is set as zero in the `Linear` experiments, and we vary its definition in three data generating processes in the `Nonlinear` experiments:

$$g(x) = \begin{cases} 2\sum_{(u,v)} w_{uv}\, x_u x_v, & \text{(interaction)}, \\ -2\sum_{r=1}^{3} a_r \sin(u_r^\top x + b_r), & \text{(sinusoid)}, \\ -2\sum_{r=1}^{3} a_r \log\big(1+\exp(u_r^\top x + b_r)\big), & \text{(softplus)}, \end{cases}$$

For the three families, the weights w_{uv}, the linear coefficients a_r, u_r, and b_r are randomly sampled at the beginning of each experiment for which the detailed processes are deferred to Appendix B.1, after which the labeled and unlabeled data are drawn conditional on them.

In the `Linear` and `Nonlinear` experiments, the temperature parameter is fixed at $\tau = 2.5$. In the evaluation of `Temperature` experiments, we focus only on the linear model with $g(x) \equiv 0$ and vary the temperature $\tau \in \{0.5, 1.5, 2.5, 3.5, 4.5, 5.5\}$.

Method implementations. As we mentioned earlier, the three competing methods are based on the same estimators (built from the training folds) for fair comparison. We first use a gradient boosting classifier construct an estimator $\hat{p}_k(y \,|\, x)$ for the per-source conditional class probability $P^{(k)}(Y = y \,|\, X = x)$. Following Section 4.2, we specify the per-source scores in MDCP via

$$s_k(X_i, Y_i) := -\hat{h}_\lambda(X_i, Y_i) = -\textstyle\sum_{k=1}^{K} \hat{\lambda}_k(X_i)\hat{p}_k(Y_i|X_i),$$

where, following the procedure in Section 4.2, we parameterize the nonnegative weight functions $\lambda_k(x)$ as spline functions, and learn $\hat{\lambda}_k(x)$ by minimizing the empirical objective (12). To facilitate fast implementation, we $\hat{p}_{\text{data}}(y \mid x) = \sum_{k=1}^{K} \hat{w}_k\, \hat{p}_k(y \,|\, x)$ with $\hat{p}_k(y \,|\, x)$ from the trained classifier, and $\hat{w}_k$ is the fraction of the k-th source data in the pooled dataset. The multipliers $\hat{\lambda}_k(x)$ are trained on the same training fold based on the fitted classifiers $\hat{p}_k(y \,|\, x)$, i.e., we reuse the training data for both stages. In both `Baseline-src-`k and `Baseline-agg`, we use the widely-used APS score [Romano et al., 2020] with the same fitted probabilities $\hat{p}_k(y \,|\, x)$ to build single-source and aggregated prediction sets, thereby serving as comparable baselines with the same fitted models without optimizing for multi-distribution efficiency.

Simulation results. Figure 2 presents the comparison between MDCP and two baselines in the `Linear` setting, in terms of average coverage, worst-case coverage, and average set size. The single-source sets lead to severe under-coverage. Due to max-p aggregation, both `Baseline-agg` and `MDCP` achieve valid worst-case coverage, yet `MDCP` delivers (i) significant efficiency improvement, and (ii) tight worst-case coverage due to (approximate) complementary slackness (Theorem 3). Relative to the max-p baseline, MDCP yields on average a 24.90% smaller set. By contrast, the max-p aggregation baseline is conservative, achieving 96.96% global and 95.44% worst-case coverage. MDCP is also more stable: the standard deviation of set size is 9.92% lower then the max-p baseline. Although we focused on the computationally convenient conditional optimal formulation, the complementary slackness is still quite strong when the coverage is evaluated marginally.

Figure 3 presents the results in the `Nonlinear` settings. Single-source calibration achieves valid coverage in the softplus setting but severely under-covers in other ones. In contrast, `MDCP` maintains tight worst-case coverage across all settings. `MDCP` again produces much smaller prediction sets relative to `Baseline-agg`. Notably, in the only softplus setting where single-source calibration is valid, MDCP achieves tighter worst-case coverage and substantially smaller set sizes, showing the benefits of both max-p aggregation and efficiency optimization: MDCP is able to adaptively concentrate coverage on regions with strong overlap across sources.

We further investigate the effect of the temperature parameter τ in Figure 4. As τ increases and the individual distributions (of the labels and conformity scores) move farther apart, maintaining valid coverage across sources becomes more challenging. Consistent with this intuition, the coverage of single-source baseline declines as τ grows. Also, the signal-to-noise ratio increases with τ, making the task progressively harder. Nevertheless, MDCP maintains tight worst-case coverage and substantial efficiency gain over `Baseline-agg`.

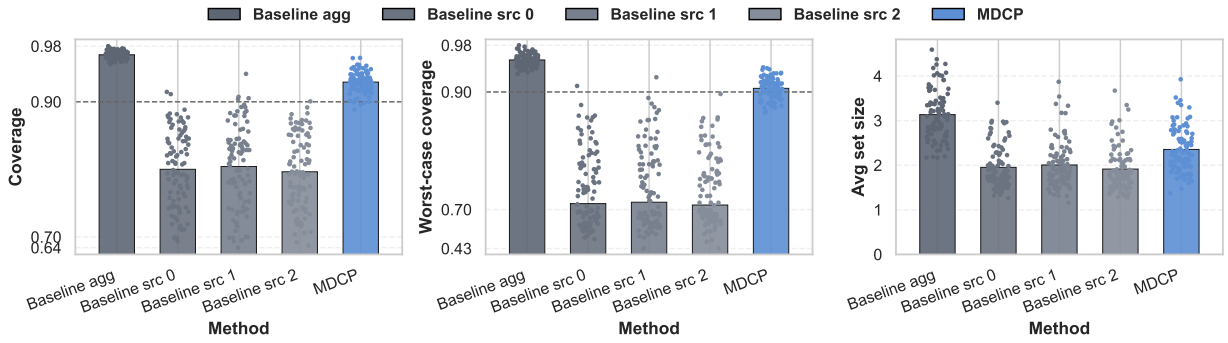


Figure 2: Performance of MDCP and baselines in the classification `Linear` experiments, where the bars represent the result of each method averaged over $N = 100$ runs, and the dots represent the result in each run. Left: coverage over all test data. Middle: worst-case coverage over single-source test data. Right: average set size over all test data.

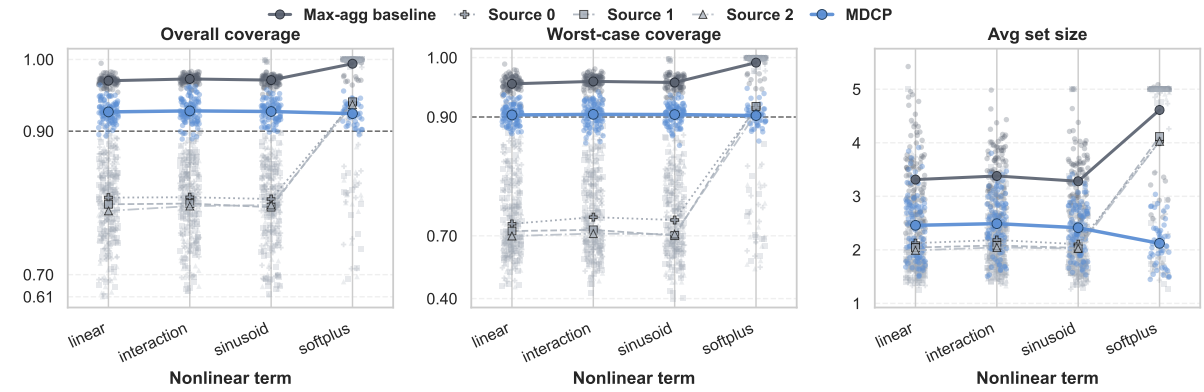


Figure 3: Performance of MDCP and baselines in the classification `Nonlinear` experiments. The x-axis is the setting of the nonlinear term $g(x)$, with the linear setting presented for comparison. The connected dots are average results colored by method, with the colored, dimmed dots being the results in each of the $N = 100$ runs. Left: coverage over all test data. Middle: worst-case coverage over single-source test data. Right: average set size over all test data.

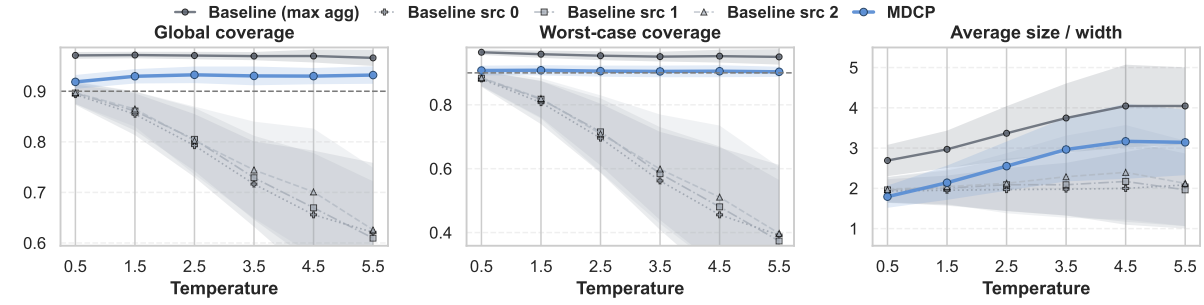


Figure 4: Performance of MDCP and baselines in the classification `Temperature` experiments. The x-axis the value of the temperature parameter τ. Each line shows the results of a method averaged over $N = 100$ runs, with shaded 95% confidence intervals. Left: coverage over all test data. Middle: worst-case coverage over single-source test data. Right: average set size over all test data.

Additional ablation study. Besides the main results, we conduct a suite of ablation study on numerical stability of the optimization process; see Appendix B.4 for details. Specifically, we consider a penalized version of the objective 16 which penalizes certain norms of Θ, and investigate a variant of Algorithm 1 that adaptively optimizes for the penalty parameter in the training process. Across all the simulation settings, the improvement of this approach is negligible, which shows that the current optimization is stable enough.

5.3 Simulations in regression problems

Data generating processes. In all regression settings, for source $k \in [K]$, we sample the labels via $Y = \mu_k(X) + \varepsilon_k$, with independent noise $\varepsilon_k \sim \mathcal{N}(0, \sigma_k^2)$. Following similar design ideas as in the classification settings, given a temperature parameter $\tau \in \mathbb{R}$, the regression function is $\mu_k(x) = \beta_k^\top x + b_k + g(x)$, where the source-specific coefficient is given by $\beta_k = \bar{\beta} + 0.2\tau \cdot \delta_k$, with $\bar{\beta}_j \sim \mathcal{N}(0,1)$ and $(\delta_k)_j \sim \mathcal{N}(0,1)$ independently drawn for $j \in \mathcal{I}$ and $\bar{\beta}_j \equiv 0$ and $(\delta_k)_j \equiv 0$ for $j \notin \mathcal{I}$, and we recall that $\mathcal{I}$ is the randomly drawn set of signals. The source-specific intercept is given by $b_k = b + \tau \cdot v_k$ with independently drawn $b \sim \mathcal{N}(0, 0.5^2)$ and $v_k \sim \mathcal{N}(0, 0.5^2)$. In each run, we randomly sample a signal-to-noise ratio in $[5, 10]$, and achieve it by adjusting the noise variance σ_k^2. Finally, the nonlinear component $g(x)$ is set to be zero in the `Linear` experiments, and we consider the same three choices of $g(x)$ in the `Nonlinear` experiments as in the classification settings (Section 5.2), with the same sampling process of the hyperparameters.

Again, we fix the temperature $\tau = 2.5$ in `Linear` and `Nonlinear` experiments. In `Temperature` setting, we focus only on the linear model and vary the temperature $\tau \in \{0.5, 1.5, 2.5, 3.5, 4.5, 5.5\}$; in addition, we sample $u \sim \mathrm{Unif}([\{1-\tau/4\}_+, 1+\tau/4])$ and multiply the SNR-calibrated σ_k by u to also let the temperature increase the noise across experiments. Again, in each run of the experiment, we sample all the hyperparameters once, and generate the multi-source data conditional on them.

Method implementations. In the regression procedure, the optimal score function relies on the condition density $f_k(y \,|\, x)$ in each source k. As mentioned in Section 4.3, to avoid the challenging conditional density estimation, we model the data as $P^{(k)}_{Y \,|\, X=x} \sim \mathcal{N}(\mu_k(x), \sigma_k(x))$ for some mean function $\mu_k(x) = \mathbb{E}^{(k)}[Y \,|\, X = x]$ and conditional standard deviation function $\sigma_k^2(x) = \mathrm{Var}^{(k)}(Y \,|\, X = x)$, and obtain their estimates $\hat{\mu}_k(\cdot)$ and $\hat{\sigma}_k(\cdot)$ via gradient boosting decision trees. Plugging in the two estimates leads to the estimated per-source conditional densities $\hat{f}_k(y \mid x)$ using gradient boosting decision trees. Following the optimality results, the conformity score for MDCP is then given by

$$s_k(X_i, Y_i) := -\textstyle\sum_{k=1}^K \hat{\lambda}_k(X_i) \hat{f}_k(Y_i | X_i),$$

where we parameterize the nonnegative weight vector $\lambda(x) \in \mathbb{R}_+^K$ as a spline function and learn $\hat{\lambda}_k(x)$ by minimizing the empirical objective (12). We use the training fold to estimate $\hat{\mu}_k$ and $\hat{\sigma}_k$, and use the estimated values to learn the dual optimizers $\hat{\lambda}_k(x)$ on the same training fold. Similar to classification, we set the marginal $\hat{f}_k(x) \equiv 1$, and then $\hat{f}_{\mathrm{data}}(y \mid x) = \sum_{k=1}^K \hat{w}_k \hat{f}_k(y|x)$, where $\hat{f}_k(y|x)$ is given by the trained regressor. Both baselines use the conformity score $V_k(x, y) = (y - \hat{\mu}_k(x))/\hat{\sigma}_k(x)$ with the same estimated functions as in MDCP, paralleling the method in [Lei et al., 2018].

Simulation results. Figure 5 shows the performance of the competing methods in the `Linear` settings. MDCP achieves tight worst-case coverage due to (approximate) complementary slackness, while the single-source baseline severely under-covers. On average, MDCP attains a 22.02% smaller set compared to the `Baseline-agg` method, with notably smaller variance of set width. while the later one is conservative, yielding 97.07% global and 95.37% worst-case coverage for regression.

In the `Nonlinear` setting presented in Figure 6, MDCP maintains valid average and worst-case coverage while achieving consistently shorter prediction sets across all settings. In contrast, the single-source baseline fails to achieve validity, and the `Baseline-agg` method is overly conservative even in the worst-case sense. MDCP strikes a balance between coverage and efficiency: it achieves much higher coverage with just slightly

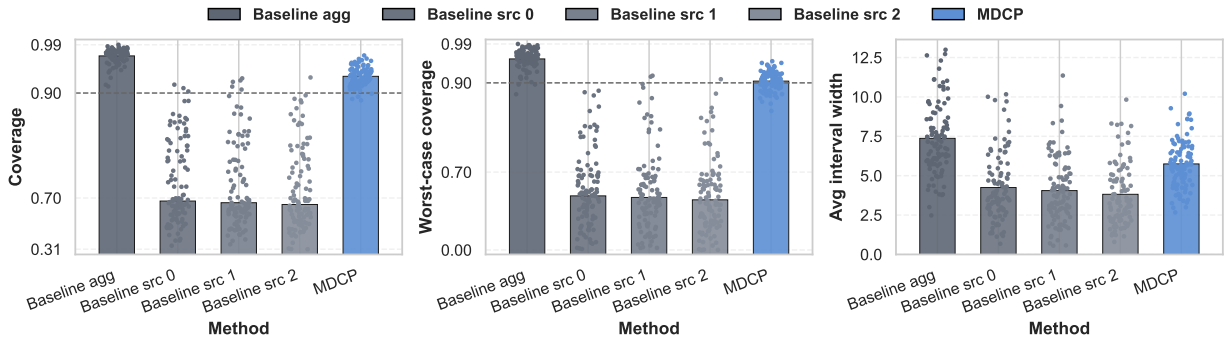


Figure 5: Evaluation with regression `Linear` suites; details are otherwise the same as Figure 2.

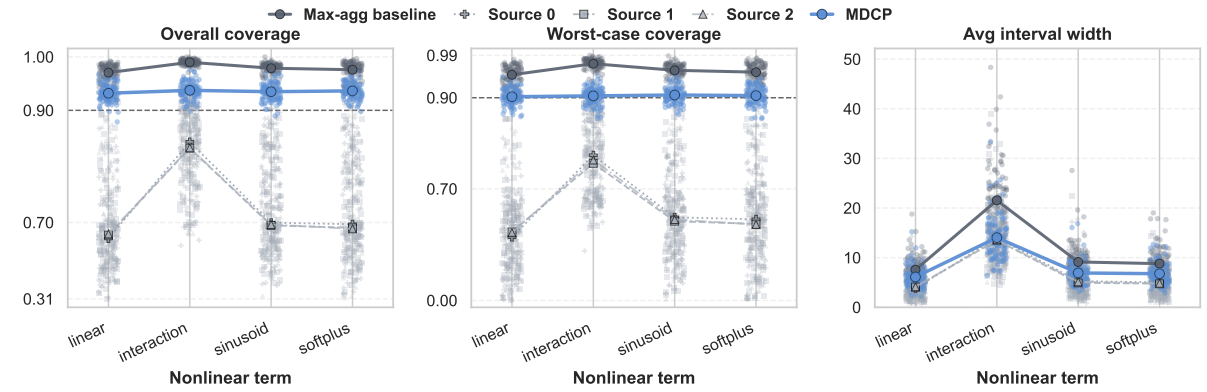


Figure 6: Evaluation with regression `Nonlinear` suites; details are otherwise the same as in Figure 3.

longer prediction sets than the single-source counterparts, and avoids the unnecessary overlap with tight coverage compared with `Baseline-agg`.

In the `Temperature` experiments where the temperature parameters governs the separation of multiple sources, as shown in Figure 7, we observe similar messages as in the classification case. The performance of `Baseline-src-`k degrades as τ increases, with lower average and worst-case coverage and larger standard deviation. By contrast, MDCP stands right at the point of tight coverage, while achieving uniformly smaller set widths than the `Baseline-agg` method by adaptively trading off the coverage across sources.

Additional ablation study. We conduct the same suite of ablation study on numerical stability of the optimization process as in the classification settings; see Appendix B.4 for details. Again, our results across all the simulation settings demonstrate the stability of the current optimization process.

6 Real-data applications

Finally, we demonstrate the broad application of MDCP through three real datasets. Section 6.1 focuses on a classification task where MDCP protects against subpopulation shift. Section 6.2 addresses uniform coverage across urban and rural areas when inferring economic information from satellite image. Section 6.3 uses a medical service dataset to ensure fairness across sensitive groups without observing the group label.

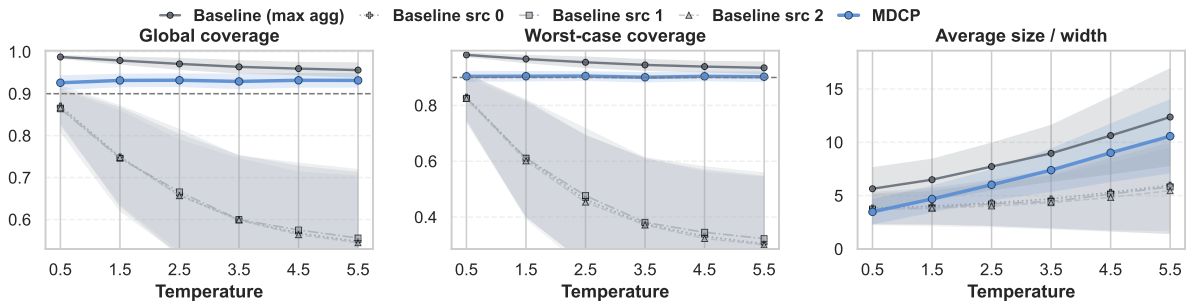


Figure 7: Evaluation with regression `Temperature` suites; details are otherwise the same as in Figure 4.

6.1 Functional category of satellite image under subpopulation shift

Satellite ML has been widely used to detect functional land uses, allocate resources, and inform risk analyses. A key challenge here is the geographic heterogeneity and acquisition variability. Here, we use MDCP to protect subpopulation shift between data-rich locations to data-poor regions due to different materials, urban morphologies, and imaging conditions.

We leverage the 2016 time slice in the Functional Map of the World (FMoW) dataset [Christie et al., 2018] with over one million images from 249 countries/regions, and the label is one of 62 functional classes. We focus on uniform coverage across regions in Africa, the Americas, Asia, Europe, Oceania and others. We treat each geographic region as a source. Let X denote the image input and Y the functional class label. In this context, uniform coverage ensures reliability under arbitrary changes in the composition of regions.

The selected data contains 140,459 samples in total. We allocate 37.5% as the model training fold $|\mathcal{D}_{\text{pre-train}}| = 52,531$. The training distribution is highly imbalanced, with 30.27% from European, 38.72% from the Americas, yet only 2.23% from Ocenia and 0.05% from Other. We train a `DenseNet-121` classifier [Huang et al., 2018] initialized with `ImageNet` weights [Deng et al., 2009] on the full pre-training set $\mathcal{D}_{\text{pre-train}}$, which pools samples from all source domains. The resulting classifier provides the per-source predictions $\hat{p}_k(y|x)$ and is thereafter treated as fixed. Next, we perform $N = 100$ random partitions of the remaining data into auxiliary train (12.5%), calibration (37.5%), test (50%) splits. For each run, we learn $\lambda(x)$ on the auxiliary training set and calibrate the MDCP sets, following the procedure in Section 4.2 and Section 5.2; the baseline methods follow the APS score based on the trained model, using the 37.5% calibration split to calibrate single-source prediction sets and evaluate on the 50% test split. The nominal coverage is set at $1 - \alpha = 0.9$, and the results are reported in Figure 8.

Due to nontrivial heterogeneity across regions, we observe unequal coverage for single-source baselines. Standard conformal prediction sets calibrated from regions like Asia and Europe achieve overall coverage above 0.9, yet still exhibit low worst-case coverage. On the other hand, the Oceania and Other regions struggle to cover other sources due to data scarcity. In contrast, MDCP remains valid across all sources, with near-tight worst-case coverage (although the average coverage is inevitably above 0.9). Moreover, `Baseline-agg` without efficiency optimization inflates the prediction set: it admits any signal deemed useful by any source, leading to over-coverage especially when certain sources have limited data. MDCP mitigates this issue by joint training across sources to balance coverage. Indeed, its set size is even smaller than single-source prediction sets, showing the significant benefit of efficiency optimization.

In Appendix B.4.2, we further examine the penalty-tuning approach similar to the ablation studies in simulations. In this task, we again observe similar results from the tuned and untuned versions of MDCP, showing the stability of our procedure.

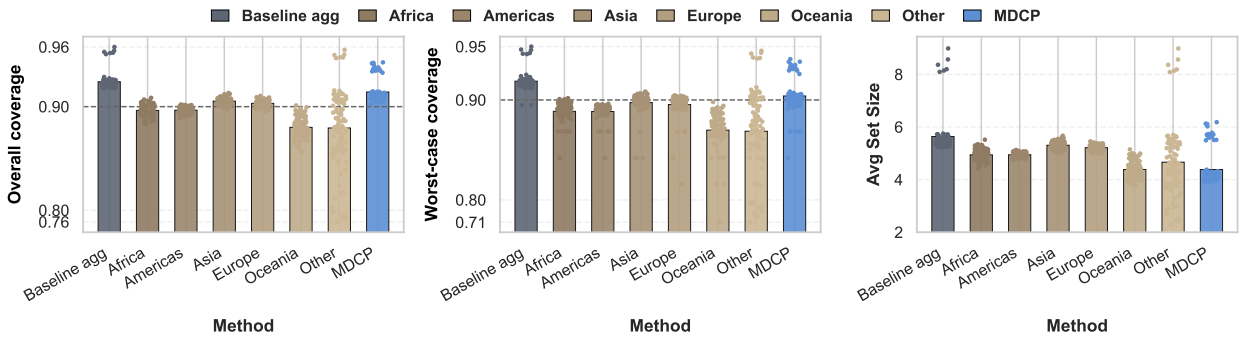
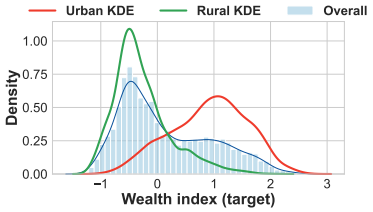
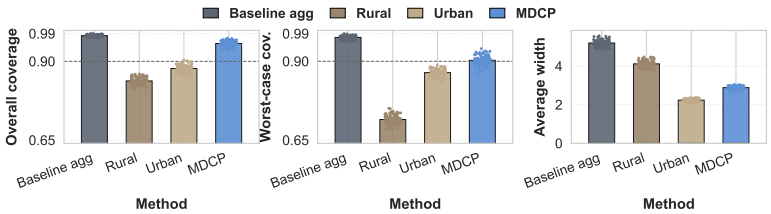


Figure 8: Performance of MDCP, `Baseline-agg`, and `Baseline-src-`k with each source region on the FMoW dataset across six sources: Africa, Americas, Asia, Europe, Oceania, and Other regions. The bars show the average results over $N = 100$ runs, and the dots show the result in each run. Left: overall coverage evaluated over the entire test set. Middle: worst-coverage over all test sources. Right: average set size over the entire test set.



(a) Per-source and pooled distribution of the label. The curves come from a kernel density estimation.



(b) Performance of MDCP and baselines averaged over $N = 100$ splits. Left: coverage evaluated over the entire test set. Middle: worst-case coverage for two sources. Right: prediction set width over the entire test set.

Figure 9: Results of MDCP and baselines in the PovertyMap dataset.

6.2 Poverty prediction under urban-rural shift

Household surveys for mapping economic well-being are infrequent or missing especially in regions where nationally representative surveys are limited by local resources [Blumenstock et al., 2015]. In these scenarios, satellite imagery offers a scalable proxy: a practical strategy is to learn from countries with the desired economic label then transfer to countries with images only [Abelson et al., 2014]. In this part, we visit the subset of a modified release of the Yeh et al. [2020] poverty-mapping dataset from 2014 to 2016 to show the application of MDCP to provide reliable uncertainty quantification across rural and urban areas. In this data, the features are the satellite image, and the label is a continuously-valued wealth index. Figure 9a visualizes the label density in the urban and rural areas which exhibits strong heterogeneity.

The dataset contains $n = 7,535$ samples, with 2,664 from the urban areas and 4,871 from rural, each treated as a source. We reserve 37.5% of the data for training and fit separate 8-channel `ResNet-18` backbones [He et al., 2015] with randomly initialization for the rural and urban subsets. In this model, a Gaussian location-scale head is attached to the last hidden layer to output the estimators $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ under the working model $\hat{f}(y \mid x) = \mathcal{N}\big(y; \hat{\mu}(x), \hat{\sigma}^2(x)\big)$. This yields two source models, $\hat{f}^{\text{Rural}}(y \mid x)$ and $\hat{f}^{\text{Urban}}(y \mid x)$. Next, we perform $N = 100$ randomly splits of the remaining data into auxiliary train (12.5%), calibration (37.5%), test (50%) folds. For each random split, we learn $\lambda(x)$ with the auxiliary training fold, use the calibration fold to calibrate the MDCP set, and evaluate on the test set. The single-source baselines (`Urban`-only and `Rural`-only) rely solely on the pre-trained source-specific models, perform calibration on the source-specific 37.5% calibration fold, and are evaluated on the full test fold.

As shown in Figure 9b, single-source models calibrated with single-source data fail to achieve valid coverage on the other domain. We see from Figure 9a that the rural distribution is more skewed; under strong heterogeneity, despite the larger sample size from the rural source, single-source calibration still produces wide intervals yet low coverage. On the other hand, `Baseline-agg`, which naively combines single-source prediction sets, is overly conservative. MDCP maintains tight worst-case coverage with significant efficiency gains. Its length is even shorter than single-source sets from the rural data, showing that the efficiency calibration step also improves upon the pre-trained model. With prediction loss minimization, the pre-trained model is clearly not optimized for efficient prediction intervals when coupled with the conformity score similar to Lei et al. [2018].

Finally, in Appendix B.4.2, we find that the penalty-tuning extension of MDCP, which uses the training data to adaptively search for an appropriate penalty parameter in the dual objective, leads to further improvement in prediction set efficiency while achieving a similar coverage on both pooled test data and worst-case coverage. This tuned variant yields a prediction set length comparable to the shortest (invalid) single-source prediction set.

6.3 Medical services utilization across sensitive groups

Our last application revisits the Medical Expenditure Panel Survey (MEPS) dataset used in Romano et al. [2019a], including Panels 19-21 [MEPS19, MEPS20, MEPS21], to address equalized coverage even without observing the sensitive group label. The dataset contains detailed individual-level information on demographics and health care utilization. The features include age, marital status, race, poverty level, and health status and insurance related covariates. The label is a continuously-valued medical service utilization score.

We follow the same pre-processing steps as Romano et al. [2019a] with one-hot encoding of categorical variables. The feature dimension for X is 139, consistent across panels. We apply a log transformation to the label due to its skewedness; without this step, the estimated variance would be excessively inflated which drastically degrades the efficiency of single-source baselines. As reported by the Romano et al. [2019b], predictive distributions vary across the sensitive attribute *race*: a neural-network predictor tends to predict higher utilization for non-White than for White individuals. Motivated by this finding, we treat *race* as the source label, assigning $k = 0$ to non-White and $k = 1$ to White, with sample sizes $n_0 = 9640$ and $n_1 = 6016$.

We split the data into training (60%), calibration (20%), and test (20%) folds. For both MDCP and the baselines, we follow the same per-source modeling procedure as in Section 5.3: conditional densities are modeled as $P^{(k)}_{Y|X=x} \sim \mathcal{N}(\mu_k(x), \sigma_k(x)^2)$, with $\hat{\mu}_k(x)$ and $\hat{\sigma}_k(x)$ estimated via gradient-boosting decision trees trained on the source-specific training fold. MDCP further fits $\lambda(x)$ using the same training data and calibrates prediction sets on the entire calibration fold. In contrast, the single-source baselines (`Non-White` only and `White` only) calibrate solely on their respective source-specific calibration fold. The `Baseline agg` combines the two single-source calibrated sets. Finally, the three methods are all evaluated on the same test fold. The above protocol is applied independently to each panel, with results reported separately.

Figure 10 reports the performance of the competing methods. The single-source baseline trained and calibrated exclusively on the non-white group exhibits systematic undercoverage (both on average and worst-case) across panels. This is because the white group is more right-skewed and the single-source baseline from the non-white group fails to cover its heavy tail. On the other hand, single-source sets trained and calibrated exclusively on the White group approximately attains worst-case coverage, yet the width of the prediction sets is exceedingly high. We conjecture that this may be due to the unreliable estimation of variance and the working model with the skewed data; of course, the prediction model was not trained with efficiency in the downstream conformal prediction set in mind. Similarly, `Baseline-agg` is overly conservative and has very wide prediction sets. Finally, MDCP achieves tight worst-case coverage, showing the role of approximate complementary slackness. Efficiency optimization also leads to significant improvement in set size, achieving even shorter sets than the single-source baseline without aggregation.

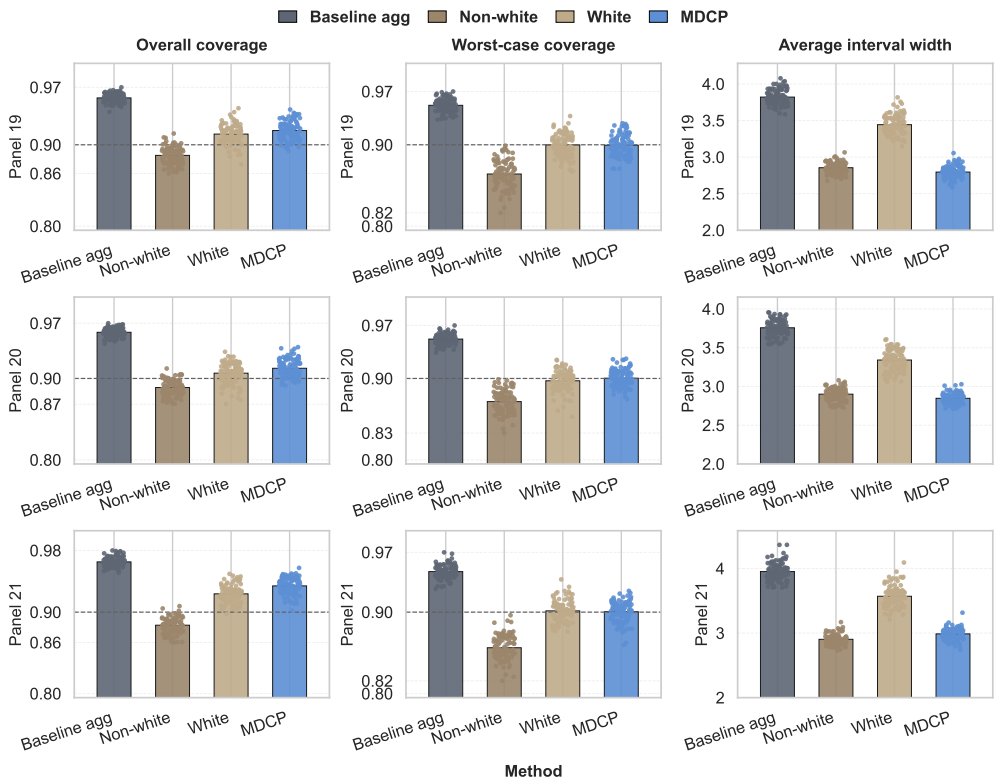


Figure 10: Results of MDCP and baselines in the MEPS dataset evaluation across three panels and two sensitive groups (sources), white and non-white. The bars show results averaged over $N = 100$ runs, and the dots show single-run results. Each row corresponds to one panel, and each column corresponds to one metric: average coverage over all test data, worst-case coverage across two sources, and average length of prediction set over all test data.

Finally, in Appendix B.4.2, we find that in this dataset, the penalty-tuning extension of MDCP again improves the prediction set efficiency upon untuned MDCP while achieving a similar coverage on both pooled test data and worst-case coverage. Combining the results in both simulations and real data applications, we recommend using this tuned variant if computation is not a concern, since it never worsens the performance.

7 Discussion

In this work, we propose the MDCP framework for constructing one single prediction set that offers valid coverage over multiple heterogeneous distributions. The key component is the max-p aggregation, which simply takes the union of single-source conformal prediction sets and therefore offers the desired coverage. While this scheme seems to simply construct a prediction set that is larger than needed for valid coverage in any single source, we show via an optimality analysis that, once coupled with a suitable conformity score, the MDCP set with max-p aggregation is both optimal and tight. We then propose concrete algorithms that learn the optimal conformity score through an empirical dual objective, combined with max-p aggregation to approach optimality while maintaining finite-sample uniform validity. Our classification algorithm only needs standard single-source classifiers, and we estimate the conditional density in regression problems via a Gaussian working model, thereby connecting to commonly-used conformity scores in the literature. Extensive simulations and real-world applications demonstrate the validity, efficiency, and tightness of MDCP, and its utility in protecting against sub-population shift, maintaining robustness across heterogeneous populations,

and ensuring equalized coverage across sensitive groups. In ablation studies, we also find the utility of a penalty-tuning approach which leads to occasional improvement in efficiency and comparable empirical coverage as the default MDCP procedure.

Several follow-up questions remain open. The first is a general formulation of multi-distribution extension with any base conformity score. Inspired by a population-level analysis, our conformity score is constructed by finding high-density regions across populations. In classification problems, it coincides with the natural idea of admitting labels into the prediction set based on predicted probability. In regression, however, it might not always be desirable to threshold a density function since it may lead to non-interval sets, and conformity scores that lead to prediction intervals by construction, such as those based on quantile functions [Romano et al., 2019b], are proven effective. Therefore, it may be meaningful to develop a general framework that learns a multi-distribution combination of pre-specified single-source conformity scores by directly optimizing efficiency-based objectives [Stutz et al., 2021, Huang et al., 2023, Xie et al., 2024].

Second, instead of max-p aggregation, another natural idea of forming a uniformly valid prediction set is to predict which group/population the new test point is from, and trade-off the coverage based on this prediction. However, it remains unclear how to manage the membership estimation error in this case, and whether this could be combined with the max-p aggregation.

Finally, when it comes to protecting against subpopulation shift, our framework still requires the knowledge of the subpopulation the labeled data are from. In practice, the subpopulation shift may change from one unknown mixture to another. How to develop robust conformal prediction sets that protect against such shifts may be a valuable problem for future investigation.

Acknowledgements

The authors thank the Wharton Research Computing team for the computational resources provided and the great support from the staff members.

References

Brian Abelson, Kush R. Varshney, and Joy Sun. Targeting direct cash transfers to the extremely poor. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1563–1572, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623335. URL https://doi.org/10.1145/2623330.2623335.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015. doi: 10.1126/science.aac4420. URL https://www.science.org/doi/abs/10.1126/science.aac4420.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world, 2018. URL https://arxiv.org/abs/1711.07846.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(57):1757–1774, 2008. URL http://jmlr.org/papers/v9/crammer08a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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

Yarin Gal et al. Uncertainty in deep learning. 2016.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The annals of Statistics*, pages 401–414, 1982.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf008, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36:26699–26721, 2023.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

Ying Jin, Zhimei Ren, and Zhengyuan Zhou. Sensitivity analysis under the f-sensitivity models: a distributional robustness perspective. *arXiv preprint arXiv:2203.04373*, 2022.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

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.

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

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

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/0e65972dce68dad4d52d063967f0a705-Paper.pdf.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, pages 6755–6764. PMLR, 2020.

Natalia L Martinez, Martin A Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. Blind pareto fairness and subgroup robustness. In *International Conference on Machine Learning*, pages 7492–7501. PMLR, 2021.

MEPS19. Medical expenditure panel survey, panel 19. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181. Accessed: Oct, 2025.

MEPS20. Medical expenditure panel survey, panel 20. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181. Accessed: Oct, 2025.

MEPS21. Medical expenditure panel survey, panel 21. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192. Accessed: Oct, 2025.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J. Candès. With malice towards none: Assessing uncertainty via equalized coverage, 2019a. URL https://arxiv.org/abs/1908.05428.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32:3543–3553, 2019b.

Yaniv Romano, Matteo Sesia, and Emmanuel J Candès. Classification with valid and adaptive coverage. *arXiv preprint arXiv:2006.02544*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

Ola Spjuth, Robin Carrión Brännström, Lars Carlsson, and Niharika Gauraha. Combining prediction intervals on multi-source non-disclosed regression datasets. In *Conformal and Probabilistic Prediction and Applications*, pages 53–65. PMLR, 2019.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International conference on artificial intelligence and statistics*, pages 2611–2619. PMLR, 2021.

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

Ran Xie, Rina Barber, and Emmanuel Candes. Boosted conformal prediction intervals. *Advances in Neural Information Processing Systems*, 37:71868–71899, 2024.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *Annals of statistics*, 50(5):2587, 2022.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 2020.

A Technical proofs

A.1 Proof of Theorem 1

Proof of Theorem 1. Since $\sup_{j\in[K]} p^{(j)}(y) \le \alpha$ implies $p^{(k)}(y) \le \alpha$, we have

$$\mathbb{P}\big(\sup_{j\in[K]} p^{(j)}(y) \le \alpha\big) \le \mathbb{P}\big(p^{(k)}(Y_{n+1}) \le \alpha\big) \le \alpha,$$

under $P^{(k)}$ and $\mathcal{D}$. The coverage statement follows by complement. The equality $\hat{C} = \bigcup_k \hat{C}^{(k)}$ is immediate from the definition of the supremum and the threshold rule:

- If $y \in \Big\{y \in \mathcal{Y} : \sup_{j\in[K]} p^{(j)}(y) > \alpha\Big\}$, then $y \in \bigcup_{j=1}^{K} \big\{y \in \mathcal{Y} : p^{(j)}(y) > \alpha\big\}$, since $\exists j$ s.t. $p^{(j)}(y) > \alpha$.
- If $y \in \bigcup_{j=1}^{K} \big\{y \in \mathcal{Y} : p^{(j)}(y) > \alpha\big\}$, then $\exists k$ s.t. $p^{(j)}(y) > \alpha$, then $y \in \Big\{y \in \mathcal{Y} : \sup_{j\in[K]} p^{(j)}(y) > \alpha\Big\}$.

This concludes the proof of Theorem 1. □

A.2 Proof of Theorem 2

Proof of Theorem 2. Write the joint density function $w_k(x, y) := r_k(x) f_k(y \mid x)$. The primal problem can be expressed as

$$\min_{I\in\{0,1\}} \iint I \, d\mu d\nu \quad \text{s.t.} \quad \iint I \, w_k \, d\mu d\nu \;\ge\; 1-\alpha, \quad k = 1, \dots, K.$$

Relax $I \in \{0, 1\}$ to $I \in [0, 1]$. Since functions $I \in [0, 1]$ form a vector space [Luenberger, 1997], we consider the Lagrangian with constant multipliers $\lambda_k \ge 0$:

$$\mathcal{L}(I, \lambda) = \iint I(x, y) \, d\mu d\nu - \sum_{k=1}^{K} \lambda_k \Big(\iint I(x, y) \, w_k(x, y) \, d\mu d\nu - (1 - \alpha) \Big).$$

Let $h_\lambda(x, y) := \sum_{k=1}^{K} \lambda_k w_k(x, y)$. Then we have

$$\mathcal{L}(I, \lambda) = \iint I(x, y) \, [1 - h_\lambda(x, y)] \, d\mu d\nu \;+\; (1 - \alpha) \sum_{k=1}^{K} \lambda_k.$$

For a fixed value of λ, minimizing over $I(x, y) \in [0, 1]$ pointwise in (x, y) yields the minimizer

$$I^*_\lambda(x, y) \in \begin{cases} \{1\}, & h_\lambda(x, y) > 1, \\ [0, 1], & h_\lambda(x, y) = 1, \\ \{0\}, & h_\lambda(x, y) < 1, \end{cases}$$

which yields the threshold form

$$C_\lambda(x) = \{y : h_\lambda(x, y) > 1\} \;\cup\; S(x), \quad S(x) \subseteq \{y : h_\lambda(x, y) = 1\}. \tag{17}$$

After minimizing over I, the dual objective is

$$\Phi(\lambda) = (1 - \alpha) \sum_{k=1}^{K} \lambda_k \;-\; \iint (h_\lambda(x, y) - 1)_+ \, d\mu(y) \, d\nu(x),$$

this gives the marginal dual objective (6) mentioned in the theorem, and the dual problem is to maximize $\Phi(\lambda)$ over $\lambda \in \mathbb{R}^K_+$.

Note that Slater's condition holds (e.g., $C(x) \equiv \mathcal{Y}$ strictly satisfies each constraint for $\alpha \in (0,1)$), so strong duality applies and a dual maximizer λ^* exists [Luenberger, 1997]. Let λ^* be a dual maximizer and define $h^*(x,y) = \sum_k \lambda_k^* r_k(x) f_k(y \mid x)$ and the tie set $T(x) = \{y : h^*(x,y) = 1\}$. There exists a primal optimizer

$$I^*(x,y) = \mathbb{1}\{h^* > 1\} + Z^*(x,y)\ \mathbb{1}\{y \in T(x)\},$$

with $Z^* : X \times Y \to [0,1]$ measurable, chosen so that

$$\begin{aligned} \lambda_k^* > 0 \quad &\Rightarrow \quad \int_{\mathcal{X}} \int_{\mathcal{Y}} I^* r_k f_k \, d\mu(y)\, d\nu(x) = 1 - \alpha, \\ \lambda_k^* = 0 \quad &\Rightarrow \quad \int_{\mathcal{X}} \int_{\mathcal{Y}} I^* r_k f_k \, d\mu(y)\, d\nu(x) \geq 1 - \alpha, \end{aligned}$$

where the covariate distribution $P_X^{(k)}$ admits a density $r_k(x)$ with respect to ν. Equivalently, writing

$$a_k := \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbb{1}\{h^* > 1\} r_k f_k \, d\mu(y)\, d\nu(x), \qquad b_k := \int_{\mathcal{X}} \int_{y \in T(x)} Z^* r_k f_k \, d\mu(y)\, d\nu(x),$$

Z^* must satisfy $a_k + b_k = 1 - \alpha$, for all k with $\lambda_k > 0$ and $a_k + b_k \geq 1 - \alpha$ for all k with $\lambda_k > 0$. When multiple constraints with their Lagrangian multiplier $\lambda_k > 0$, achieving all equalities generally requires a non-constant Z^* (for example, using randomized inclusion on the boundary). In cases where the boundary has measure zero $(\nu \otimes \mu)(T) = 0$, one can implement Z^* deterministically as an indicator of a measurable subset of the tie set; in cases where the boundary has non-zero measure $(\nu \otimes \mu)(T) > 0$, this corresponds to randomized tie-breaking.

Accordingly, complementary slackness yields, with $g_k(C) := \iint_{\mathcal{X} \times \mathcal{Y}} I w_k \, d\mu \, d\nu - (1 - \alpha) \geq 0$, it must hold that

$$\lambda_k^*\, g_k(C^*) = 0, \quad \forall k. \tag{18}$$

Thus: (i) if $\lambda_k^* > 0$ then $P^{(k)}(Y \in C^*(X)) = 1 - \alpha$, and (ii) if $\lambda_k^* = 0$ then $P^{(k)}(Y \in C^*(X)) \geq 1 - \alpha$.

We show next that at least one coordinate of λ^* is strictly positive, i.e., statement (iii). Let $\rho := \nu \otimes \mu$ and recall that for each k,

$$w_k(x,y) := r_k(x)\, f_k(y \mid x) \geq 0$$

is integrable with respect to ρ and satisfies $\iint w_k \, d\rho = 1$ since it is the joint density of (X,Y) under $P^{(k)}$ with respect to ρ.

Notice that for the dual objective (6) with $h_\lambda(x,y) = \sum_k \lambda_k w_k(x,y)$, we have $\Phi(0) = 0$. Fix any $j \in \{1, \ldots, K\}$ and consider $\lambda = t e_j$ with $t > 0$, where e_j is the j-th unit vector. Then $h_\lambda = t w_j$ and

$$\Phi(t e_j) = (1 - \alpha)t - \iint (t w_j - 1)_+ \, d\rho.$$

For any $a \geq 0$ and $t > 0$, $(ta - 1)_+ \leq ta\, \mathbb{1}\{a \geq 1/t\}$. Applying this pointwise with $a = w_j(x,y)$ and integrating,

$$\iint (t w_j - 1)_+ \, d\rho \leq t \iint w_j\, \mathbb{1}\{w_j \geq 1/t\} \, d\rho.$$

Define $T_j(t) := \iint w_j\, \mathbb{1}\{w_j \geq 1/t\} \, d\rho$. Since w_j is integrable, $T_j(t) \to 0$ as $t \downarrow 0$ (the tail of an integrable function vanishes). Therefore,

$$\iint (t w_j - 1)_+ \, d\rho \leq t T_j(t) = o(t),$$

and hence

$$\Phi(te_j) \geq t\left[(1-\alpha) - T_j(t)\right].$$

Because $T_j(t) \to 0$ and $1-\alpha > 0$, there exists $t_0 > 0$ such that for all $t \in (0, t_0)$,

$$\Phi(te_j) \geq t\frac{1-\alpha}{2} > 0.$$

Thus $\sup_{\lambda \geq 0} \Phi(\lambda) > 0$, so a dual maximizer cannot be $\lambda^* = 0$. Consequently, $\sum_k \lambda_k^* > 0$ and there exists at least one k^* with $\lambda_{k^*}^* > 0$. By complementary slackness (18),

$$\hat{P}^{(k^*)}(Y \in C^*(X)) = 1 - \alpha.$$

This proves item (iii).

Finally, if $\mu(\{y : h_{\lambda^*}(x, y) = 1\}) = 0$ for ν-almost every x, then the boundary set is μ-null almost surely, making the optimizer unique up to $(\nu \otimes \mu)$-null sets. □

A.3 Proof of Theorem 3

Proof of Theorem 3. Fix $x \in X$ and write $I(y) := \mathbb{1}\{y \in C(x)\}$. The conditional program is

$$\begin{aligned} &\min_{I \in \{0,1\}} \quad \int I(y)\, d\mu(y) \\ &\text{subject to} \quad \int I(y) f_k(y \mid x)\, d\mu(y) \geq 1 - \alpha, \quad \text{for } k = 1, \ldots, K. \end{aligned}$$

Relax $I \in \{0, 1\}$ to $I \in [0, 1]$. Similar to the marginal problem, we can form the Lagrangian with multipliers $\lambda(x) = (\lambda_1(x), \ldots, \lambda_K(x)) \in \mathbb{R}_+^K$ (here x is treated as fixed, yet we write the argument in x for clarity):

$$\mathcal{L}_x(I, \lambda(x)) = \int I(y)\, d\mu(y) - \sum_k \lambda_k(x) \left(\int I(y) f_k(y \mid x)\, d\mu(y) - (1-\alpha) \right).$$

Let $h_{\lambda(x)}(y) := \sum_k \lambda_k(x) f_k(y \mid x)$. Then

$$\mathcal{L}_x(I, \lambda(x)) = \int I(y)[1 - h_{\lambda(x)}(y)]\, d\mu(y) + (1-\alpha) \sum_k \lambda_k(x).$$

For fixed $\lambda(x)$, minimization over $I \in [0, 1]$ is pointwise in y. Any minimizer has the threshold form

$$C_{\lambda(x)}(x) = \{y : h_{\lambda(x)}(y) > 1\} \cup S(x), \quad \text{with } S(x) \subseteq \{y : h_{\lambda(x)}(y) = 1\}. \tag{19}$$

The dual function is

$$\Phi_x(\lambda(x)) = (1-\alpha) \sum_k \lambda_k(x) - \int \left(h_{\lambda(x)}(y) - 1\right)_+ d\mu(y),$$

this gives the conditional dual objective (8) mentioned in the theorem, and the dual problem is to maximize Φ_x over $\lambda(x) \in \mathbb{R}_+^K$.

Slater's condition holds (e.g., $C(x) \equiv \mathcal{Y}$ yields strict feasibility since $\alpha \in (0, 1)$), so strong duality applies and a dual maximizer $\lambda^*(x)$ exists. Thresholding $h_{\lambda^*(x)}$ yields a primal optimum $C^*(x)$. Complementary slackness gives, for each k,

$$\lambda_k^*(x) \left[\int \mathbb{1}\{y \in C^*(x)\} f_k(y \mid x)\, d\mu(y) - (1-\alpha) \right] = 0. \tag{20}$$

Hence:

- If $\lambda_k^*(x) > 0$, then $P^{(k)}(Y_{n+1} \in C^*(x) \mid X_{n+1} = x) = 1 - \alpha$.
- If $\lambda_k^*(x) = 0$, then $P^{(k)}(Y_{n+1} \in C^*(x) \mid X_{n+1} = x) \geq 1 - \alpha$.

We now show that at least one coordinate of $\lambda^*(x)$ is strictly positive. Note that $\Phi_x(0) = 0$. Fix any $j \in \{1, \ldots, K\}$ and consider $\lambda(x) = te_j$ with $t > 0$, where e_j is the j-th unit vector. Then $h_{\lambda(x)}(y) = tf_j(y \mid x)$ and

$$\Phi_x(te_j) = (1 - \alpha)t - \int (tf_j(y \mid x) - 1)_+ d\mu(y).$$

For any $a \geq 0$ and $t > 0$, $(ta - 1)_+ \leq ta \cdot \mathbb{1}\{ta - 1 \geq 0\} = ta \cdot \mathbb{1}\{a \geq 1/t\}$. Applying this pointwise with $a = f_j(y \mid x)$ and integrating,

$$\int (tf_j - 1)_+ d\mu \leq t \int f_j \mathbb{1}\{f_j \geq 1/t\} d\mu.$$

Because $f_j(\cdot \mid x)$ is a density (or probability mass function), $\int f_j d\mu = 1$. The set $\{f_j \geq 1/t\}$ shrinks to the empty set as $t \downarrow 0$, and $0 \leq f_j \mathbb{1}\{f_j \geq 1/t\} \leq f_j$. By dominated convergence,

$$T_j(t) := \int f_j \mathbb{1}\{f_j \geq 1/t\} d\mu \to 0 \text{ as } t \downarrow 0.$$

Therefore,

$$\Phi_x(te_j) \geq (1 - \alpha)t - tT_j(t) = t\left[(1 - \alpha) - T_j(t)\right].$$

Since $T_j(t) \to 0$ and $1 - \alpha > 0$, there exists $t_0 > 0$ such that for all $t \in (0, t_0)$,

$$\Phi_x(te_j) \geq t(1 - \alpha)/2 > 0.$$

Thus $\sup_{\lambda(x) \geq 0} \Phi_x(\lambda(x)) > 0$, so a dual maximizer cannot be $\lambda^*(x) = 0$. Consequently, $\sum_k \lambda_k^*(x) > 0$ and there exists some k^* with $\lambda_{k^*}^*(x) > 0$. By complementary slackness (20),

$$P^{(k^*)}(Y_{n+1} \in C^*(x) \mid X_{n+1} = x) = 1 - \alpha.$$

Additionally, if $\mu(\{y : h_{\lambda^*(x)}(y) = 1\}) = 0$, then the boundary set is μ-null, making the optimizer unique up to μ-null sets. □

A.4 Proof of Theorem 4

We begin by introducing the notation employed throughout the proofs, as well as several auxiliary lemmas that will be relied upon in the main results. Proofs of the lemmas are deferred to the Appendix A.4.2. We begin with some useful definitions.

Definition 11 (Lévy distance). *For CDFs F and G on $\mathbb{R}$, we denote the Lévy distance as*

$$d_L(F, G) := \inf\left\{\varepsilon > 0 : \forall x \in \mathbb{R},\ F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon\right\}. \tag{21}$$

Definition 12 (Generalized quantile). *For $\alpha \in (0, 1)$, define the generalized α-quantile set of a CDF G as*

$$Q_\alpha(G) := \left\{q \in \mathbb{R} : G(q^-) \leq \alpha \leq G(q)\right\}. \tag{22}$$

We term each $q \in Q_\alpha(G)$ as a generalized α-quantile.

Definition 13 (Randomized quantile). *Given scores $W_1, \ldots, W_n \in \mathbb{R}$ and an auxiliary $U \sim \mathrm{Unif}(0, 1)$ independent of the data, define the randomized empirical CDF*

$$\hat{G}_U(t) := \frac{\#\{i : W_i < t\} + (1 + \#\{i : W_i = t\})U}{n + 1}.$$

We define the randomized empirical α-quantile of $\{W_1, \ldots, W_n\}$ as

$$\hat{q}_\alpha := \inf\left\{t \in \mathbb{R} : \hat{G}_U(t) \geq \alpha\right\}. \tag{23}$$

Lemma 14 (Quantile stability under uniform CDF convergence). *Let G be a CDF on $\mathbb{R}$ and let G_n be CDFs with $\sup_t |G_n(t) - G(t)| \to 0$ as $n \to \infty$. Fix $\alpha \in (0,1)$, and let $Q_\alpha(G)$ denote the generalized α-quantile set defined in Equation (22). Suppose $q_n \in \mathbb{R}$ satisfies $G_n(q_n-) \le \alpha \le G_n(q_n)$, then*

$$\operatorname{dist}(q_n, Q_\alpha(G)) := \inf_{q \in Q_\alpha(G)} |q_n - q| \to 0.$$

In particular, if $Q_\alpha(G) = \{q^*\}$ (i.e., G is continuous at the α-quantile and there is no flat segment at level α), then $q_n \to q^*$.

Lemma 15 (Lévy-to-quantile-set continuity). *Let d_L denote the Lévy distance as defined in Equation (21). If $d_L(G_n, G) \le \varepsilon$ and $Q_\alpha(G) = [a, b]$, then every $q \in Q_\alpha(G_n)$ satisfies $q \in [a - \varepsilon, b + \varepsilon]$. In particular,*

$$\sup_{q \in Q_\alpha(G_n)} \operatorname{dist}(q, Q_\alpha(G)) \le \varepsilon.$$

A.4.1 Proof of Theorem 4

Proof of Theorem 4. To clearly denote the asymptotic regime as $n \to \infty$, throughout this proof, we add the superscript (n) to the estimated quantities $\hat{\lambda}$, $\hat{f}_k$, and $\hat{h}$.

Since both f_k and $\hat{\lambda}^{(n)}$ are bounded, and by assumption, $\sup_x \|\hat{\lambda}^{(n)}(x) - \lambda^*(x)\|_\infty \xrightarrow{p} 0$, $\sup_{x,y} |\hat{f}_k^{(n)}(y|x) - f_k(y|x)| \xrightarrow{p} 0$, decompose

$$\sup_{x,y} \left|\hat{h}^{(n)}(x,y) - h^*(x,y)\right| \le \sum_k \left[\sup_x \left|\hat{\lambda}_k^{(n)}(x) - \lambda_k^*(x)\right| \cdot \sup_{x,y} f_k(y|x) + \sup_x \left|\hat{\lambda}_k^{(n)}(x)\right| \cdot \sup_{x,y} \left|\hat{f}_k^{(n)}(y|x) - f_k(y|x)\right| \right].$$

Each term tends to 0 in probability, hence

$$\sup_{x,y} \left|\hat{h}^{(n)}(x,y) - h^*(x,y)\right| \xrightarrow{p} 0 \tag{24}$$

Fix k and condition on $\hat{h}^{(n)}$. Let $W_{k,i} := \hat{h}^{(n)}(X_i^{(k)}, Y_i^{(k)})$ and $W_{k,\text{test}} := \hat{h}^{(n)}(x,y)$. Since $V = -\hat{h}$, the randomized p-value from the Equation (10) is the randomized empirical CDF of W at the test point:

$$p_k^{(n)}(x,y) = \frac{\#\{i : W_{k,i} < W_{k,\text{test}}\} + (1 + \#\{i : W_{k,i} = W_{k,\text{test}}\})\, U_k}{n_k + 1}, \quad \text{with } U_k \sim \text{Unif}(0,1).$$

Thus, the single-source prediction set is based on thresholding $\hat{h}^{(n)}$:

$$\{y : p_k^{(n)}(x,y) \ge \alpha\} = \{y : \hat{h}^{(n)}(x,y) > \hat{q}_{k,\alpha}^{(n)}\},$$

where $\hat{q}_{k,\alpha}^{(n)}$ is the randomized empirical α-quantile of $W_{k,i}$:

$$\hat{q}_{k,\alpha}^{(n)} := \inf\left\{ t \in \mathbb{R} : \frac{\#\{i : W_{k,i} < t\} + (1 + \#\{i : W_{k,i} = t\}) \cdot U_k}{n_k + 1} \ge \alpha \right\},$$

Aggregating K sources yields

$$\hat{C}^{(n)}(x) = \{y : \hat{h}^{(n)}(x,y) \ge \hat{q}_{\min,\alpha}^{(n)}\},$$

where $\hat{q}_{\min,\alpha}^{(n)} := \min_k \hat{q}_{k,\alpha}^{(n)}$.

Let $F_k(t)$ be the CDF of $h^*(X,Y)$ under $P^{(k)}$ and $F_k^{(n)}(t)$ the CDF of $\hat{h}^{(n)}(X,Y)$. Conditional on the training data, the calibration scores are i.i.d. from a distribution with CDF $F_k^{(n)}$. By the DKW inequality,

$$\sup_t \left|\hat{F}_k^{(n)}(t) - F_k^{(n)}(t)\right| \xrightarrow{p} 0.$$

Moreover, since $\sup|\hat{h}^{(n)} - h^*| \xrightarrow{p} 0$, we can show that

$$F_k(t - \varepsilon_n) \le F_k^{(n)}(t) \le F_k(t + \varepsilon_n)$$

with $\varepsilon_n \xrightarrow{p} 0$, hence $d_L(F_k^{(n)}, F_k) \to 0$ in probability (equivalently, $F_k^{(n)}(t) \to F_k(t)$ at continuity points). Therefore,

$$d_L(\hat{F}_k^{(n)}, F_k) \to 0$$

in probability, and in particular $\hat{F}_k^{(n)}(t) \to F_k(t)$ at all continuity points t (uniformly in probability). Since our randomized p-value inversion selects an empirical generalized α-quantile $\hat{q}_{k,\alpha}^{(n)} \in Q_\alpha(\hat{F}_k^{(n)})$, apply Lemma 14 yields

$$\operatorname{dist}\left(\hat{q}_{k,\alpha}^{(n)}, Q_\alpha(F_k^{(n)})\right) \xrightarrow{p} 0.$$

apply Lemma 15 with $G_n = F_k^{(n)}$, $G = F_k$ and $\varepsilon = \varepsilon_n$ yields

$$\operatorname{dist}(Q_\alpha(F_k^{(n)}), Q_\alpha(F_k)) \le \varepsilon_n \xrightarrow{p} 0.$$

By triangle inequality,

$$\operatorname{dist}(\hat{q}_{k,\alpha}^{(n)}, Q_\alpha(F_k)) \le \operatorname{dist}(\hat{q}_{k,\alpha}^{(n)}, Q_\alpha(F_k^{(n)})) + \operatorname{dist}(Q_\alpha(F_k^{(n)}), Q_\alpha(F_k)) \xrightarrow{p} 0.$$

That is,

$$\operatorname{dist}\left(\hat{q}_{k,\alpha}^{(n)}, Q_\alpha(F_k)\right) \xrightarrow{p} 0.$$

In particular, if 1 is the unique generalized α-quantile of F_k, then $\hat{q}_{k,\alpha}^{(n)} \xrightarrow{p} 1$.

By KKT conditions and strong duality, we know the optimal rule thresholds at 1; that is, it includes all (x, y) with $h^*(x, y) > 1$ and, at most, a randomized fraction of those with $h^*(x, y) = 1$. Under $P^{(k)}$, the maximal coverage achievable without lowering the threshold is

$$\mathbb{P}^{(k)}\left(h^*(X, Y) \ge 1\right) = 1 - F_k(1^-).$$

The coverage constraint $\mathbb{P}^{(k)}(y \in C^*) \ge 1 - \alpha$ therefore forces $1 - F_k(1^-) \ge 1 - \alpha$, i.e., $F_k(1^-) \le \alpha$ for every k; otherwise the threshold-1 solution would be infeasible, contradicting strong duality.

By conclusion (iii) of Theorem 3, there exists at least one k with $\lambda_k > 0$, such that α lies in the jump $[F_k(1^-), F_k(1)]$, hence the generalized α-quantile is unique and equals 1. Consequently, for each k, any α-quantile of F_k is no smaller than 1, and for at least one k, it is exactly equal to 1. Therefore,

$$\hat{q}_{\min,\alpha}^{(n)} \xrightarrow{p} 1. \tag{25}$$

Let

$$\delta_n := \left|\hat{q}_{\min,\alpha}^{(n)} - 1\right| + \sup_{x,y}\left|\hat{h}^{(n)}(x, y) - h^*(x, y)\right|.$$

Then $\delta_n \xrightarrow{p} 0$ by Equations (25) and (24). Also note that,

- If $h^*(x, y) > 1 + 2\delta_n$, then $\hat{h}^{(n)}(x, y) > 1 + \delta_n \ge \hat{q}_{\min,\alpha}^{(n)}$, so $(x, y) \in \hat{C}^{(n)}$.
- If $h^*(x, y) < 1 - 2\delta_n$, then $\hat{h}^{(n)}(x, y) < 1 - \delta_n \le \hat{q}_{\min,\alpha}^{(n)}$, so $(x, y) \notin \hat{C}^{(n)}$.

Hence,

$$\{(x, y) : h^*(x, y) > 1 + 2\delta_n\} \subseteq \hat{C}^{(n)} \subseteq \{(x, y) : h^*(x, y) \ge 1 - 2\delta_n\} \cup \hat{B}_n,$$

where

$$\hat{B}_n \subseteq \{(x,y) : |\hat{h}^{(n)}(x,y) - \hat{q}^{(n)}_{\min,\alpha}| = 0\} \subseteq \{(x,y) : |h^*(x,y) - 1| \le \delta_n\},$$

which follows from the Equation (24) derived earlier. Taking symmetric differences with $\{(x,y) : h^*(x,y) \ge 1\}$ and letting $n \to \infty$,

$$\limsup_n \rho\left(\hat{C}^{(n)} \triangle \{(x,y) : h^*(x,y) \ge 1\}\right) \le \limsup_n \rho\left(\{(x,y) : |h^*(x,y) - 1| \le 2\delta_n\}\right) \le |T| \tag{26}$$

since $T = \{(x,y) : h^*(x,y) = 1\}$ and $|T| = \rho(T) = \int_{\mathcal{X}} \mu(T(x))d\nu(x)$, and the measure of shrinking neighborhoods of T tends to $|T|$.

Finally, write

$$|\hat{C}^{(n)}| - |C^*| = \left(|\hat{C}^{(n)}| - |\{h^* \ge 1\}|\right) + \left(|\{h^* \ge 1\}| - |C^*|\right).$$

The first bracket is bounded in absolute value by $\rho(\hat{C}^{(n)} \triangle \{h^* \ge 1\}) \le |T|$ from Inequality (26). The second bracket equals $|T| - |S^*| + |\{h^* > 1\}| - |\{h^* > 1\}| = |T| - |S^*|$, whose absolute value is $\le |T|$. Therefore,

$$\limsup_{n\to\infty} \Big| |\hat{C}^{(n)}| - |C^*| \Big| \le |T|.$$

Moreover, Inequality (26) shows there exists a subsequence $\{n_j\}$ and a measurable set $S_\infty \subseteq T := \{(x,y) : h^*(x,y) = 1\}$ such that

$$\rho\left(\hat{C}^{(n_j)} \,\Delta\, (\{h^* > 1\} \cup S_\infty)\right) \to 0.$$

Consequently, choosing the oracle set $C^*(x) = \{y : h^*(x,y) > 1\} \cup S_\infty(x)$ yields $|\hat{C}^{(n_j)}| \to |C^*|$. □

A.4.2 Proof of Lemma 14

Proof of Lemma 14. For a monotone right-continuous H, denote the left limit by $H(x-) := \sup_{t<x} H(t)$. If $\sup_t |H_n(t) - H(t)| \le \varepsilon$, then also $\sup_x |H_n(x-) - H(x-)| \le \varepsilon$, because

$$H_n(x-) = \sup_{t<x} H_n(t) \ge \sup_{t<x}[H(t) - \varepsilon] = H(x-) - \varepsilon,$$
$$H_n(x-) = \sup_{t<x} H_n(t) \le \sup_{t<x}[H(t) + \varepsilon] = H(x-) + \varepsilon.$$

Let $a := \inf\{t : G(t) \ge \alpha\}$ and $b := \sup\{t : G(t) \le \alpha\}$; then $a \le b$ and $Q_\alpha(G) = [a,b]$. Then

- For any $\delta > 0$, $G(a-\delta) < \alpha$. Define $\gamma_L(\delta) := \alpha - G(a-\delta) > 0$.
- For any $\delta > 0$, for all $x \ge b + \delta$ we have $G(x-) > \alpha$. Indeed, for any such x pick s with $b < s < x$; then $G(s) > \alpha$ by definition of b, so $G(x-) \ge G(s) > \alpha$. Hence define $\gamma_R(\delta) := G(b + \delta/2) - \alpha > 0$.

Left bound. Fix $\delta > 0$ and choose n large so that $\sup_t |G_n(t) - G(t)| \le \varepsilon_n$ with $\varepsilon_n < \gamma_L(\delta)/2$. If $q \le a - \delta$ then

$$G_n(q) \le G(q) + \varepsilon_n \le G(a-\delta) + \varepsilon_n = \alpha - \gamma_L(\delta) + \varepsilon_n < \alpha,$$

contradicting the requirement $\alpha \le G_n(q)$ for $q \in Q_\alpha(G_n)$. Therefore any $q \in Q_\alpha(G_n)$ must satisfy $q > a - \delta$.

Right bound. With the same n and ε_n and the $\gamma_R(\delta)$ defined above, if $q \ge b + \delta$ then

$$G_n(q-) \ge G(q-) - \varepsilon_n \ge (\alpha + \gamma_R(\delta)) - \varepsilon_n > \alpha,$$

contradicting the requirement $G_n(q-) \le \alpha$ for $q \in Q_\alpha(G_n)$. Therefore any $q \in Q_\alpha(G_n)$ must satisfy $q < b + \delta$.

Therefore, for any fixed $\delta > 0$ and all sufficiently large n we have

$$Q_\alpha(G_n) \subset (a - \delta, b + \delta).$$

In particular, our selected $q_n \in Q_\alpha(G_n)$ lies within δ of the closed set $[a, b] = Q_\alpha(G)$, so $\operatorname{dist}(q_n, Q_\alpha(G)) \le \delta$. Because $\delta > 0$ was arbitrary, $\operatorname{dist}(q_n, Q_\alpha(G)) \to 0$.

For the unique-quantile case $Q_\alpha(G) = \{q^*\}$, the distance convergence implies $q_n \to q^*$. □

A.4.3 Proof of Lemma 15

Proof of Lemma 15. $d_L \le \varepsilon$ means $F(x - \varepsilon) - \varepsilon \le G_n(x) \le F(x + \varepsilon) + \varepsilon$ for all x. If $q \in Q_\alpha(G_n)$ then $G_n(q-) \le \alpha \le G_n(q)$, hence

$$F(q - \varepsilon) - \varepsilon \le \alpha \le F(q + \varepsilon) + \varepsilon.$$

If $q < a - \varepsilon$, then $q + \varepsilon < a$ and $F(q + \varepsilon) \le \alpha$, contradicting the right inequality. If $q > b + \varepsilon$, then $q - \varepsilon > b$ and $F(q - \varepsilon) \ge \alpha$, contradicting the left inequality. So $q \in [a - \varepsilon, b + \varepsilon]$. □

A.5 Optimality of the integrated dual problem

In this part, we formalize the discussion at the beginning of Section 4.1 on the optimal $\lambda^*(x)$ as the solution to an integrated dual objective.

Proposition 16 (Equivalence of integrated dual and conditional dual)**.** *For $k = 1, \dots, K$, let $f_k(\cdot \mid x)$ be the conditional density/pmf of $Y \mid X = x$ with respect to μ. Fix $\alpha \in (0, 1)$. For $\lambda \in \mathbb{R}_+^K$ and $x \in \mathcal{X}$, define $h_\lambda(x, y) := \sum_{k=1}^K \lambda_k f_k(y \mid x)$, and*

$$\varphi_x(\lambda) := (1 - \alpha) \sum_{k=1}^K \lambda_k - \int_{\mathcal{Y}} (h_\lambda(x, y) - 1)_+ \, d\mu(y).$$

Let $\tilde{\nu}$ be a σ-finite measure on $(\mathcal{X}, \mathcal{A})$ with Radon–Nikodym density $w(x) := \frac{d\tilde{\nu}}{d\nu}$ satisfying $0 < w(x) < \infty$ for ν-a.e. x. Consider the integrated dual objective

$$\Phi_{\tilde{\nu}}(\lambda(\cdot)) := \int_{\mathcal{X}} \left[(1 - \alpha) \sum_{k=1}^K \lambda_k(x) - \int_{\mathcal{Y}} (h_\lambda(x, y) - 1)_+ \, d\mu(y) \right] d\tilde{\nu}(x).$$

Then $\Phi_{\tilde{\nu}}(\lambda(\cdot)) = \int_{\mathcal{X}} w(x) \, \varphi_x(\lambda(x)) \, d\nu(x)$, and

(i) A measurable $\lambda^*(\cdot)$ maximizes $\Phi_{\tilde{\nu}}$ if and only if

$$\lambda^*(x) \in \arg\max_{\lambda \in \mathbb{R}_+^K} \varphi_x(\lambda) \qquad \text{for } \nu\text{-a.e. } x.$$

Hence the set of maximizers is independent of the particular choice of $\tilde{\nu}$, as long as $d\tilde{\nu}/d\nu > 0$ ν-a.e.

(ii) For ν-a.e. x, any maximizer $\lambda^*(x)$ is a dual maximizer of the x-conditional problem (7). Thresholding h_{λ^*} at level 1 gives the conditionally optimal set

$$C^*(x) = \{y \in Y : h_{\lambda^*}(x, y) > 1\} \cup S(x), \quad \text{with } S(x) \subseteq \{y : h_{\lambda^*}(x, y) = 1\},$$

as in Theorem 3. Thus the optimal score and set do not depend on $\tilde{\nu}$.

Proof of Proposition 16. By the Radon–Nikodym theorem, $d\tilde{\nu} = w \, d\nu$ with $w > 0$ ν-a.e. Substituting,

$$\Phi_{\tilde{\nu}}(\lambda(\cdot)) = \int_{\mathcal{X}} \left[(1 - \alpha) \sum_k \lambda_k(x) - \int_{\mathcal{Y}} (h_\lambda(x, y) - 1)_+ \, d\mu(y) \right] w(x) \, d\nu(x) = \int_{\mathcal{X}} w(x) \, \varphi_x(\lambda(x)) \, d\nu(x).$$

For any measurable $\lambda(\cdot)$,

$$\Phi_{\tilde{\nu}}(\lambda(\cdot)) = \int_{\mathcal{X}} w(x)\,\varphi_x(\lambda(x))\,d\nu(x) \le \int_{\mathcal{X}} w(x)\,\sup_{\lambda\ge 0}\varphi_x(\lambda)\,d\nu(x), \tag{27}$$

with equality iff $\varphi_x(\lambda(x)) = \sup_{\lambda\ge 0}\varphi_x(\lambda)$ for ν-a.e. x. Note that $\hat{\lambda}(\cdot)$ with $\hat{\lambda}(x) \in \operatorname{argmax}\varphi_x(\cdot)$ exists and attains the upper bound, for this $\hat{\lambda}$,

$$\Phi_{\tilde{\nu}}(\hat{\lambda}) = \int_{\mathcal{X}} w(x)\varphi_x(\hat{\lambda}(x))\,d\nu(x) = \int_{\mathcal{X}} w(x)\,\sup_{\lambda\ge 0}\varphi_x(\lambda)\,d\nu(x),$$

which matches the upper bound in Equation (27) and is therefore optimal. Moreover, if a candidate $\lambda(\cdot)$ fails to maximize φ_x at a set of x with positive $\tilde{\nu}$-measure, replacing it by a pointwise maximizer on that set always strictly increases $\Phi_{\tilde{\nu}}$, proving the necessity. Because $w(x) > 0$, multiplying by $w(x)$ does not change the pointwise argmax sets, so the maximizers are independent of $\tilde{\nu}$.

By definition, $\varphi_x(\cdot)$ is the dual objective of the x-conditional problem (7). Therefore, a pointwise maximizer $\lambda^*(x)$ is a dual maximizer for (7). By KKT conditions for (7), thresholding $h_{\lambda^*}(x,\cdot)$ at 1 yields the conditionally optimal set stated above. Independence from $\tilde{\nu}$ follows from (i). □

A.6 Proof of Theorem 10

Proof of Theorem 10. First, following exactly the same conditions and proof in Jin et al. [2022, Theorem 1] applied to the loss function $\hat{\ell}(\cdot)$, we can show that $\|\hat{\lambda} - \bar{\lambda}^*\|_{L_2} = O_P\big((\frac{\log n}{n})^{p/(2p+d)}\big)$ and $\|\hat{\lambda} - \bar{\lambda}^*\|_{\infty} = O_P\big((\frac{\log n}{n})^{2p^2/(2p+d)^2}\big)$. Then, by triangle inequality and Assumption 8, we obtain the desired results. □

B Simulation details

B.1 Hyperparameter sampling

We detail the sampling process of the hyperparameters in the simulations in Section 5.3.

For the interaction family, we draw the weights i.i.d. from $w_{uv} \sim \mathcal{N}(0, 1.1^2)$ for $(u,v) \in \mathcal{I}\times\mathcal{I}$. For both the sinusoidal and softplus families, each unit $r = 1,2,3$ uses a projection vector $u_r \in \mathbb{R}^d$ constructed as follows: we first sample a support $S_r \subset \mathcal{I}$ of size 3 uniformly at random and draw a magnitude $M_r \sim \mathrm{Unif}(0.375, 0.875)$, sample a random unit vector $d_r \in \mathbb{R}^3$ and define $u_r[S_r] = M_r d_r$ with $u_r[\mathcal{I}\setminus S_r] = 0$. The sinusoidal component samples $b_r \sim \mathrm{Unif}(-\pi/3, \pi/3)$ and $a_r \sim \mathrm{Unif}(0.5, 1.5)$, independently across r. The softplus component utilizes the same construction for u_r as above, but with $b_r \sim \mathrm{Unif}(-0.5, 0.5)$ and $a_r \sim \mathrm{Unif}(0.75, 2.0)$, again independently across r.

B.2 Algorithm instantiation

This subsection includes omitted implementation details in the classification and regression algorithms in our experiments.

Classification algorithm In Section 5.2, we use nonparametric methods for probability estimations (in our case, we use gradient-boosted trees), and calibrate their probabilistic outputs with stratified cross-validation and an isotonic mapping. We solve for the optimizer λ using minibatch updates. The optimization is implemented in PyTorch with automatic differentiation, where precomputed spline features, data densities, source weights, and related terms are used to form an objective on the minibatch, with a trainable spline parameter matrix. Gradients are then computed via autodifferentiation, and parameters are updated with Adam. After each epoch update, we do full-data evaluations to allow early stopping, improving efficiency and mitigating overfitting.

Regression algorithm In Section 5.3, for each source k we fit a heteroskedastic Gaussian plug-in model for the conditional density. We first learn a regression function $\hat{\mu}_k(x)$ using flexible nonparametric estimators (in our case, we use gradient-boosted trees). To model dispersion, we obtain out-of-fold predictions $\tilde{\mu}$ for the mean model via K-fold: we partition the data into K folds, for each of the K folds, fit the model on $K-1$ folds and predict on the held-out fold and compute the squared residuals $\hat{r}^2 = (Y - \tilde{\mu})^2$. we then fit a second regressor on residual variance to the logarithm of the residual squares using all K folds. At prediction time we evaluate $\hat{\sigma}_k(x)$ from this variance model, and form

$$\hat{f}_k(y \mid x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_k(x)} \exp\left(-\frac{1}{2}\left(\frac{y - \hat{\mu}_k(x)}{\hat{\sigma}_k(x)}\right)^2\right).$$

As in classification, we treat the marginal density of X as constant and use $\hat{f}_k(x, y) := \hat{f}_k(y \mid x)$ throughout. And we also use minibatch optimizer, softplus nonnegativity, and early stopping.

B.3 Grid search algorithm and guarantee

Let $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{calib}}$ be the training and calibration data pooled across all K sources. Define:

$$y_L := \min\{Y_i : (X_i, Y_i) \in \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{calib}}\},$$
$$y_U := \max\{Y_i : (X_i, Y_i) \in \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{calib}}\}.$$

Fix an integer $M \geq 2$ (e.g., $M = 100$). Define a uniform grid on $[y_L, y_U]$ with $\Delta := \frac{y_U - y_L}{M-1}$:

$$y^{(j)} := y_L + j\,\Delta, \quad \text{for } j = 0, 1, \ldots, M-1. \tag{28}$$

We call $y^{(j)}$ and $y^{(j+1)}$ are adjacent grid points. A subset B of indices is called *consecutive* if it contains no gaps; equivalently, B can be written as $\{a, a+1, \ldots, b\}$ for some integers $a \leq b$. For example, $\{3, 4, 5\}$ is consecutive, while $\{3, 5\}$ is not. For a test covariate x, we include a candidate y-grid point y if the aggregated MDCP p-value $p(y) := \max_k p^{(k)}(x, y) >= \alpha$, where each $p^{(k)}$ is derived using formula (10). Let

$$J := \{j \in \{0, \ldots, M-1\} : p(y^{(j)}) \geq \alpha\} \tag{29}$$

be the set of included grid indices. We decompose J into consecutive blocks $B_r = \{j_{r,L}, \ldots, j_{r,R}\}$ for $r = 1, \ldots, R$. We say a decomposition is *maximal*, if the decomposed blocks are consecutive, disjoint and cannot be enlarged by adding adjacent indices from J.

Algorithm 2 Grid-Search Algorithm (Regression)

Input: Number of sources K, pooled calibration data $\mathcal{D} = \cup_{k=1}^{K} \mathcal{D}^{(k)}$, test input x, grid endpoints y_L, y_U, grid size M, grid spacing Δ, significance level α

// Grid construction
Construct grid points $y^{(j)}$, $j = 0, \ldots, M-1$ over $[y_L, y_U]$ as in (28).
// Evaluate aggregated p-values on the grid
for $j = 0$ **to** $M-1$ **do**
  Compute $p(y^{(j)}) = \max_k p^{(k)}(x, y^{(j)})$
end for
Collect included grid points, form J following (29).
// Merge included grid points into blocks
Decompose J into maximal consecutive blocks $B_r = \{j_{r,L}, \ldots, j_{r,R}\}$ for $r = 1, \ldots, R$.
// Extend each block by one grid spacing
for each block B_r **do**
  Create interval $I_r := [y^{(j_{r,L})} - \Delta,\ y^{(j_{r,R})} + \Delta]$
end for
Taking unions of all intervals $C_{\text{grid}}(x) := \bigcup_r I_r$

Output: Regression prediction set $C_{\text{grid}}(x)$ for test input x

Let C_{MDCP} denote the MDCP set we want to construct with score $s_k(x, y) = -\sum_{k=1}^{K} \lambda_k(x; \hat{\Theta}) \hat{f}_k(y \mid x)$ and p-value (10). Let C_{grid} denote the conformal set constructed from the grid search Algorithm 2.

Proposition 17 (Superset on the grid range)**.** *Let $C := C_{\text{MDCP}}(x) \cap [y_L, y_U]$. Suppose each connected component of C is a closed interval $[\ell, r]$ that intersects the grid, i.e., $[\ell, r] \cap \{y^{(j)}\} \neq \varnothing$. Then*

$$C \subseteq C_{\text{grid}}(x) \cap [y_L - \Delta,\ y_U + \Delta].$$

Proof of Proposition 17. Fix a connected component $[\ell, r] \subseteq C$ with $[\ell, r] \cap \{y^{(j)}\} \neq \varnothing$. Let

$$j_1 := \min\{j : y^{(j)} \in [\ell, r]\}, \qquad j_2 := \max\{j : y^{(j)} \in [\ell, r]\}.$$

Since $y^{(j_1)}, y^{(j_2)} \in [\ell, r] \subseteq C$, we have $p(y^{(j_1)}) \geq \alpha$ and $p(y^{(j_2)}) \geq \alpha$, so $j_1, j_2 \in J$ and all indices $j \in [j_1, j_2]$ belong to the same consecutive block B_r.

By grid spacing, $y^{(j_1-1)} = y^{(j_1)} - \Delta$ (if $j_1 > 0$), and $y^{(j_2+1)} = y^{(j_2)} + \Delta$ (if $j_2 < M-1$).

(i). Because j_1 is the first grid index inside $[\ell, r]$, we have $y^{(j_1-1)} < \ell \leq y^{(j_1)}$, hence $\ell \geq y^{(j_1)} - \Delta$.

(ii). Because j_2 is the last grid index inside $[\ell, r]$, we have $y^{(j_2)} \leq r < y^{(j_2+1)}$, hence $r \leq y^{(j_2)} + \Delta$.

Therefore $[\ell, r] \subseteq [y^{(j_1)} - \Delta,\ y^{(j_2)} + \Delta] = I_r$, where I_r is an interval produced from the Algorithm 2. Taking the union over all components yields $C \subseteq \bigcup_r I_r = C_{\text{grid}}(x)$. Finally, by construction $I_r \subseteq [y_L - \Delta,\ y_U + \Delta]$, so

$$C \subseteq C_{\text{grid}}(x) \cap [y_L - \Delta,\ y_U + \Delta].$$

That is, within the observed y-range, the grid merge-and-extend procedure never excludes any MDCP-accepted value and may only enlarge the set. □

B.4 Ablation study on optimization

For the ablation study, we examine the difficulty of optimizing the dual objective (11) and assess the stability and reliability of the optimization procedure. To this end, we introduce the following penalty terms and

investigate their effect, recall in (15), $\lambda_j(x) = \text{softplus}\left(\Lambda(x)^\top \theta_j\right)$ for $j \in [K]$, where $\Lambda(x) \in \mathbb{R}^m$ is a vector of spline basis functions and $\theta_j \in \mathbb{R}^m$ are trainable coefficients:

$$\hat{\mathbb{E}}_{\text{train}}\left[\frac{(1-h_\lambda)_-}{\hat{p}_{\text{data}}}\right] + (1-\alpha)\hat{\mathbb{E}}_{\text{train}}\left[\textstyle\sum_k \lambda_k\right] - \gamma \underbrace{\left(\hat{\mathbb{E}}_{\text{train}}\left[\textstyle\sum_k \lambda_k^2\right] + \textstyle\sum_k \|D\theta_k\|^2\right)}_{\text{Penalty}},$$

where D is the second order difference operator:

$$(D\theta_k)_i = \theta_{k,i} - 2\,\theta_{k,i+1} + \theta_{k,i+2}, \qquad i = 1, \ldots, m-2,$$

which serves as a discrete analogue of penalizing the curvature of the underlying function $\lambda_k(\cdot)$, and $\theta_{k,i}$ is the i-th parameter in the spline feature space.

We select the hyperparameter γ over the grid $[0.0, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0]$. To use the data efficiently, we split the training data into a *mimic calibration set* and a *mimic test set*. For each individual run, we calibrate the method on the mimic calibration set for every candidate γ, evaluate performance on the mimic test set, and choose the γ that yields the smallest average set size on this mimic test set. Denote this selected value by γ^*. We then fix γ^* and run MDCP with the original calibration and test data. Since the calibration and test data are not involved in this optimization process, the uniform coverage guarantee of MDCP still follows. Moreover, we expect the selected hyperparameter to perform at least as well as, and potentially better than, the non-penalized version (i.e., $\gamma = 0$) in terms of the chosen efficiency criterion. We compare the results from the penalized MDCP with data-driven γ^* side by side with the non-penalized version ($\gamma = 0$). We evaluate this approach across all the simulation studies and real data applications.

B.4.1 Simulation results

For the classification simulations, using the setup in Section 5.2, we evaluate performance on the three suites from Section 5.1: `Linear` (Figure 11), `Nonlinear` (Figure 12), and `Temperature` (Figure 13). After the initial training step, we split the training data into equal-sized mimic calibration and mimic test sets (50%/50%) and apply the parameter-selection procedure described above. Across all three suites, tuning the penalty parameter γ produces at most negligible gains in set efficiency. This suggests that the MDCP optimization step is already stable and no additional penalty is required in most of the simulation settings.

In the regression simulations, under the same setup as Section 5.3, we examine performance on the three suites defined in Section 5.1: `Linear` (Figure 14), `Nonlinear` (Figure 15), and `Temperature` (Figure 16). Analogous to the classification experiments, once the model has been trained, we divide the training data evenly into a mimic calibration set and a mimic test set, and subsequently perform the parameter selection procedure described above. Across all three suites, data-driven tuning of the penalty parameter γ produces, at best, marginal improvements in set efficiency. This finding indicates that the baseline MDCP optimization procedure is already sufficiently robust, and that, in most simulated scenarios, the dual optimization problem can be solved reliably without introducing an additional penalty term.

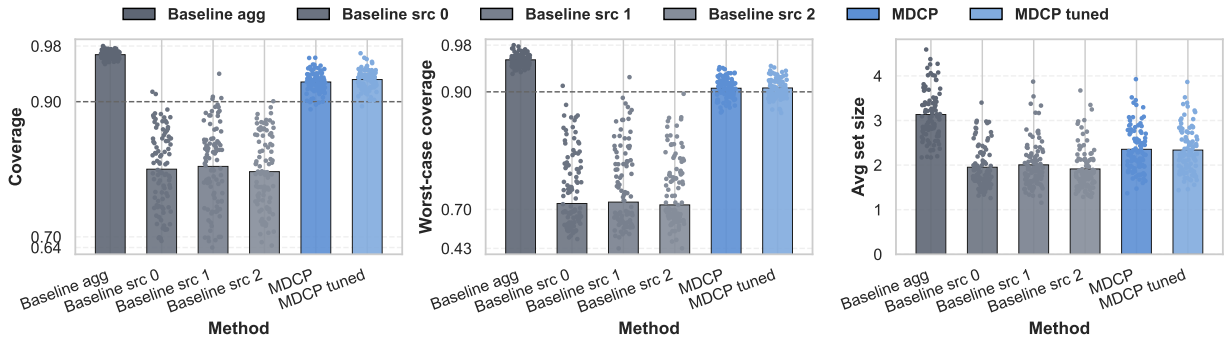


Figure 11: Evaluation on the classification `Linear` suites, where MDCP with data-driven γ^* is labeled as "MDCP tuned". All other experimental settings are identical to those in Figure 2.

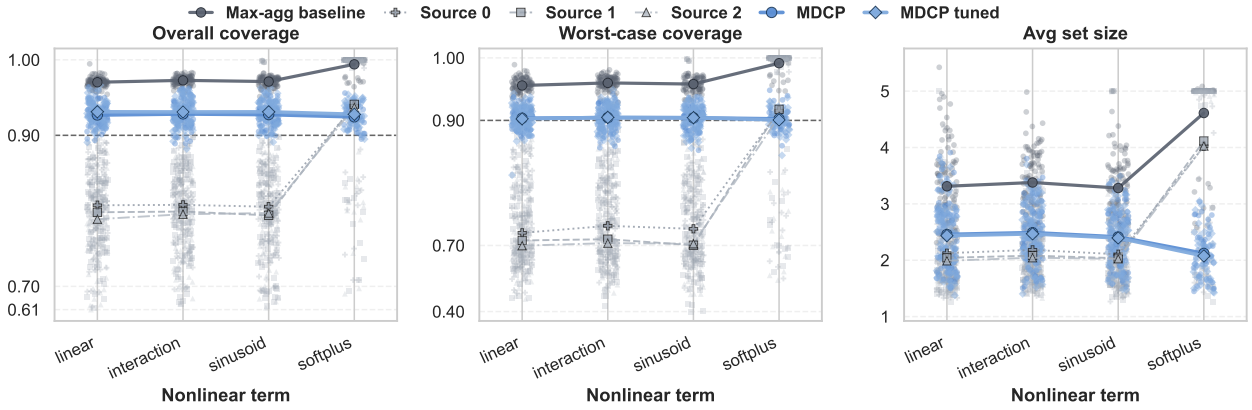


Figure 12: Evaluation on the classification `Nonlinear` suites. Experimental settings are identical to Figure 3. The differences between vanilla MDCP and tuned MDCP are small across all nonlinear term settings.

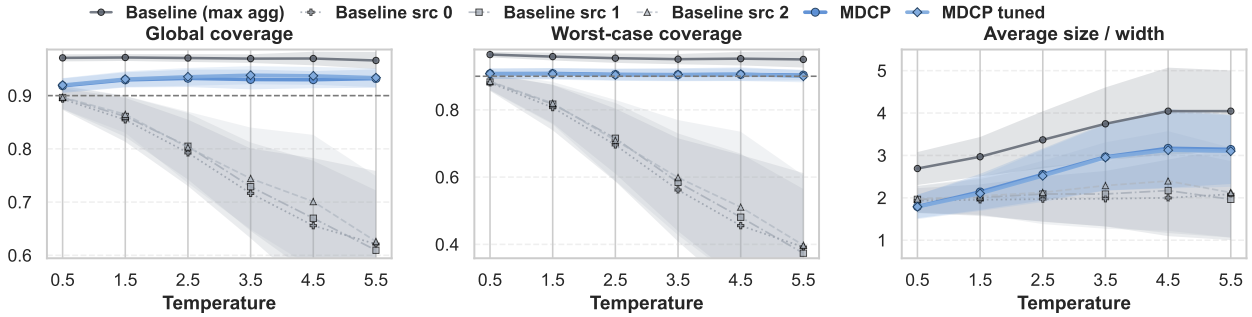


Figure 13: Evaluation on the classification `Temperature` suites. Experimental settings are identical to Figure 4. Vanilla MDCP and tuned MDCP exhibit only minor differences across all temperature parameter settings.

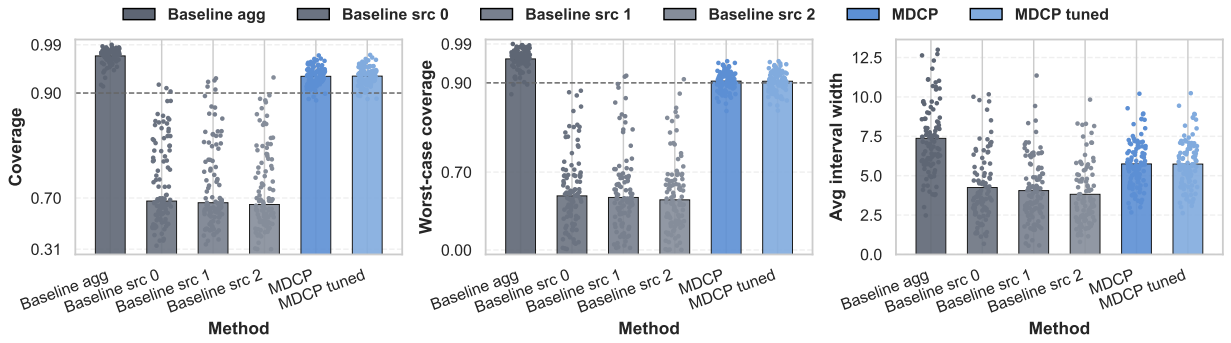


Figure 14: Results on the regression `Linear` suites, where MDCP with the selected penalty strength parameter γ^* appears as "MDCP tuned". All other experimental settings match those in Figure 5.

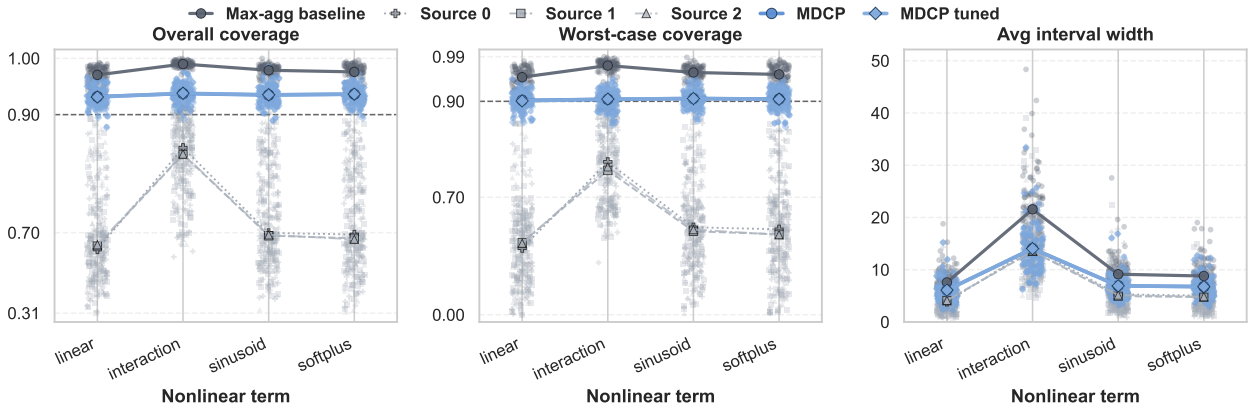


Figure 15: Results on the regression `Nonlinear` suites Experimental settings match those in Figure 6. Across all choices of the nonlinear term, MDCP and tuned MDCP behave very similarly.

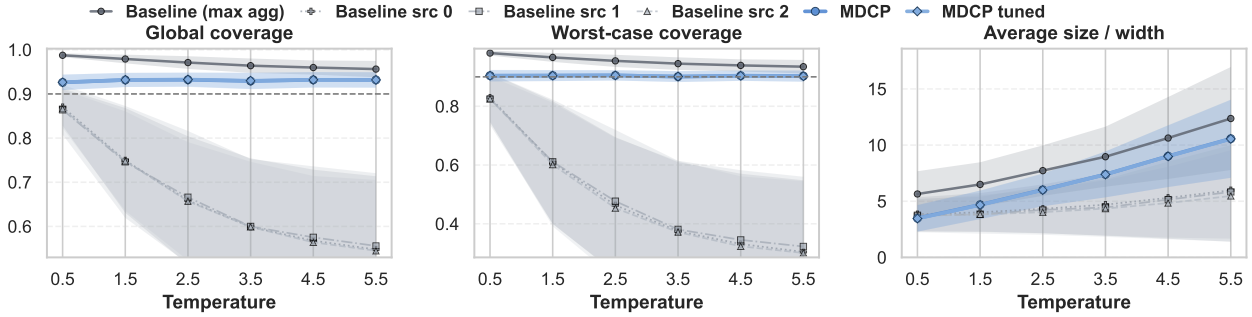


Figure 16: Results on the regression `Temperature` suites. Experimental settings match those in Figure 7. For all temperature parameter values, the gap between MDCP and tuned MDCP is negligible.

B.4.2 Real data results

We also assess the impact of γ on the real-world datasets. Following Section 6, we repeat the procedures for the FMoW, PovertyMap, and MEPS datasets, now including γ as an additional tuning parameter. The candidate values match those used above, ranging from 0.001 to 1000, with $\gamma = 0$ corresponding to the non-penalized version. For all three datasets, the training set is split 50%/50% into mimic calibration

and mimic test subsets. For the FMoW dataset (Figure 17), the original MDCP procedure is already stable, and introducing the penalty term yields little to no improvement. For the PovertyMap dataset (Figure 18), introducing γ improves set efficiency, yielding tighter conformal sets. To investigate this further, we summarize per-source coverage in Figure 19 by evaluating each source separately as the test set. Both MDCP and tuned MDCP are tight on the urban source; however, tuned MDCP, by reweighting the shared points through the λ_k, produces even tighter sets when evaluated on both sources. For the MEPS dataset (Figure 20), the low-density regions of the highly skewed target distribution are particularly challenging for baseline methods using score functions similar to Lei et al. [2018]. In this setting, introducing the penalty term has a substantial effect: it prevents the λ_k from becoming excessively large in low-density regions in order to cover a few difficult points, and instead allows MDCP to concentrate on higher-density, and thus practically more relevant, regions.

These results show that the mimic-split strategy can yield performance gains when the density is difficult to estimate or the optimization problem is challenging, while remaining simple to implement with a 50%/50% calibration–test split. Nonetheless, the vanilla MDCP procedure is already sufficiently robust to serve as the default choice in most settings.

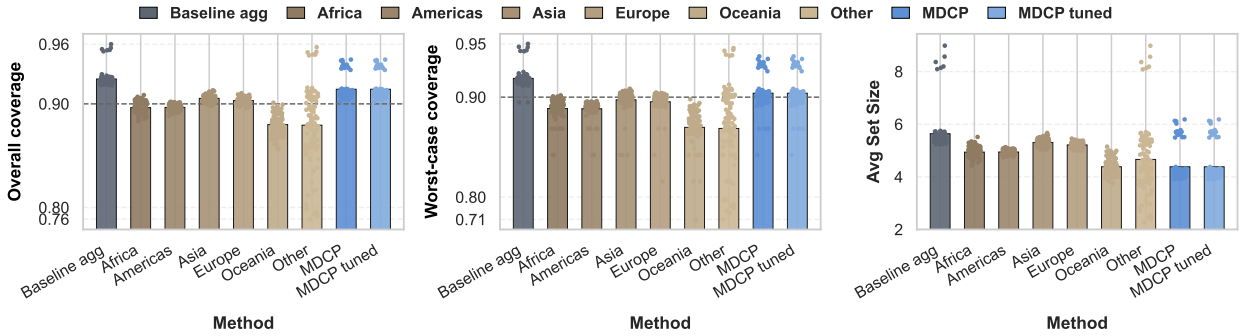


Figure 17: Results on the FMoW data, using the algorithmic procedure described in Section 6.1. MDCP and tuned MDCP produce closely aligned performance in this case.

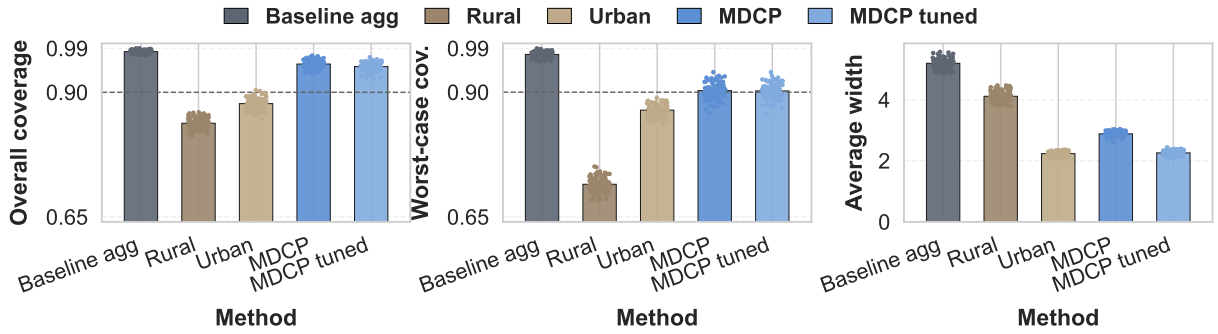


Figure 18: Results on the PovertyMap data, using the algorithmic procedure described in Section 6.2. Introducing the parameter γ improves set efficiency.

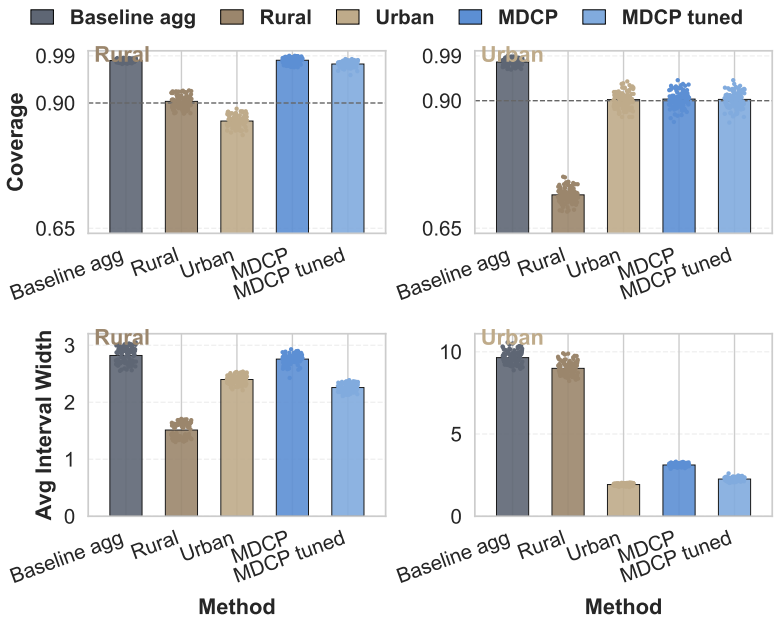


Figure 19: PovertyMap results from the same experiment in Figure 18. Each subplot is labeled (top left) with its evaluation target distribution. In this setting, tuning γ improves MDCP performance for both subset targets.

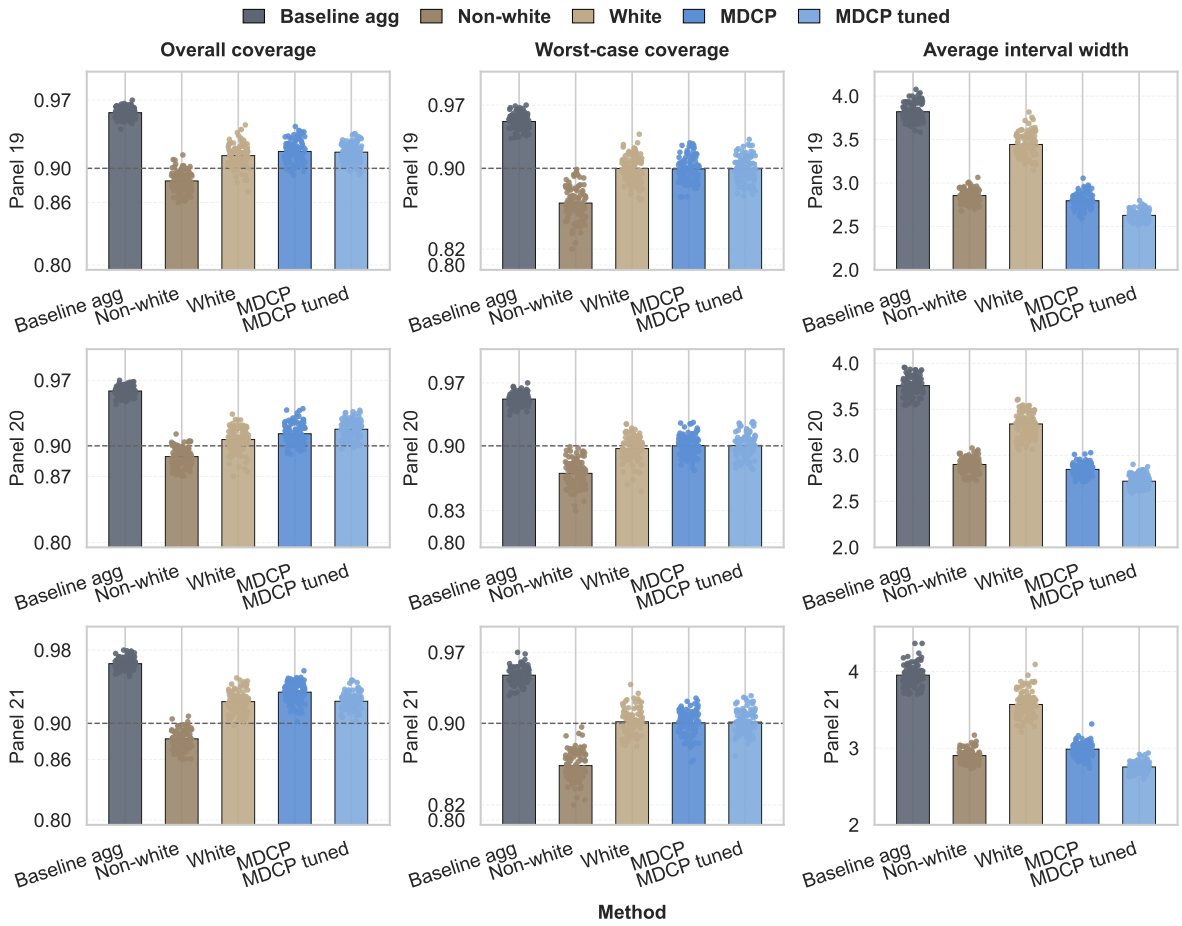


Figure 20: Results on the MEPS data, using the algorithmic procedure described in Section 6.3. Introducing γ further improves tuned MDCP, enhancing its robustness on this highly skewed dataset.