Non-exchangeable Conformal Prediction with Optimal Transport: Tackling Distribution Shifts with Unlabeled Data

Alvaro H.C. Correia[†] **Christos Louizos**[†]
Qualcomm AI Research[*]
Amsterdam, The Netherlands
{acorreia, clouizos}@qti.qualcomm.com

Abstract

Conformal prediction is a distribution-free uncertainty quantification method that has gained popularity in the machine learning community due to its finite-sample guarantees and ease of use. Its most common variant, dubbed split conformal prediction, is also computationally efficient as it boils down to collecting statistics of the model predictions on some calibration data not yet seen by the model. Nonetheless, these guarantees only hold if the calibration and test data are exchangeable, a condition that is difficult to verify and often violated in practice due to so-called distribution shifts. The literature is rife with methods to mitigate the loss in coverage in this non-exchangeable setting, but these methods require some prior information on the type of distribution shift to be expected at test time. In this work, we study this problem via a new perspective, through the lens of optimal transport, and show that it is possible to estimate the loss in coverage and mitigate arbitrary distribution shifts, offering a principled and broadly applicable solution.

1 Introduction

Conformal prediction [45] (CP) works under the assumption that calibration and test data are exchangeable. Exchangeability is a weaker requirement than the more common i.i.d. assumption but still implies that samples are identically distributed, which is hard to verify and ensure in practical applications. Therefore, it is important to develop conformal methods capable of adapting to potential distribution shifts or, at least, quantifying the gap in coverage caused by such shifts. In this paper, we study the effect of distribution shifts in conformal prediction through the lens of optimal transport, which proved instrumental in not only quantifying coverage gaps, but also alleviating them via a reweighting of the calibration data.

We start by introducing the notion of *total coverage gap*: the expected coverage gap over all possible target coverage rates in $[0, 1]$. This metric captures the aggregate effect of distribution shift on conformal prediction and motivates the subsequent contributions, which build on this concept to provide theoretical bounds and practical strategies for mitigating the gap.

1. We derive two new upper bounds to the total coverage gap, formulated in terms of optimal transport distances between the distributions of calibration and test nonconformity scores.
2. We show that one of our upper bounds, which requires only unlabeled samples from the test distribution, can be used to learn weights $\boldsymbol{w} = \{w_i\}_{i=1}^n$ for the calibration data. These weights can then be used in CP to reduce the gap in coverage during test time.

[*]Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. [†]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-5278

3. We evaluate our methods on a (toy) regression task and on the ImageNet-C and iWildCam datasets. For the classification tasks we also consider the more challenging setting which includes covariate and label shift.

The paper is structured as follows. Section 2 provides the necessary background for our main theoretical results. Section 3 presents our two new upper bounds on the total coverage gap, and Section 4 demonstrates their application to reduce the coverage gap. Finally, we review related work in Section 5, report experimental results in Section 6, and conclude in Section 7.

2 Background

In this section, we lay out the background necessary for our main results. We start with a brief introduction to conformal prediction followed by an overview of optimal transport.

2.1 Conformal Prediction

Conformal prediction [45] is a framework to extract prediction sets from predictive models that satisfy finite-sample coverage guarantees under specific assumptions[2]. More precisely, consider the calibration set $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from an unknown distribution P on $\mathcal{X} \times \mathcal{Y}$, formally $(X_1, Y_1, \ldots, X_n, Y_n) \sim P^n$. Given this set, conformal prediction constructs prediction sets $\mathcal{C}(X_t)$ for new points $X_t \sim P$ such that the marginal coverage property holds for any $\alpha \in [0, 1]$:

$$\mathbb{P}(Y_t \in \mathcal{C}(X_t)) \geq 1 - \alpha, \tag{1}$$

where the probability is taken over the randomness of $\{(X_i, Y_i)\}_{i=1}^n$ and (X_t, Y_t). The prediction set $\mathcal{C}(X_t)$ is constructed using *(non)conformity* scores, which quantify how well a sample fits within other samples in a set. One of the most common conformal prediction methods is that of split-conformal prediction (SCP) [29]. In SCP, a score $s(X_i, Y_i)$ is obtained for each point in a calibration set $\{(X_i, Y_i)\}_{i=1}^n$, and at test time $\mathcal{C}(X_t)$ is constructed as

$$\mathcal{C}(X_t) = \left\{ y \in \mathcal{Y} : s(X_t, y) \leq \mathbb{Q}_\alpha \left(\{s(X_i, Y_i)\}_{i=1}^n \right) \right\},$$

with $\mathbb{Q}_\alpha$ the $1 - \alpha$ quantile of the empirical distribution defined by the set of scores $\{s(X_i, Y_i)\}_{i=1}^n$.

The main assumption for the marginal coverage guarantee to hold is that of *exchangeability*; the new point (X_t, Y_t) needs to be exchangeable with the points in the calibration set $\{(X_i, Y_i)\}_{i=1}^n$, i.e., it should follow the same distribution $P(X, Y)$. Unfortunately, violations of the exchangeability assumption are all too common [24] and the naive application of standard SCP when (X_t, Y_t) comes from another distribution $Q(X, Y)$ could produce misleading prediction sets that do not achieve the desired coverage rate [2, 42, 45]. Proper usage of conformal prediction in these settings requires methods to (i) quantify the coverage gap caused by the distribution shift, and (ii) mitigate the effect of the shift on the CP procedure itself to get as close as possible to the target coverage rate.

We refer the reader to [1, 38] for more thorough introductions to conformal prediction.

2.2 Optimal Transport: Couplings and Wasserstein Distance

Consider a complete and separable metric space $(\mathcal{Z}, c)$, where $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is a metric. Let $\mathcal{P}_p(\mathcal{Z})$ be the set of all probability measures P on $(\mathcal{Z}, c)$ with finite moments of order $p \geq 1$, i.e., $\int_{\mathcal{Z}} c(z_0, z)^p dP(z) < \infty$ for some $z_0 \in \mathcal{Z}$. The p-Wasserstein distance is a metric on $\mathcal{P}_p(\mathcal{Z})$ that is defined for any measures P and Q in $\mathcal{P}_p(\mathcal{Z})$ as

$$W_p(P, Q) = \left(\inf_{\pi \in \Gamma(P,Q)} \int_{\mathcal{Z} \times \mathcal{Z}} c(z, z')^p d\pi(z, z') \right)^{1/p} \tag{2}$$

where $\Gamma(P, Q)$ denotes the collection of all measures on $\mathcal{Z} \times \mathcal{Z}$ with marginals P and Q. We refer to any probability measure in $\Gamma(P, Q)$ as a coupling of P and Q and use $\pi^*(P, Q)$ to denote p-Wasserstein optimal couplings, i.e., any coupling that attains the infimum in (2).

[2]Conformal prediction is often described as an uncertainty quantification method, but it may be more accurately viewed as an uncertainty representation technique: it conveys uncertainty through the size of prediction sets rather than assigning a numerical value to uncertainty.

Importantly, Wasserstein distances are also defined for discrete measures, and empirical measures in particular. Let $\hat{P}_n$ and $\hat{Q}_m$ denote the empirical distributions of samples $\{z_i\}_{i=1}^n, z_i \sim P$ and $\{z'_j\}_{j=1}^m, z'_j \sim Q$, which induce empirical measures $\hat{P}_n = \sum_{i=1}^n \delta_{z_i}$ and $\hat{Q}_m = \sum_{j=1}^m \delta_{z'_j}$. In that case, a coupling can be captured by a matrix Γ, with $\Gamma_{i,j}$ the mass to be transported from z_i to z'_j. Similarly, for empirical measures the cost function c induces a cost matrix with $C_{i,j} = \left\| z_i - z'_j \right\|^p$ such that the transportation problem is given by

$$\min_{\Gamma} \sum_{i,j} C_{i,j}\Gamma_{i,j} \qquad \text{subject to} \quad \sum_{i=1}^n \Gamma_{i,j} = 1/m \ \ \forall j \in [\![m]\!], \ \sum_{j=1}^m \Gamma_{i,j} = 1/n \ \ \forall i \in [\![n]\!].$$

In this work, we will be concerned with the distribution over nonconformity scores, which are typically one-dimensional. In this case, the p-Wasserstein simplifies to

$$W_p(P,Q) = \left(\int_0^1 \left\| F_P^{-1}(q) - F_Q^{-1}(q) \right\|^p dq \right)^{1/p}, \tag{3}$$

where F_P^{-1} (resp. F_Q^{-1}) is the quantile function, i.e., the inverse of the cumulative distribution function (CDF) F_P (resp. F_Q) under measure P (resp. Q). For $p = 1$, we can also express the Wasserstein distance in terms of the respective CDFs

$$W_1(P,Q) = \int_{\mathbb{R}} \left| F_P(z) - F_Q(z) \right| dz. \tag{4}$$

In this paper, we focus on the 1-Wasserstein distance. Our main results rely solely on the definitions and properties outlined here, plus basic properties like the triangle inequality. Nevertheless, the interested reader will be well served by the excellent introductions to optimal transport in [31, 43].

3 Theoretical Results

In this section, we define the notion of total coverage gap and introduce our main theoretical results that allow us to upper bound it. For the sake of conciseness we defer the proofs to Appendix A.

We start by laying out the necessary notation. Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be input and output variables and $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a (non)conformity score function. Regardless of the type of distribution shift (e.g., covariate or label shifts) its effect on the conformal prediction guarantees will manifest itself in the distribution over calibration and test scores. Therefore, in this paper we will directly manipulate the distribution of scores $S = s(X,Y)$, and to that end we use $s_\sharp P$ and $s_\sharp Q$ to denote the calibration and test distributions over the scores, i.e., $s_\sharp P = s_*(P)$ is the pushforward measure of P by s. When $s_\sharp P$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, we write its density as $p_{s_\sharp P}$. We also use subscripts to distinguish between scores observed during calibration S_c and at test time S_t. As usual, we will use uppercase letters for random variables and lowercase letters for their realizations, e.g., $S_t = s_t$. We reserve calligraphic letters for sets, e.g., $\mathcal{S}_c = \{s(X_i, Y_i)\}_{i=1}^n$ is the set of calibration scores obtained from a sample of size n drawn from P^n, and $\mathcal{C}(X_t)$ denotes a prediction set for test variable X_t.

3.1 Total Variation Distance Bound

With this notation, we write the coverage under P and Q as

$$P(Y_t \in \mathcal{C}(X_t)) = P(S_t \leq \mathbb{Q}_\alpha(\mathcal{S}_c)) = \mathbb{E}_{S_t \sim s_\sharp P}\left[\mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}[\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha(\mathcal{S}_c)\right)] \right]$$
$$Q(Y_t \in \mathcal{C}(X_t)) = Q(S_t \leq \mathbb{Q}_\alpha(\mathcal{S}_c)) = \mathbb{E}_{S_t \sim s_\sharp Q}\left[\mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}[\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha(\mathcal{S}_c)\right)] \right],$$

with $\mathbb{Q}_\alpha(\mathcal{S}_c)$ the $1-\alpha$ quantile of the empirical distribution defined by a set of calibration scores $\mathcal{S}_c$.

In general, one cannot guarantee valid coverage under arbitrary test distributions Q, i.e., we cannot ensure $Q(Y_{test} \in \mathcal{C}(X_{test})) \geq 1 - \alpha$. Therefore, it is important to quantify the gap in coverage

 https://doi.org/10.52202/085713-5278

induced by the change in distribution from P to Q. To that end, let $\Delta(\alpha)$ denote the coverage gap for a specific α value

$$\begin{aligned}\Delta_{P,Q}(\alpha) &:= \left|P(S_t \leq \mathbb{Q}_\alpha(S_c)) - Q(S_t \leq \mathbb{Q}_\alpha(S_c))\right| \\ &= \left|\mathbb{E}_{S_t \sim s_\sharp P}\left[\mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\left[\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha(\mathcal{S}_c)\right)\right]\right] - \mathbb{E}_{S_t \sim s_\sharp Q}\left[\mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\left[\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha(\mathcal{S}_c)\right)\right]\right]\right|,\end{aligned}$$

where $\mathbf{1}(\cdot)$ is the indicator function. It is easy to show the coverage gap is upper bounded by the total variation distance. We first restate the following well-known result for the total variation between two distributions P and Q (see e.g. Farinhas et al. [10]):

$$D_{TV}(P, Q) \geq |\mathbb{E}_P[g] - \mathbb{E}_Q[g]|,$$

for some function g such that $|g| \leq 1$. It suffices to take g as $g(x) = \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}[\mathbf{1}\left(x \leq \mathbb{Q}_\alpha(\mathcal{S}_c)\right)]$, which is clearly bounded with $|g(x)| \leq 1$ for all x, to get $\Delta_{P,Q}(\alpha) \leq D_{TV}(P, Q)$. Unfortunately, estimating the total variation distance between P and Q without access to their respective densities is impractical. Instead, in the following we will propose two different ways to get around this difficulty and effectively quantify the coverage gap. In both cases, we leverage optimal transport, which defines valid distances even for empirical measures, i.e., when we only have access to P and Q via samples.

3.2 Upper Bound on the Total Coverage Gap

We begin by defining the total coverage gap as follows.

Definition 3.1 (Total coverage gap)**.** The expected coverage gap across all possible values $\alpha \in [0, 1]$ given by

$$\Delta_{P,Q} := \int_0^1 \Delta_{P,Q}(\alpha) d\alpha = \mathbb{E}_{p(\alpha)}[\Delta(\alpha)],$$

with $p(\alpha)$ being the uniform distribution in $[0, 1]$.

The following result establishes that the total coverage gap between P and Q is upper bounded by a weighted CDF distance and the 1-Wasserstein distance between them, $W_1(P, Q)$.

Theorem 3.2. *Let P and Q be probability measures on $\mathcal{X} \times \mathcal{Y}$ with $s_\sharp P$ and $s_\sharp Q$ their respective pushforward measures by a score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Assume $s_\sharp P$ is absolutely continuous with respect to the Lebesgue measure with density $p_{s_\sharp P}(s_c)$. Then the total coverage gap can be upper bounded as follows*

$$\Delta_{P,Q} \leq \int_{\mathbb{R}} p_{s_\sharp P}(s_c) \left|F_{s_\sharp P}(s_c) - F_{s_\sharp Q}(s_c)\right| ds_c \tag{5}$$

$$\leq \Big(\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c)\Big) W_1(s_\sharp P, s_\sharp Q). \tag{6}$$

Naturally, the upper bound of Theorem 3.2 is tight if there is no distribution shift, in which case $W_1(s_\sharp P, s_\sharp Q)$ evaluates to zero and the coverage gap is also zero by definition. Both (5) and (6) are valid upper bounds to the total coverage gap that are easy to compute in practice. It suffices to estimate the density of calibration scores—e.g., via kernel density estimation (KDE)—and compute the 1-Wasserstein distance in (6) or the difference of CDFs in (5), all of which are easily computable from samples, especially since nonconformity scores are typically unidimensional.

3.3 Upper Bound on the Total Coverage Gap without Labels

While the above bounds are informative, they come with one *crucial* drawback; they require *labeled* samples from Q, which might be hard to obtain in practice. To overcome the need for labels, we present another upper bound to the total coverage gap that can be computed with unlabeled data from the test distribution Q. The main insight is that, although we may not have access to the score of the ground truth label, in the classification setting, we generally know the scores of *all* possible labels. This gives us meaningful information on the distribution of scores under the shifted test distribution Q, which we use to construct auxiliary distributions $s_\sharp Q^\downarrow$ and $s_\sharp Q^\uparrow$, whose CDFs satisfy $F_{s_\sharp Q^\uparrow}(t) \leq F_{s_\sharp Q}(t) \leq F_{s_\sharp Q^\downarrow}(t)$ for all $t \in \mathbb{R}$. This sandwiching of the CDF of $F_{s_\sharp Q}(t)$ corresponds to a stochastic dominance relationship, denoted $s_\sharp Q^\uparrow \succcurlyeq s_\sharp Q \succcurlyeq s_\sharp Q^\downarrow$, and allows us to construct bounds in the form of (5) and (6) even without access to the unknown $s_\sharp Q$.

Theorem 3.3. *Let P and Q be two probability measures on $\mathcal{X} \times \mathcal{Y}$ with $s_\sharp P$ and $s_\sharp Q$ their respective pushforward measures by the score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Assume $s_\sharp P$ is absolutely continuous with respect to the Lebesgue measure with density $p_{s_\sharp P}(s_c)$. Further let $s_\sharp Q^\downarrow$ and $s_\sharp Q^\uparrow$ be such that $s_\sharp Q^\uparrow \succcurlyeq s_\sharp Q \succcurlyeq s_\sharp Q^\downarrow$. Then, we have that*

$$\Delta_{P,Q} \leq \frac{1}{2} \int p_{s_\sharp P}(s_c) \Bigg(\Big| F_{s_\sharp P}(s_c) - F_{s_\sharp Q^\uparrow}(s_c) \Big| + \Big| F_{s_\sharp P}(s_c) - F_{s_\sharp Q^\downarrow}(s_c) \Big| + F_{s_\sharp Q^\downarrow}(s_c) - F_{s_\sharp Q^\uparrow}(s_c) \Bigg) ds_c \tag{7}$$

$$\leq \frac{1}{2} \Big(\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c) \Big) \Bigg(W_1(s_\sharp P, s_\sharp Q^\uparrow) + W_1(s_\sharp P, s_\sharp Q^\downarrow) + \mathbb{E}_{s_\sharp Q^\uparrow}[S] - \mathbb{E}_{s_\sharp Q^\downarrow}[S] \Bigg). \tag{8}$$

Theorem 3.3 tells us that we can upper bound the total coverage gap between the calibration distribution P and an unknown test distribution Q, if we can somehow find two auxiliary distributions over the test scores $s_\sharp Q$, such that $s_\sharp Q^\uparrow \succcurlyeq s_\sharp Q \succcurlyeq s_\sharp Q^\downarrow$. Fortunately, in the classification setting, we typically evaluate the scores of all possible classes (e.g., by computing the probability of all classes with a softmax activation). Thus, although the true score $s(x, y)$ is not observed, we know it must be contained in the set $\boldsymbol{s}(x) = \{s(x, y') : y' \in \mathcal{Y}\}$. We can then use a set of m unlabeled samples from Q and their corresponding scores $\{\boldsymbol{s}(x_i)\}_{i=1}^m$ to construct empirical distributions $s_\sharp \hat{Q}_m^\uparrow$ and $s_\sharp \hat{Q}_m^\downarrow$ with the required stochastic dominance relation to the unknown $s_\sharp \hat{Q}_m$. A natural solution is to take the minimum and maximum scores of each instance x_i to get the following empirical distributions

$$s_\sharp \hat{Q}_m^{\min} := \frac{1}{m} \sum_{i=1}^m \delta_{\min \boldsymbol{s}(x_i)} \qquad s_\sharp \hat{Q}_m^{\max} := \frac{1}{m} \sum_{i=1}^m \delta_{\max \boldsymbol{s}(x_i)}, \tag{9}$$

with $\mathbf{1}(\cdot)$ the indicator function and δ_z the delta function at a value z. It is easy to see the empirical distributions obey $s_\sharp \hat{Q}_m^{\max} \succcurlyeq s_\sharp \hat{Q} \succcurlyeq s_\sharp \hat{Q}_m^{\min}$ as needed.

In the common setting where predictions come from a classifier f that outputs class probabilities, a practical alternative we found effective is to construct the auxiliary distributions $s_\sharp \hat{Q}_m^U$ and $s_\sharp \hat{Q}_m^f$ as

$$s_\sharp \hat{Q}_m^U := \frac{1}{m} \sum_{i=1}^m \delta_{s(x_i, y_i')}, y_i' \sim U(Y) \qquad s_\sharp \hat{Q}_m^f := \frac{1}{m} \sum_{i=1}^m \delta_{s(x_i, y_i')}, y_i' \sim Q_f(Y|x_i), \tag{10}$$

where we take the score of a random label j sampled either from a uniform distribution $U(Y)$ or from $Q_f(Y|X)$, the conditional distribution given by model f. The motivation here is that $s_\sharp \hat{Q}_m^U$ captures the scenario where the model is uninformative of the correct label, while $s_\sharp \hat{Q}_m^f$ reflects the scenario in which the model perfectly captures the true distribution $Q(Y|X)$. Although we cannot guarantee $s_\sharp \hat{Q}_m^U \succcurlyeq s_\sharp \hat{Q} \succcurlyeq s_\sharp \hat{Q}_m^f$, we empirically observe this relation to hold in most cases (see Fig. 1).

Naturally, the tightness of the upper bounds in Theorem 3.3 depend heavily on how close $s_\sharp Q^\downarrow$ and $s_\sharp Q^\uparrow$ are to the unknown $s_\sharp Q$. In the absence of prior knowledge about the nature of the distribution shift, the best we can do is rely on the general auxiliary distributions described above, which may yield relatively loose bounds. Nevertheless, the bounds constructed using $(s_\sharp \hat{Q}_m^{\min}, s_\sharp \hat{Q}_m^{\max})$ or $(s_\sharp \hat{Q}^f, s_\sharp \hat{Q}^U)$ serve as effective optimization objectives for reducing the coverage gap in practice, through a reweighting of the calibration samples, as we explain in the following section.

Before proceeding, we note that upper bounds on the coverage gap can also be derived for restricted ranges of the miscoverage rate α. In Appendix A.4, we present a bound for α ranging between α^- and α^+, with $0 \leq \alpha^- \leq \alpha^+ \leq 1$. Additionally, Appendix A.5 provides a bound for a fixed miscoverage rate α, denoted $\Delta_{P,Q}(\alpha)$. While these bounds are less effective as optimization objectives, they offer useful theoretical insights and are detailed in Appendix A.

4 Learning

In the conformal prediction literature, it is common to address non-exchangeability by reweighing the calibration points [2, 42]. In practice, this implies that the quantile of the scores is computed on a

 https://doi.org/10.52202/085713-5278

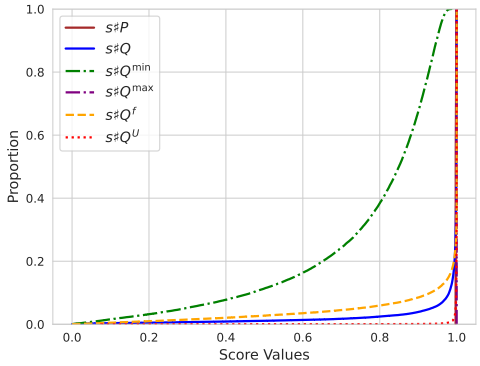

Figure 1: Empirical CDFs of nonconformity scores in ImageNet-C Gaussian noise under the calibration $s_\sharp \hat{P}$, test $s_\sharp \hat{Q}$, and auxiliary distributions. We can visually verify $s_\sharp \hat{Q}^{\max} \succcurlyeq s_\sharp \hat{Q} \succcurlyeq s_\sharp \hat{Q}^{\min}$ and $s_\sharp \hat{Q}^{U} \succcurlyeq s_\sharp \hat{Q} \succcurlyeq s_\sharp \hat{Q}^{f}$.

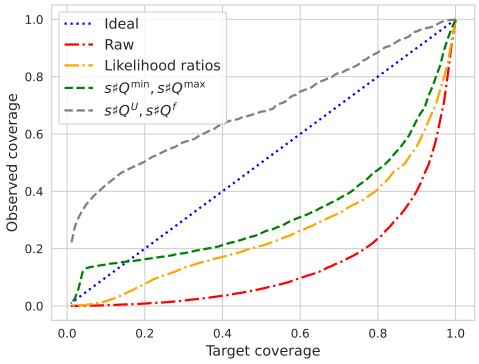

Figure 2: Total coverage gap in ImageNet-C Fog with weights learned via likelihood ratio estimation (orange), optimal transport with $(s_\sharp \hat{Q}^{\min}, s_\sharp \hat{Q}^{\max})$ in green, and $(s_\sharp \hat{Q}^{f}, s_\sharp \hat{Q}^{U})$ in gray.

weighted empirical distribution of the calibration scores $s_\sharp P_n^{\boldsymbol{w}}$ with the following empirical CDF

$$s_\sharp \hat{P}_n^{\boldsymbol{w}} = \sum_{i=1}^{n} w_i \delta_{s(x_i, y_i)}, \tag{11}$$

$w_i \geq 0$ are properly normalized weights associated with calibration samples (x_i, y_i). In the case of covariate shifts, Tibshirani et al. [42] show that we can recover proper coverage by setting the weights in (11) proportionally to the likelihood ratio, i.e., $w_i \propto {dQ(x_i)}/{dP(x_i)}$. Barber et al. [2] also rely on a weighted empirical distribution in the form of (11) but assume the weights to be fixed based on some prior knowledge of the likely deviations from exchangeability.

Motivated by these ideas, we propose instead to learn the distribution $s_\sharp P_n^{\boldsymbol{w}}$ directly by minimizing the upper bounds of Theorem 3.3 with respect to its weights $\boldsymbol{w}$. We replace the Wasserstein and CDF distances in these bounds with their empirical counterparts (see Appendix A.3 for details), enabling optimization from samples. Specifically, we assume access to n labeled samples $\{(x_i, y_i)\}_{i=1}^{n}$ from P and m unlabeled samples $\{x_i\}_{i=n+1}^{n+m}$ from the test distribution Q. These are used to construct $s_\sharp \hat{Q}^{\downarrow}$ and $s_\sharp \hat{Q}^{\uparrow}$. As discussed in Section 3, two alternatives we consider are to take the pair $(s_\sharp \hat{Q}^{\min}, s_\sharp \hat{Q}^{\max})$ as in (9) or $(s_\sharp \hat{Q}^{f}, s_\sharp \hat{Q}^{U})$ as in (10). However, other constructions, potentially leveraging prior information about the distribution shift or the application domain, are possible, as long as $s_\sharp \hat{Q}^{\uparrow} \succcurlyeq s_\sharp \hat{Q} \succcurlyeq s_\sharp \hat{Q}^{\downarrow}$. We note that, although using unlabeled samples from the test distribution is uncommon in CP, in many cases it is easy to collect such samples in practice.

Evaluating the bound of Theorem 3.3 admits an efficient exact solution for empirical distributions: it suffices to sort the samples (n from P and m from Q) to compute the difference between the empirical CDFs. Crucially, when computing weighted empirical CDFs as in (11), the weights $\boldsymbol{w}$ only come into play after the score values are sorted, and thus the operation is trivially differentiable with respect to $\boldsymbol{w}$, with no relaxation needed. Finally, we estimate the density of $p_{s_\sharp P}$ by fitting a Gaussian kernel density estimator (KDE) to the calibration scores. It is easy to fit Gaussian KDEs to weighted samples, and the estimated density is also differentiable with respect to the weights.

Having established how to evaluate the bounds efficiently and differentiate through the weighting, we now turn to how these weights are parameterized. We consider two strategies:

- **Free-form weights**. We directly optimize a set of unnormalized weights $\{\tilde{w}_i\}_{i=1}^{n}$, each one tied to a specific calibration point (x_i, y_i). After optimization, these weights are normalized to form the weighted empirical distribution in (11). This method is simple and effective, offering maximum flexibility for a fixed calibration set, but it remains restricted to that set.
- **Learnable weight function**. Alternatively, we learn a parametric function $w_\theta : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with $\tilde{w}_i = w_\theta\big(s(x_i, y_i)\big)$ where θ are the function parameters (e.g., a small neural network).

Algorithm 1 Learning Weights for Non-exchangeable Conformal Prediction via Optimal Transport

Input:
n labeled samples $\{(x_i, y_i)\}_{i=1}^n$ from P
m unlabeled samples $\{x_j\}_{j=n+1}^{n+m}$ from Q
score function s
Initialize unnormalized weights $\tilde{\boldsymbol{w}} = \{\tilde{w}_i\}_{i=1}^n$ or weight function w_θ
Compute calibration scores $\{s(x_i, y_i)\}_{i=1}^n$
Compute test score vectors $\{\boldsymbol{s}(x_j)\}_{j=n+1}^{n+m}$ // includes all candidate labels $\boldsymbol{s}(x) = \{s(x, y) : y \in \mathcal{Y}\}$
repeat
Construct $\hat{Q}^\downarrow$ and $\hat{Q}^\uparrow$ from $\{s(x_j)\}$ // e.g., $(\min, \max)$ or (f, U)
Compute normalized weights: $\boldsymbol{w} = \mathrm{softmax}(\tilde{\boldsymbol{w}})$
Fit KDE to $\{s(x_i, y_i)\}_{i=1}^n$ with weights $\boldsymbol{w}$
Update $\tilde{\boldsymbol{w}}$ or w_θ to minimize (7) or (8) // weighted-CDF or 1-Wasserstein bound
until convergence

Unlike the free-form approach, this formulation allows computing weights for additional calibration points or test candidates, thereby recovering the standard weighted split-CP setting and aligning with the importance-weighting principle of Tibshirani et al. [42].

Both parametrizations involve trade-offs. We focus most of our analysis on free-form weights because they align naturally with our bounds-based objectives, require only simple differentiable operations, and avoid extra modeling assumptions. This makes them stable and data-efficient in the small-sample regime, which is our primary concern. Nonetheless, these weights are tied to the specific calibration samples used during training, so the same set must be retained for calibration. This introduces dependencies among calibration points and breaks exchangeability with the test set. Under distribution shift, exchangeability is already compromised, so this violation is less critical. In this setting, the focus naturally shifts from preserving exchangeability to mitigating its effects, which is exactly what we achieve by optimizing calibration weights directly. Weight functions, by contrast, offer a more general solution: by mapping scores to weights through a parametric model, they can assign weights to unseen calibration points and recover the standard weighted split-CP setting. This flexibility comes at a cost: training the model requires additional labeled samples from P and careful specification of its architecture. Despite these differences, both approaches achieve comparable empirical performance (see experiments in Section 6 and Appendix C).

See Algorithm 1 for a sketch of how we optimize the total coverage gap from Theorem 3.3. A more complete algorithm, including how this optimization fits into split CP is given in Algorithm 2, in Appendix C. In practice, during optimization unnormalized weights $\{\tilde{w}_i\}$ are mapped to normalized weights via a softmax, regardless of whether $\tilde{w}_i$ come from a learnable vector or a weight function w_θ. This normalization ensures differentiability and proper scaling before computing the empirical distribution used to evaluate the 1-Wasserstein or weighted-CDF bounds and fit the KDE. The computational cost remains the same in both cases: evaluating the 1-Wasserstein distance requires $\mathcal{O}((m+n)\log(m+n))$ for sorting, while fitting a Gaussian KDE on n calibration samples costs $\mathcal{O}(k \cdot n)$, where k is the number of evaluation points.

Regression setting In principle, our upper bound to the total coverage gap in Theorem 3.3 is directly applicable to regression tasks. The only caveat is that the true score $s(X_t, Y_t)$ might no longer be restricted to a finite set of known values, as in the classification setting, and designing $s_\sharp Q^\downarrow$ and $s_\sharp Q^\uparrow$ is more challenging. One can always construct these auxiliary distributions based on some prior information about the task or the underlying distribution shift, but more generally it is possible to leverage the regression-as-classification framework [16], and again use $(s_\sharp \hat{Q}^{\min}, s_\sharp \hat{Q}^{\max})$ or $(s_\sharp \hat{Q}^f, s_\sharp \hat{Q}^U)$, as we do successfully in the experiment described in Section 6.1.

5 Related Work

Several works have studied conformal prediction in the non-exchangeable setting, especially in the context of time series, where the exchangeability assumption is violated by the very autoregressive nature of these problems [6, 13, 14, 18, 28, 46, 47, 50]. Closer to our work, Tibshirani et al. [42]

 https://doi.org/10.52202/085713-5278

and Barber et al. [2] have also proposed to mitigate the coverage gap by reweighing the calibration samples. However, in [42] their weights only address covariate shifts and correspond to the unknown likelihood ratio $dQ(x)/dP(x)$, which is hard to estimate in practice, especially under severe distribution shifts and the density chasm problem [34]. In our experiments, our methods compare favorably to learned likelihood ratios as proposed in [42], attesting to the difficulty of learning accurate ratios.

To our knowledge, our methods and the reweighing scheme of [2] are the only capable of tackling arbitrary distribution shifts in split conformal prediction. Still, their approach involves data-independent weights that must be designed a priori using some prior knowledge about the underlying distribution shift, whereas we learn appropriate weights directly from a few unlabeled samples from the test distribution. With the exception of the work of [32], where label shift is also tackled via likelihood ratios in a similar fashion to [42], most other works focus on covariate shift [19, 21, 25]. Among these, [15, 23, 33, 49] are notable for tackling covariate shifts by approximating conditional coverage guarantees, i.e., by approximately satisfying $\mathbb{P}(Y_t \in \mathcal{C}(X_t)|X_t) \geq 1 - \alpha$. These achieve impressive results but are computationally expensive or limited to specific types of covariate shifts.

Of special note is the work of [15], which proposes to adapt the conformal threshold for each test point, providing conditional coverage guarantees if the distribution shift comes from a known function class. In our experiments, where it is not clear how to define such function class, their method—implemented in its most general form via radial basis function (RBF) kernels—produced larger prediction sets than our methods at a much larger computational cost at test time. We also compare against entropy scaled conformal prediction or ECP [21]. ECP consists in dividing the threshold, i.e., $\mathbb{Q}_\alpha(\mathcal{S}_c)$, by the $(1-\alpha)$ quantile of the entropy over class predictions for the test points, with the intuition that high entropy (indicating high uncertainty) will decrease the threshold, leading to larger prediction sets. While this heuristic proved effective for covariate shift, the improvements in coverage were modest in the context of label shift. In contrast, our methods demonstrate greater robustness to different types of shift (see Table 6).

Finally, our methods are closely related to the concurrent work of [48], which also explores the relationship between the coverage gap and the $W_1(P, Q)$. However, their results are derived through a different approach and bound $\Delta_{P,Q}(\alpha)$ for any α. Unfortunately, their bound does not depend on α, and is thus loose for most target coverage rates. Moreover, in practice, their methods are only applicable to distribution shifts where the test distribution is a mixture of different calibration distributions. In contrast, our bound from Theorem 3.3 can be applied to any type distribution shift.

6 Experiments

In this section, we describe and analyze a set of experiments designed to evaluate and validate our methods. In each of them, we have two sets of samples, $\mathcal{D}_P$ distributed according to some calibration distribution P and $\mathcal{D}_Q$ according to some test distribution Q, with P differing from Q via some form of distribution shift. We divide $\mathcal{D}_Q$ into two, with $\mathcal{D}_Q^{(1)}$ reserved for fitting a density ratio estimator or learning the weights in our method as in Algorithm 1, and $\mathcal{D}_Q^{(2)}$ used for testing. When no pretrained model is available, we also split $\mathcal{D}_P$, using $\mathcal{D}_P^{(1)}$ for training a model and $\mathcal{D}_P^{(2)}$ for calibration in split CP. Main experiments use 300 samples for $\mathcal{D}_P^{(2)}$ and $\mathcal{D}_Q^{(1)}$, with $\mathcal{D}_P^{(2)}$ extended to 600 for weight functions (split evenly for fitting and calibration). Sample size effects are analyzed in Appendix C.

For regression tasks, we optimize the 1-Wasserstein variant of our bounds (8), while for image classification tasks, we adopt the weighted-CDF formulation (7). These choices reflect empirical findings; each variant performs best in its respective domain, as discussed in Appendix C.5. In all cases, we define nonconformity scores as one minus the probabilities assigned by the model.

6.1 Regression setting with synthetic data

We start with the synthetic data experiment proposed in [49], where the ground truth likelihood ratios are known. We use the regression-as-classification method of Guha et al. [16], splitting the output space into 50 equally spaced bins. In Figure 3, our methods significantly enhance coverage in most cases, with no notable difference between the two variants. In this low-dimensional setting, learning the likelihood ratios also proved effective, albeit with a slight tendency to under-cover. In contrast, our method exhibited a mild bias toward over-coverage. See Appendix C for experimental details.

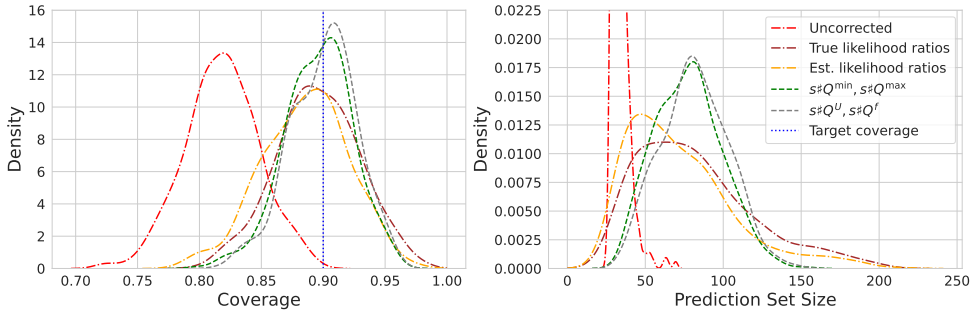

Figure 3: Distribution of coverage and prediction set sizes for the synthetic regression task across 500 simulations and target coverage rate of 90% (blue vertical line). For ease of visualization, we plot the density estimated with a KDE fit to the 500 observations.

6.2 Image classification

Imagenet-C We use the established ImageNet-C [17] dataset to test our methods under the covariate and label shift settings. ImageNet-C contains a total of 15 covariate shifts at different severity levels, from 1 (least severe) to 5 (most severe), but no label shift. Since our methods apply to any type of distribution shift, we also simulate label shift in ImageNet-C. The details can be found in Appendix C. In Table 1 (and its extended version in Appendix C) we observe a drastic drop in coverage for the uncorrected scores; from the target coverage of 90%, we drop to an average of $\sim 80\%$ for severity level 1, and to an average of $\sim 30\%$ for severity level 5. By introducing weighting to the calibration set, our OT methods improves coverage across the board, achieving similar coverage with and without label shift. In contrast, learned likelihood ratios provided modest improvements in coverage and the ECP method of [21] only produced competitive results in the absence of label shift. Finally, the method of [15] produced good results in terms of coverage in ImageNet-C but at the cost of excessive large prediction sets (see Section 6.3 for a discussion on prediction set sizes).

iWildCam We further use iWildCam [3] as one more dataset that contains natural distribution shifts. We can see in Table 1 that the distribution shift incurs roughly a 10% drop in coverage when not correcting the scores. While likelihood ratio weighting produce only modest improvements over the uncorrected scores, both of our OT settings improve coverage. The $(\min, \max)$ setting increased coverage by more than 7% on average, while the (f, U) setting improved coverage by almost 10%. Finally, the ECP method [21] got almost perfect coverage, even under label shift. Interestingly, the method of [15] actually hurt coverage. This could be either because the distribution shift in iWildCam is not well captured by the class of shifts considered in their method (given by RBF kernels in this experiment) or due to severe class imbalance in iWildCam, which might hamper the optimization.

Notably, the results were consistent across parametrization choices, with both the free-form and weight-function variants yielding very similar performance overall. The only meaningful differences emerged under relatively mild distribution shifts: the weight-function approach performed better on iWildCam, while the free-form variant showed stronger results on ImageNet-C at severity level one.

6.3 Discussion

Prediction set sizes. Our methods, like most other approaches to non-exchangeable CP, including the likelihood ratios proposed by Tibshirani et al. [42], and the weights introduced by Barber et al. [2], do not alter the observed test scores, thereby preserving the ranking of classes under the shifted distribution. Consequently, similar to these other works, our analysis focuses on coverage. The underlying CP algorithm (split CP) remains unchanged, and the variance in prediction set sizes is attributed to miscoverage. The best achievable performance mirrors what would be obtained if labeled samples from Q were available; undercoverage results in smaller-than-optimal prediction sets, while overcoverage leads to larger-than-optimal sets. The only exception to this is the method of Gibbs et al., which aims for conditional coverage and thus changes the conformal threshold for each test point. This decouples coverage and prediction set size, but in our experiments their methods produced larger prediction sets than other methods despite getting close to the target coverage.

 https://doi.org/10.52202/085713-5278

Table 1: **Average coverage and prediction set size on image classification tasks, with and without label shift**. Results for uncorrected distributions, calibrating and testing on Q (Oracle), likelihood ratios (LR), the methods of Kasa et al. [21] and Gibbs et al. [15], and our methods with weighted-CDF objective (7), including $(\min, \max)$ and (f, U) variants, and free-form (FF) and weight function (WF) parametrizations. The target coverage is set to 90%. Extended version in Table 6.

		iWildCam		ImageNet-C Sev. 1		ImageNet-C Sev. 3		ImageNet-C Sev. 5	
		Cov.	Size	Cov.	Size	Cov.	Size	Cov.	Size
no label shift	Uncorrected	$78.2_{\pm 3.0}$	$22.1_{\pm 4.4}$	$78.7_{\pm 6.0}$	$2.7_{\pm 0.7}$	$58.7_{\pm 14.3}$	$3.3_{\pm 1.2}$	$29.8_{\pm 18.9}$	$3.3_{\pm 1.5}$
	Oracle	$89.7_{\pm 1.4}$	$50.5_{\pm 5.3}$	$89.8_{\pm 1.6}$	$10.7_{\pm 7.2}$	$89.9_{\pm 1.7}$	$80.1_{\pm 71.4}$	$89.7_{\pm 1.9}$	$338.6_{\pm 190.9}$
	LR	$79.0_{\pm 2.6}$	$23.3_{\pm 4.8}$	$84.1_{\pm 4.9}$	$5.2_{\pm 3.8}$	$71.3_{\pm 11.9}$	$12.3_{\pm 13.1}$	$47.4_{\pm 20.2}$	$27.7_{\pm 37.0}$
	Kasa et al.	$95.2_{\pm 1.2}$	$79.7_{\pm 9.0}$	$96.6_{\pm 1.1}$	$42.4_{\pm 19.2}$	$93.7_{\pm 3.2}$	$119.3_{\pm 65.4}$	$86.1_{\pm 8.4}$	$282.0_{\pm 159.1}$
	Gibbs et al.	$67.7_{\pm 6.0}$	$109.6_{\pm 6.3}$	$88.9_{\pm 2.8}$	$548.5_{\pm 36.7}$	$85.6_{\pm 3.6}$	$635.7_{\pm 66.3}$	$84.9_{\pm 5.6}$	$754.4_{\pm 102.7}$
	FF $(\min, \max)$	$85.2_{\pm 4.1}$	$37.5_{\pm 10.9}$	$91.4_{\pm 3.9}$	$15.1_{\pm 9.7}$	$88.2_{\pm 7.7}$	$63.9_{\pm 41.3}$	$71.3_{\pm 17.0}$	$128.5_{\pm 93.0}$
	WF $(\min, \max)$	$89.8_{\pm 2.9}$	$51.7_{\pm 11.1}$	$94.6_{\pm 6.3}$	$34.7_{\pm 20.6}$	$87.0_{\pm 14.1}$	$69.7_{\pm 46.6}$	$71.5_{\pm 17.1}$	$130.5_{\pm 94.1}$
	FF (f, U)	$88.2_{\pm 3.4}$	$46.2_{\pm 11.9}$	$93.0_{\pm 3.0}$	$19.2_{\pm 11.4}$	$90.3_{\pm 5.9}$	$79.8_{\pm 48.9}$	$79.5_{\pm 12.1}$	$205.7_{\pm 164.4}$
	WF (f, U)	$88.9_{\pm 4.6}$	$49.7_{\pm 14.1}$	$95.7_{\pm 2.2}$	$36.6_{\pm 20.1}$	$90.1_{\pm 6.2}$	$79.7_{\pm 50.1}$	$78.7_{\pm 13.1}$	$200.4_{\pm 164.2}$
with label shift	Uncorrected	$79.2_{\pm 7.1}$	$20.9_{\pm 5.4}$	$79.2_{\pm 8.2}$	$2.7_{\pm 0.8}$	$59.2_{\pm 15.7}$	$3.2_{\pm 1.2}$	$29.8_{\pm 20.1}$	$3.3_{\pm 1.5}$
	Oracle	$90.2_{\pm 4.3}$	$43.7_{\pm 20.7}$	$90.3_{\pm 4.5}$	$15.0_{\pm 19.9}$	$90.2_{\pm 3.8}$	$88.1_{\pm 87.4}$	$90.5_{\pm 4.0}$	$349.0_{\pm 220.0}$
	LR	$81.5_{\pm 5.7}$	$24.6_{\pm 4.4}$	$84.5_{\pm 7.1}$	$6.1_{\pm 5.5}$	$72.5_{\pm 12.9}$	$14.3_{\pm 15.8}$	$47.3_{\pm 21.2}$	$27.7_{\pm 33.5}$
	Kasa et al.	$88.6_{\pm 6.3}$	$35.9_{\pm 17.1}$	$79.2_{\pm 8.1}$	$2.7_{\pm 0.8}$	$59.7_{\pm 15.2}$	$3.4_{\pm 1.4}$	$31.2_{\pm 19.9}$	$4.1_{\pm 2.3}$
	Gibbs et al.	$38.0_{\pm 15.6}$	$44.9_{\pm 23.5}$	$88.6_{\pm 3.8}$	$552.3_{\pm 46.0}$	$85.5_{\pm 4.6}$	$638.6_{\pm 70.2}$	$84.8_{\pm 5.8}$	$753.5_{\pm 100.4}$
	FF $(\min, \max)$	$83.1_{\pm 5.8}$	$22.1_{\pm 12.6}$	$91.0_{\pm 5.7}$	$14.3_{\pm 10.2}$	$88.4_{\pm 8.6}$	$62.6_{\pm 41.4}$	$72.0_{\pm 18.8}$	$127.9_{\pm 94.4}$
	WF $(\min, \max)$	$91.3_{\pm 4.3}$	$52.2_{\pm 15.2}$	$93.7_{\pm 9.3}$	$33.1_{\pm 21.5}$	$87.3_{\pm 12.8}$	$69.4_{\pm 47.3}$	$70.4_{\pm 19.2}$	$129.2_{\pm 96.3}$
	FF (f, U)	$88.2_{\pm 4.7}$	$33.5_{\pm 12.9}$	$93.5_{\pm 3.6}$	$20.7_{\pm 14.6}$	$90.3_{\pm 6.8}$	$78.9_{\pm 51.7}$	$79.8_{\pm 13.9}$	$206.0_{\pm 168.3}$
	WF (f, U)	$91.0_{\pm 3.8}$	$49.8_{\pm 10.9}$	$95.1_{\pm 7.1}$	$36.3_{\pm 21.8}$	$90.4_{\pm 7.0}$	$80.0_{\pm 51.4}$	$79.2_{\pm 14.6}$	$201.3_{\pm 168.6}$

Bound variants. In Theorems 3.2 and 3.3, we have two flavors of upper bounds: one expressed in the terms of a weighted distance of CDFs, and another using the 1-Wasserstein distance. The former is always tighter and, likely for this reason, has shown superior performance in image classification tasks. Conversely, the 1-Wasserstein bounds, while generally looser, establish a more natural connection to optimal transport theory and has shown better empirical performance in regression tasks. This may be due to the weighting by $p_{s_\sharp P}(s_c)$ in (5) and (7), which might complicate optimization.

Number of samples. Our methods are effective across varying numbers of labeled samples from P and unlabeled samples from Q. As shown in Appendix C.4.3, coverage improves with more samples as expected, but meaningful gains are observed even with as few as 30 samples from each distribution.

Limitations. The tightness of the unlabeled upper bounds in Theorem 3.3 depends on the auxiliary distributions $s_\sharp Q^\downarrow$ and $s_\sharp Q^\uparrow$. When these closely approximate $s_\sharp Q$, the resulting bounds are tight, and optimizing either (7) or (8) is predictably highly effective in reducing the coverage gap. However, in the practically interesting setting of a general and unknown distribution shift we consider, the available choices for $s_\sharp Q^\downarrow$ and $s_\sharp Q^\uparrow$ are likely less informative and yield a necessarily looser bound. Having said that, the bounds computed with $(s_\sharp \hat{Q}^{\min}, s_\sharp \hat{Q}^{\max})$ and $(s_\sharp \hat{Q}^f, s_\sharp \hat{Q}^U)$ still perform surprisingly well in reducing the coverage gap when used for reweighing the calibration data, demonstrating their practical value and broad applicability. Nevertheless, care must be taken: in cases of minimal or no distribution shift, $s_\sharp P$ may already offer better coverage than the solution to our objective. It is therefore advisable to first assess the presence of a distribution shift, potentially using unlabeled samples from Q, before applying our methods. We explore this issue further in Appendix C.

7 Conclusion

In this work, we employ optimal transport theory to study the effect of distribution shifts on conformal prediction. Specifically, we derive upper bounds on the total coverage gap induced by a shift from the calibration distribution P to the test distribution Q, expressed in terms of (weighted) CDF and Wasserstein distances. Recognizing that labeled examples from Q are often unavailable in practice, we extend our analysis by leveraging the structure inherent in the nonconformity scores of the unlabeled test data. To this end, we construct auxiliary distributions $s_\sharp \hat{Q}^\downarrow$ and $s_\sharp \hat{Q}^\uparrow$, which enable label-free bounds of the coverage gap. Furthermore, we utilize these bounds as optimization objectives to learn importance weights over the calibration data. Empirically, our approach significantly reduces the coverage gap across a range of distribution shift settings.

References

[1] Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[2] Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

[3] Beery, S., Cole, E., and Gjoka, A. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

[4] Bellotti, A. Optimized conformal classification using gradient descent approximation. *arXiv preprint arXiv:2105.11255*, 2021.

[5] Caprio, M., Stutz, D., Li, S., and Doucet, A. Conformalized credal regions for classification with ambiguous ground truth. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=L7sQ8CW2FY`.

[6] Chernozhukov, V., Wüthrich, K., and Yinchu, Z. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference on learning theory*, pp. 732–749. PMLR, 2018.

[7] Correia, A., Massoli, F. V., Louizos, C., and Behboodi, A. An information theoretic perspective on conformal prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 6 November 2024.

[8] De Angelis, M. and Gray, A. Why the 1-wasserstein distance is the area between the two marginal cdfs. *arXiv preprint arXiv:2111.03570*, 2021.

[9] Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

[10] Farinhas, A., Zerva, C., Ulmer, D. T., and Martins, A. Non-exchangeable conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2): 455–482, 2021.

[12] Garg, S., Erickson, N., Sharpnack, J., Smola, A., Balakrishnan, S., and Lipton, Z. C. Rlsbench: Domain adaptation under relaxed label shift. In *International Conference on Machine Learning*, pp. 10879–10928. PMLR, 2023.

[13] Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

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

[15] Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

[16] Guha, E. K., Natarajan, S., Möllenhoff, T., Khan, M. E., and Ndiaye, E. Conformal prediction via regression-as-classification. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[18] Jensen, V., Bianchi, F. M., and Anfinsen, S. N. Ensemble conformalized quantile regression for probabilistic time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

 https://doi.org/10.52202/085713-5278

[19] Jin, Y., Ren, Z., and Candès, E. J. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6): e2214889120, 2023.

[20] Jones, M. C. Simple boundary correction for kernel density estimation. *Statistics and computing*, 3:135–146, 1993.

[21] Kasa, K., Zhang, Z., Yang, H., and Taylor, G. W. Adapting conformal prediction to distribution shifts without labels. *arXiv preprint arXiv:2406.01416*, 2024.

[22] Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Kiyani, S., Pappas, G., and Hassani, H. Length optimization in conformal prediction. *arXiv preprint arXiv:2406.18814*, 2024.

[24] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

[25] Lei, L. and Candès, E. J. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

[26] Liu, Y., Kothari, P., Van Delft, B., Bellot-Gurlet, B., Mordan, T., and Alahi, A. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021.

[27] maintainers, T. and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

[28] Oliveira, R. I., Orenstein, P., Ramos, T., and Romano, J. V. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38, 2024.

[29] Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pp. 345–356. Springer, 2002.

[30] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[31] Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[32] Podkopaev, A. and Ramdas, A. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pp. 844–853. PMLR, 2021.

[33] Qiu, H., Dobriban, E., and Tchetgen Tchetgen, E. Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1680–1705, 2023.

[34] Rhodes, B., Xu, K., and Gutmann, M. U. Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916, 2020.

[35] Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

[36] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[37] Scott, D. W. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[38] Shafer, G. and Vovk, V. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[39] Stutz, D., Dvijotham, K. D., Cemgil, A. T., and Doucet, A. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022.

[40] Stutz, D., Roy, A. G., Matejovicova, T., Strachan, P., Cemgil, A. T., and Doucet, A. Conformal prediction under ambiguous ground truth. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=CAd6V2qXxc`.

[41] Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

[42] Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[43] Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[44] Vovk, V. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

[45] Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, second edition, December 2022. ISBN 978-3-031-06648-1. doi: 10.10007/978-3-031-06649-8.

[46] Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

[47] Xu, C. and Xie, Y. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11575–11587, 2023.

[48] Xu, R., Chen, C., Sun, Y., Venkitasubramaniam, P., and Xie, S. Wasserstein-regularized conformal prediction under general distribution shift. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=aJ3tiX1Tu4`.

[49] Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkae009, 2024.

[50] Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR, 2022.

 https://doi.org/10.52202/085713-5278

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our theoretical results are all proved in the supplementary material, and our experiments support the claim that our methods are effective at mitigating the coverage gap in conformal prediction under a variety of distribution shift settings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We elaborate on the limitations of our method in Section 6.3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are accompanied by the necessary assumptions and thorough derivations. The complete proofs are all in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our experimental settings in details in the supplementary material. We also intend to make the source code available, subject to company approval and applicable policies.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

 https://doi.org/10.52202/085713-5278

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used in this paper are already open source and readily available. We intend to make the source code available, subject to company approval and applicable policies.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All relevant details for all experiments are described in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We repeat all experiments across different random seeds. For the image classification tasks, we use 10 random seeds and report the mean and standard deviation of these results. For the synthetic regression task, we use 500 random seeds and plot the distribution of empirical coverage and prediction set sizes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run in a single commercial GPU, as mentioned in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: All paper is primarily theoretical in nature and of limited direct societal impacts. Yet, we do discuss its broader impact in the supplementary material.

https://doi.org/10.52202/085713-5278

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All resources used in the paper are appropriately cited and referenced.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce any new assets at the moment. We intend to make the source code available, subject to company approval and applicable policies. That would come with the proper documentation for reproducing the experiments.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer:[NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/085713-5278

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core methods of this research do not relate to LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Broader Impact

This work introduces new methods to enhance the coverage of conformal prediction under distribution shifts, which frequently occur in real-world applications. We believe our contributions will have a positive impact, encouraging practitioners to adopt uncertainty quantification techniques like conformal prediction, provided the underlying guarantees are well understood.

A Proofs and Additional Theoretical Results

In this section, we provide detailed proofs of our new upper bounds to the coverage gap as well as extra theoretical results. We start by providing the proofs of Theorems 3.2 and 3.3.

A.1 Bound to the Total Coverage Gap

Theorem 3.2. *Let P and Q be probability measures on $\mathcal{X} \times \mathcal{Y}$ with $s_\sharp P$ and $s_\sharp Q$ their respective pushforward measures by a score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Assume $s_\sharp P$ is absolutely continuous with respect to the Lebesgue measure with density $p_{s_\sharp P}(s_c)$. Then the total coverage gap can be upper bounded as follows*

$$\Delta_{P,Q} \leq \int_{\mathbb{R}} p_{s_\sharp P}(s_c) \left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q}(s_c) \right| ds_c \tag{5}$$

$$\leq \Big(\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c) \Big) W_1(s_\sharp P, s_\sharp Q). \tag{6}$$

Proof.

$$\Delta_{P,Q} = \int_0^1 \Bigg| \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n} \left[\mathbb{E}_{S_t \sim s_\sharp P} [\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha(\mathcal{S}_c)\right)] \right] - \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n} \left[\mathbb{E}_{S_t \sim s_\sharp Q} [\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha(\mathcal{S}_c)\right)] \right] \Bigg| d\alpha \tag{12}$$

$$= \int_0^1 \left| \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n} \left[F_{s_\sharp P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha(\mathcal{S}_c)) \right] \right| d\alpha \tag{13}$$

$$\leq \int_0^1 \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n} \left[\left| F_{s_\sharp P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha(\mathcal{S}_c)) \right| \right] d\alpha \tag{14}$$

$$= \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n} \left[\sum_{i=1}^n \frac{1}{n} \left| F_{s_\sharp P}(s_c^{(i)}) - F_{s_\sharp Q}(s_c^{(i)}) \right| \right] \tag{15}$$

$$= \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n} \left[\left| F_{s_\sharp P}(S_c) - F_{s_\sharp Q}(S_c) \right| \right] \tag{16}$$

$$= \int_{\mathbb{R}} p_{s_\sharp P}(s_c) \left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q}(s_c) \right| ds_c \tag{17}$$

$$\leq \left[\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c) \right] \int_{\mathbb{R}} \left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q}(s_c) \right| ds_c \tag{18}$$

$$= \left[\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c) \right] W_1(s_\sharp P, s_\sharp Q). \tag{19}$$

where the first inequality in (14) is due to Jensen's inequality and (15) holds because the empirical quantile $\mathbb{Q}_\alpha(\mathcal{S}_c)$ must evaluate to one of the n values in $\mathcal{S}_c = \{s_c^{(1)}, \ldots, s_c^{(n)}\}$, each of which takes $\frac{1}{n}$ of the $[0, 1]$ range. In (15), we have the expectation of the sample mean, which equals the expectation of the population as in (16). Lastly, (18) holds because $p_{s_\sharp P}(s_c)$ is a density and thus non-negative everywhere, and (19) follows directly from the definition of the 1-Wasserstein distance, as in (4). □

https://doi.org/10.52202/085713-5278

Remark A.1 (On the expectation under a weighted measure). The bound in Theorem 3.2 extends to any weighted calibration measure P_ρ with density $\rho(s)p_{s_\sharp P}(s)$, where $\rho : \mathbb{R} \to [0, \infty)$ satisfies

$$\int_{\mathbb{R}} \rho(s)p_{s_\sharp P}(s)\, ds = 1.$$

In this case, the total coverage gap is upper bounded by

$$\Delta_{P_\rho,Q} \leq \int_{\mathbb{R}} \rho(s)p_{s_\sharp P}(s) \left|F_{s_\sharp P_\rho}(s) - F_{s_\sharp Q}(s)\right| ds \leq \Big(\sup_{s\in\mathbb{R}} \rho(s)p_{s_\sharp P}(s)\Big) W_1(s_\sharp P_\rho, s_\sharp Q).$$

The proof follows identically by replacing $p_{s_\sharp P}$ with $\rho(s)p_{s_\sharp P}(s)$ in the argument.

Remark A.2 (On practical weighting). In experiments (Section 4), weights are applied at the sample level, forming a weighted empirical measure rather than a continuous density. These weights are not globally normalized; instead, normalization is enforced through a softmax or similar constraint during optimization. Our theoretical result assumes a normalized weighting function ρ, ensuring P_ρ is a probability measure. This is a stronger condition than what is used in practice, but the empirical approach approximates this normalization. Moreover, the bound holds for the empirical distribution defined by a calibration dataset as shown in Proposition A.3 below.

Proposition A.3 (Empirical weighted bound)**.** Let $s_\sharp \hat{P}_n^{\boldsymbol{w}} = \sum_{i=1}^n w_i\, \delta_{s_c^{(i)}}$ be the weighted empirical measure over calibration scores $\{s^{(i)}\}_{i=1}^n$, with weights $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$. Let $F_{s_\sharp \hat{P}_n^{\boldsymbol{w}}}$ and $F_{s_\sharp Q}$ denote the CDFs of $s_\sharp \hat{P}_n^{\boldsymbol{w}}$ and an arbitrary test distribution over scores $s_\sharp Q$, respectively. Define the weighted quantile function $\mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha) := F^{-1}_{s_\sharp \hat{P}_n^{\boldsymbol{w}}}(\alpha)$ and the empirical weighted coverage gap

$$\widehat{\Delta}_{w,Q} := \int_0^1 \left|\mathbb{E}_{\mathcal{S}_c\sim s_\sharp \hat{P}_n^{\boldsymbol{w}}} \Big[\mathbb{E}_{S_t\sim s_\sharp P}\big[\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha)\right)\big] - \mathbb{E}_{S_t\sim s_\sharp Q}\big[\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha)\right)\big]\Big]\right| d\alpha.$$

Then

$$\widehat{\Delta}_{w,Q} \;\leq\; \sum_{i=1}^n w_i \left|F_{s_\sharp \hat{P}_n^{\boldsymbol{w}}}(s_c^{(i)}) - F_{s_\sharp Q}(s_c^{(i)})\right|.$$

Proof. By definition and the tower property,

$$\widehat{\Delta}_{w,Q} = \int_0^1 \left|\mathbb{E}_{\mathcal{S}_c\sim s_\sharp \hat{P}_n^{\boldsymbol{w}}} \left[F_{s_\sharp \hat{P}_n^{\boldsymbol{w}}}(\mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha))\right]\right| d\alpha.$$

Applying Jensen's inequality (absolute value is convex),

$$\widehat{\Delta}_{w,Q} \leq \mathbb{E}_{\mathcal{S}_c\sim s_\sharp \hat{P}_n^{\boldsymbol{w}}} \left[\int_0^1 \left|F_{s_\sharp \hat{P}_n^{\boldsymbol{w}}}(\mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha))\right| d\alpha\right].$$

Since $s_\sharp \hat{P}_n^{\boldsymbol{w}}$ is a discrete distribution with atoms $\{s_c^{(i)}\}_{i=1}^n$ and masses $\{w_i\}_{i=1}^n$, its quantile map $\alpha \mapsto \mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha)$ takes the value $s_c^{(i)}$ on an interval of length exactly w_i. Therefore, for any fixed realization of $\mathcal{S}_c$,

$$\int_0^1 \left|F_{s_\sharp \hat{P}_n^{\boldsymbol{w}}}(\mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha^{\boldsymbol{w}}(\alpha))\right| d\alpha = \sum_{i=1}^n w_i \left|F_{s_\sharp \hat{P}_n^{\boldsymbol{w}}}(s_c^{(i)}) - F_{s_\sharp Q}(s_c^{(i)})\right|.$$

Taking expectation with respect to $\mathcal{S}_c \sim s_\sharp \hat{P}_n^{\boldsymbol{w}}$ does not change the right-hand side, which is deterministic given $\hat{P}_n^{\boldsymbol{w}}$, and the claim follows. □

A.2 Unlabeled Bound to the Total Coverage Gap

Before proving Theorem 3.3, we begin by recalling the concept of stochastic dominance, which plays a key role in the argument. Specifically, if A and B are two probability distributions on $\mathbb{R}$, we say A dominates B, denoted $A \succcurlyeq B$, if $F_A(t) \leq F_B(t)$ for all $t \in \mathbb{R}$. This notion is especially useful in our context, as will be made clear in the derivations. In particular, stochastic dominance also simplifies the computation of the 1-Wasserstein distance, as captured by the following well-known result.

Lemma A.4 (De Angelis & Gray [8]). *Let A and B be two probability distributions on $\mathbb{R}$, with $A \succcurlyeq B$, then*

$$W_1(A, B) = \mathbb{E}_A[X] - \mathbb{E}_B[X]$$

Theorem 3.3. *Let P and Q be two probability measures on $\mathcal{X} \times \mathcal{Y}$ with $s_\sharp P$ and $s_\sharp Q$ their respective pushforward measures by the score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Assume $s_\sharp P$ is absolutely continuous with respect to the Lebesgue measure with density $p_{s_\sharp P}(s_c)$. Further let $s_\sharp Q_m^\downarrow$ and $s_\sharp Q_m^\uparrow$ be such that $s_\sharp Q_m^\uparrow \succcurlyeq s_\sharp Q \succcurlyeq s_\sharp Q_m^\downarrow$. Then, we have that*

$$\Delta_{P,Q} \leq \frac{1}{2} \int p_{s_\sharp P}(s_c) \Bigg(\left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \right| + \left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\downarrow}(s_c) \right| + F_{s_\sharp Q_m^\downarrow}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \Bigg) ds_c \tag{7}$$

$$\leq \frac{1}{2} \left[\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c) \right] \left(W_1(s_\sharp P, s_\sharp Q_m^\uparrow) + W_1(s_\sharp P, s_\sharp Q_m^\downarrow) + \mathbb{E}_{s_\sharp Q_m^\uparrow}[S] - \mathbb{E}_{s_\sharp Q_m^\downarrow}[S] \right). \tag{8}$$

Proof of (7). We start from (5) and apply the triangle inequality twice to get

$$\Delta_{P,Q} \leq \int p_{s_\sharp P}(s_c) \left(\left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \right| + \left| F_{s_\sharp Q}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \right| \right) ds_c$$

$$\Delta_{P,Q} \leq \int p_{s_\sharp P}(s_c) \left(\left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\downarrow}(s_c) \right| + \left| F_{s_\sharp Q}(s_c) - F_{s_\sharp Q_m^\downarrow}(s_c) \right| \right) ds_c$$

Since all values in these inequalities are non-negative, we can add them up to get

$$\begin{aligned}
\Delta_{P,Q} &\leq \frac{1}{2} \int p_{s_\sharp P}(s_c) \Bigg(\left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \right| + \left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\downarrow}(s_c) \right| \\
&\qquad + \left| F_{s_\sharp Q}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \right| + \left| F_{s_\sharp Q}(s_c) - F_{s_\sharp Q_m^\downarrow}(s_c) \right| \Bigg) ds_c \\
&= \frac{1}{2} \int p_{s_\sharp P}(s_c) \Bigg(\left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \right| + \left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\downarrow}(s_c) \right| \\
&\qquad + F_{s_\sharp Q}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) + F_{s_\sharp Q_m^\downarrow}(s_c) - F_{s_\sharp Q}(s_c) \Bigg) ds_c \\
&= \frac{1}{2} \int p_{s_\sharp P}(s_c) \Bigg(\left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \right| + \left| F_{s_\sharp P}(s_c) - F_{s_\sharp Q_m^\downarrow}(s_c) \right| \\
&\qquad + F_{s_\sharp Q_m^\downarrow}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c) \Bigg) ds_c,
\end{aligned}$$

where the first equality holds because of the stochastic dominance relationship, which tells us that $F_{s_\sharp Q_m^\uparrow}(t) \leq F_{s_\sharp Q}(t)$ and $F_{s_\sharp Q}(t) \leq F_{s_\sharp Q_m^\downarrow}(t)$ for all $t \in \mathbb{R}$. □

Proof of (8). The result follows from (7) by taking the maximum of $s_\sharp P$ out of the integral as in the proof of Theorem 3.2. Alternatively, we can also prove (8) directly from (6) in Theorem 3.2. For that, it suffices to show $W_1(s_\sharp P, s_\sharp Q)$ is upper bounded by the term in parentheses in (8) and the proof immediately follows from Theorem 3.2. We start by applying the triangle inequality twice to get

$$W_1(s_\sharp P, s_\sharp Q) \leq W_1(s_\sharp P, s_\sharp Q_m^\uparrow) + W_1(s_\sharp Q, s_\sharp Q_m^\uparrow)$$
$$W_1(s_\sharp P, s_\sharp Q) \leq W_1(s_\sharp P, s_\sharp Q_m^\downarrow) + W_1(s_\sharp Q, s_\sharp Q_m^\downarrow).$$

Since all terms are non-negative, we can sum both inequalities, which gives us

$$2W_1(s_\sharp P, s_\sharp Q) \leq W_1(s_\sharp P, s_\sharp Q_m^\uparrow) + W_1(s_\sharp P, s_\sharp Q_m^\downarrow) + W_1(s_\sharp Q, s_\sharp Q_m^\uparrow) + W_1(s_\sharp Q, s_\sharp Q_m^\downarrow). \tag{20}$$

 https://doi.org/10.52202/085713-5278

Using $s_\sharp Q_m^\uparrow \succcurlyeq s_\sharp Q \succcurlyeq s_\sharp Q_m^\downarrow$, we have from Lemma A.4 that

$$W_1(s_\sharp Q, s_\sharp Q_m^\uparrow) = \mathbb{E}_{s_\sharp Q_m^\uparrow}[S] - \mathbb{E}_{s_\sharp Q}[S] \quad \text{and} \quad W_1(s_\sharp Q, s_\sharp Q_m^\downarrow) = \mathbb{E}_{s_\sharp Q}[S] - \mathbb{E}_{s_\sharp Q_m^\downarrow}[S],$$

and by summing both equalities, the unknown expectations $\mathbb{E}_{s_\sharp Q}[S]$ cancel out, giving us

$$W_1(s_\sharp Q, s_\sharp Q_m^\uparrow) + W_1(s_\sharp Q, s_\sharp Q_m^\downarrow) = \mathbb{E}_{s_\sharp Q_m^\uparrow}[S] - \mathbb{E}_{s_\sharp Q_m^\downarrow}[S]. \tag{21}$$

Finally, it suffices to plug eq. 21 into eq. 20 and the proof follows directly from Theorem 3.2. □

Remark A.5. It is interesting to note that, if $s_\sharp Q_m^\uparrow \succcurlyeq s_\sharp P \succcurlyeq s_\sharp \hat{Q}_m^\downarrow$, the upper bound simplifies to

$$\Delta_{P,Q} \leq \int p_{s_\sharp P}(s_c)\Big(F_{s_\sharp Q_m^\downarrow}(s_c) - F_{s_\sharp Q_m^\uparrow}(s_c)\Big)ds_c.$$

A.3 Estimating the Upper Bounds to the Total Coverage Gap from Samples

In practice, we often do not have direct access to the distributions themselves and have to rely only on samples. Next, we show how to estimate our bounds from samples by leveraging the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [9] to get finite-sample guarantees. We now proceed to show how the upper bounds from Theorem 3.2 can be estimated from samples.

Theorem A.6. *Let P and Q be two probability measures on $\mathcal{X} \times \mathcal{Y}$ with $s_\sharp P$ and $s_\sharp Q$ their respective pushforward measures by the score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Let $s_\sharp \hat{P}_n$ and $s_\sharp \hat{Q}$ denote their empirical distributions constructed from n and m samples, respectively. Then, we have with probability at least $1 - 2d$ that*

$$\begin{aligned}\Delta_{P,Q} &\leq \int p_{s_\sharp P}(s_c)\left|F_{s_\sharp \hat{P}_n}(s_c) - F_{s_\sharp \hat{Q}}(s_c)\right| ds_c + \sqrt{\frac{\log(2/d)}{2n}} + \sqrt{\frac{\log(2/d)}{2m}} \\ &\leq \Big(\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c)\Big) W_1(s_\sharp \hat{P}_n, s_\sharp \hat{Q}) + \sqrt{\frac{\log(2/d)}{2n}} + \sqrt{\frac{\log(2/d)}{2m}}.\end{aligned}$$

Proof. We start by applying the triangle inequality twice to get

$$\begin{aligned}\Delta_{P,Q} &\leq \int p_{s_\sharp P}(s_c)\left|F_{s_\sharp P}(s_c) - F_{s_\sharp \hat{Q}}(s_c)\right| ds_c + \int p_{s_\sharp P}(s_c)\left|F_{s_\sharp Q}(s_c) - F_{s_\sharp \hat{Q}}(s_c)\right| ds_c \\ &\leq \int p_{s_\sharp P}(s_c)\left|F_{s_\sharp \hat{P}_n}(s_c) - F_{s_\sharp \hat{Q}}(s_c)\right| ds_c \\ &\quad + \int p_{s_\sharp P}(s_c)\left|F_{s_\sharp P}(s_c) - F_{s_\sharp \hat{P}_n}(s_c)\right| ds_c + \int p_{s_\sharp P}(s_c)\left|F_{s_\sharp Q}(s_c) - F_{s_\sharp \hat{Q}}(s_c)\right| ds_c\end{aligned}$$

From here, we get the weighted CDF version of the bound by applying the DKW inequality to the last two terms. This gives us that, with probability at least $1 - 2d$

$$\begin{aligned}\Delta_{P,Q} &\leq \int p_{s_\sharp P}(s_c)\left|F_{s_\sharp \hat{P}_n}(s_c) - F_{s_\sharp \hat{Q}}(s_c)\right| ds_c \\ &\quad + \int p_{s_\sharp P}(s_c)\sqrt{\frac{\log(2/d)}{2n}} ds_c + \int p_{s_\sharp P}(s_c)\sqrt{\frac{\log(2/d)}{2m}} ds_c \\ &= \int p_{s_\sharp P}(s_c)\left|F_{s_\sharp \hat{P}_n}(s_c) - F_{s_\sharp \hat{Q}}(s_c)\right| ds_c + \sqrt{\frac{\log(2/d)}{2n}} + \sqrt{\frac{\log(2/d)}{2m}}.\end{aligned}$$

The last equality holds since the DKW correction can be pulled outside of the integral and the integrals then sum to one due to $p_{s_\sharp P}(s_c)$ being a probability density. Finally, once more we can use the fact that $p_{s_\sharp P}(s_c)$ is non-negative everywhere to get

$$\int p_{s_\sharp P}(s_c)\left|F_{s_\sharp \hat{P}_n}(s_c) - F_{s_\sharp \hat{Q}}(s_c)\right| ds_c \leq \Big(\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c)\Big) W_1(s_\sharp \hat{P}_n, s_\sharp \hat{Q}),$$

which gives us the bound expressed in terms of the 1-Wasserstein distance. □

Remark A.7 (Extension to unlabeled bound). The upper bounds in Theorem 3.3 can also be estimated from samples in a similar manner. In fact, to get the unlabeled version of Theorem A.6, it suffices to construct two auxiliary empirical distributions such that $s_\sharp \hat{Q}_m^{\uparrow} \succcurlyeq s_\sharp \hat{Q} \succcurlyeq s_\sharp \hat{Q}_m^{\downarrow}$ and follow the same arguments used to derive Theorem 3.3 from Theorem 3.2.

Remark A.8. The DKW inequality only applies to i.i.d. samples. Therefore, to compute the bound after having optimized the weights in $s_\sharp \hat{P}_n^{\boldsymbol{w}}$, we first resample $n_{\boldsymbol{w}}$ samples from this weighted distribution, where $n_{\boldsymbol{w}} = 1/\sum w_i^2$ is the effective sample size of $s_\sharp \hat{P}_n^{\boldsymbol{w}}$. We then evaluate the bound in Theorem A.6 using these new $n_{\boldsymbol{w}}$ samples and replacing n with $n_{\boldsymbol{w}}$.

A.4 Restricted Total Coverage Gap

Thus far, we have discussed the total coverage gap, $\Delta_{P,Q}$, which considers miscoverage rates over the full range $[0, 1]$ and underpins the main results of this paper, as well as the coverage gap for specific miscoverage rates, $\Delta_{P,Q}(\alpha)$, introduced in Appendix A.5.

In some scenarios, interest may lie in a restricted range of miscoverage rates rather than the entire interval $[0, 1]$. To accommodate this, the definition can be extended to a range $[\alpha^-, \alpha^+]$ with $0 \leq \alpha^- \leq \alpha^+ \leq 1$ such that

$$\Delta_{P,Q}(\alpha^-, \alpha^+) := \int_{\alpha^-}^{\alpha^+} \frac{\Delta_{P,Q}(\alpha)}{\alpha^+ - \alpha^-} d\alpha \tag{22}$$

For the restricted coverage above, we have the following result.

Proposition A.9 (Restricted total coverage gap)**.** Let P and Q be probability measures on $\mathcal{X} \times \mathcal{Y}$, and let $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable score function with pushforward measures $s_\# P$ and $s_\# Q$. Assume $s_\# P$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ with density $p_{s_\# P}$ and CDF $F_{s_\# P}$. For $0 \leq \alpha^- \leq \alpha^+ \leq 1$, define

$$\Delta_{P,Q}(\alpha^-, \alpha^+) \;:=\; \frac{1}{\alpha^+ - \alpha^-} \int_{\alpha^-}^{\alpha^+} \Delta_{P,Q}(\alpha)\, d\alpha,$$

Then $\Delta_{P,Q}(\alpha^-, \alpha^+)$ is upper bounded by

$$\frac{1}{\alpha^+ - \alpha^-} \int_{\mathbb{R}} p_{s_\# P}(s_c) \left| F_{s_\# P}(s_c) - F_{s_\# Q}(s_c) \right| \mathbf{1}\left\{ s_c \in \left[F_{s_\# P}^{-1}(\alpha^-),\, F_{s_\# P}^{-1}(\alpha^+) \right] \right\} ds_c. \tag{23}$$

Proof. The proof is close to that of Theorem 3.2, following similar steps. By definition and the Jensen's inequality, we know that

$$\begin{aligned}
\Delta_{P,Q}(\alpha) &= \left| \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n} \left[F_{s_\sharp P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha(\mathcal{S}_c)) \right] \right| \\
&\leq \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n} \left[\left| F_{s_\sharp P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha(\mathcal{S}_c)) \right| \right]
\end{aligned}$$

Averaging over $\alpha \in [\alpha^-, \alpha^+]$ and applying Fubini to swap the order of integration,

$$\Delta_{P,Q}(\alpha^-, \alpha^+) \leq \mathbb{E}_{\mathcal{S}_c \sim s_\# P} \left[\frac{1}{\alpha^+ - \alpha^-} \int_{\alpha^-}^{\alpha^+} \left| F_{s_\# P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\# Q}(\mathbb{Q}_\alpha(\mathcal{S}_c)) \right| d\alpha \right]$$

Now fix $\mathcal{S}_c = \{s_c^{(1)}, \ldots, s_c^{(n)}\}$ sorted increasingly. Over the full range $[0, 1]$, each calibration score occupies an interval of length $1/n$ in the quantile map. Restricting to $[\alpha^-, \alpha^+]$ simply zeroes out scores whose CDF lies outside this range. Thus

$$\begin{aligned}
&\int_{\alpha^-}^{\alpha^+} \left| F_{s_\# P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\# Q}(\mathbb{Q}_\alpha(\mathcal{S}_c)) \right| d\alpha = \\
&\qquad \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}\left(F_{s_\# P}(s_c^{(i)}) \in [\alpha^-, \alpha^+] \right) \left| F_{s_\# P}(s_c^{(i)}) - F_{s_\# Q}(s_c^{(i)}) \right|.
\end{aligned}$$

https://doi.org/10.52202/085713-5278

Note that this is the same argument of (15) in the proof of Theorem 3.2, but here we take extra care to restrict the range to $[\alpha^-, \alpha^+]$. Plugging back we have

$$\Delta_{P,Q}(\alpha^-, \alpha^+) \leq \mathbb{E}_{\mathcal{S}_c \sim s_\# P}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left(F_{s_\# P}(s_c^{(j)}) \in [\alpha^-, \alpha^+]\right) \left|F_{s_\# P}(s_c^{(j)}) - F_{s_\# Q}(s_c^{(j)})\right|\right].$$

Since the calibration data point is identically distributed, the expectation of the sample mean equals the population mean:

$$\mathbb{E}_{\mathcal{S}_c \sim s_\# P}\left[\frac{1}{n}\sum_{j=1}^{n} h(s_c^{(j)})\right] = \mathbb{E}_{\mathcal{S}_c \sim s_\# P}[h(S)],$$

with $h(s) = \mathbf{1}\left(F_{s_\# P}(s_c^{(i)}) \in [\alpha^-, \alpha^+]\right) \left|F_{s_\# P}(s_c^{(i)}) - F_{s_\# Q}(s_c^{(i)})\right|$. Therefore,

$$\Delta_{P,Q}(\alpha^-, \alpha^+) \leq \frac{1}{\alpha^+ - \alpha^-}\int_{\mathbb{R}} p_{s_\# P}(s) \left|F_{s_\# P}(s) - F_{s_\# Q}(s)\right| \mathbf{1}\left(F_{s_\# P}(s_c^{(i)}) \in [\alpha^-, \alpha^+]\right) ds,$$

which is the desired bound. □

Remark A.10 (On Wasserstein relaxations). It is also possible to connect the result of Proposition A.9 to the 1-Wasserstein distance. Since $\mathbf{1}(\cdot) \leq 1$, a loose relaxation of (23) gives

$$\begin{aligned}\Delta_{P,Q}(\alpha^-, \alpha^+) &\leq \frac{1}{\alpha^+ - \alpha^-}\Big(\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c)\Big)\int_{\mathbb{R}} \left|F_{s_\# P}(s) - F_{s_\# Q}(s)\right| ds \\ &= \frac{1}{\alpha^+ - \alpha^-}\Big(\sup_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c)\Big) W_1(s_\# P, s_\# Q),\end{aligned}$$

which may be overly conservative in practice.

While one could optimize the upper bound in Proposition A.9 directly, preliminary experiments show that this approach yields only marginal improvements in coverage. We see two likely reasons for this. First, optimizing a bound restricted to a specific coverage range may be inherently more challenging; for instance, the pointwise bound for a fixed α in Appendix A.5 also failed to deliver better empirical performance. Second, the total coverage gap already provides a strong and well-behaved objective, leaving little room for alternative formulations to offer significant gains. Nevertheless, these more targeted objectives remain an interesting direction for future work, particularly in applications where coverage guarantees over a narrow range of α are critical.

A.5 Upper Bound to the Coverage Gap for a Specific Target Miscoverage Rate

As discussed in the main paper, similar techniques can also be employed to derive an upper bound on $\Delta_{P,Q}(\alpha)$, the coverage gap corresponding to a given miscoverage rate α. This result is formalized in Theorem A.11, which also outlines how it can be estimated from samples using the DKW inequality.

Theorem A.11. *Let P and Q be two probability measures on $\mathcal{X} \times \mathcal{Y}$ with $s_\sharp P$ and $s_\sharp Q$ their respective pushforward measures by the score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Let $s_\sharp \hat{P}_n$ and $s_\sharp \hat{Q}_m$ denote their empirical distributions constructed from n and m samples, respectively. Further, let $s_\sharp \hat{Q}_m^{\downarrow}$ and $s_\sharp \hat{Q}_m^{\uparrow}$ be such that $s_\sharp \hat{Q}_m^{\uparrow} \succcurlyeq s_\sharp \hat{Q}_m \succcurlyeq s_\sharp \hat{Q}_m^{\downarrow}$. Then, we have with probability at least $1 - 2d$ that*

$$\begin{aligned}\Delta_{P,Q}(\alpha) \leq \frac{1}{2}\mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\Bigg[&\left|F_{s_\sharp \hat{P}_n}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp \hat{Q}_m^{\downarrow}}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right| \\ &+ \left|F_{s_\sharp \hat{P}_n}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp \hat{Q}_m^{\uparrow}}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right| + \left|F_{s_\sharp \hat{Q}_m^{\downarrow}}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp \hat{Q}_m^{\uparrow}}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right|\Bigg] \\ &+ \sqrt{\frac{\log(2/d)}{2n}} + \sqrt{\frac{\log(2/d)}{2m}}\end{aligned}$$

Proof. Per definition the coverage gap for a specific target miscoverage rate α is given by

$$\begin{aligned}
\Delta_{P,Q}(\alpha) &:= \left|P(S_t \leq \mathbb{Q}_\alpha(S_c)) - Q(S_t \leq \mathbb{Q}_\alpha(S_c))\right| \\
&= \Bigg|\mathbb{E}_{S_t \sim s_\sharp P}\left[\mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\left[\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha(S_c)\right)\right]\right] \\
&\qquad\qquad - \mathbb{E}_{S_t \sim s_\sharp Q}\left[\mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\left[\mathbf{1}\left(S_t \leq \mathbb{Q}_\alpha(S_c)\right)\right]\right]\Bigg|, \\
&= \left|\mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\left[F_{s_\sharp P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right]\right| \\
&\leq \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\left[\left|F_{s_\sharp P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right|\right]
\end{aligned}$$

where the last inequality follows from Jensen's inequality. At this point we introduce the empirical distributions $s_\sharp \hat{P}_n$ and $s_\sharp \hat{Q}_m$ by applying the triangle inequality twice.

$$\begin{aligned}
\Delta_{P,Q}(\alpha) &\leq \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\Bigg[\left|F_{s_\sharp \hat{P}_n}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp Q}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right| \\
&\qquad\qquad + \left|F_{s_\sharp P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp \hat{P}_n}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right|\Bigg] \\
&\leq \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\Bigg[\left|F_{s_\sharp \hat{P}_n}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp \hat{Q}_m}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right| \\
&\quad + \left|F_{s_\sharp P}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp \hat{P}_n}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right| + \left|F_{s_\sharp Q}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp \hat{Q}_m}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right|\Bigg]
\end{aligned}$$

We then apply the DKW inequality to get with probability $1 - 2d$

$$\Delta_{P,Q}(\alpha) \leq \mathbb{E}_{\mathcal{S}_c \sim s_\sharp P^n}\left[\left|F_{s_\sharp \hat{P}_n}(\mathbb{Q}_\alpha(\mathcal{S}_c)) - F_{s_\sharp \hat{Q}_m}(\mathbb{Q}_\alpha(\mathcal{S}_c))\right|\right] + \sqrt{\frac{\log(2/d)}{2n}} + \sqrt{\frac{\log(2/d)}{2m}}.$$

Finally, we introduce the two auxiliary distributions $s_\sharp Q_m^\uparrow \succcurlyeq s_\sharp Q \succcurlyeq s_\sharp Q_m^\downarrow$ by once more applying the triangle inequality twice and summing the inequalities to get the final result. □

Unfortunately, preliminary experiments indicate that the bound presented in Theorem A.11 is loose and of limited practical utility, unless the auxiliary distributions are close to the true score distribution under Q. Further research is required to derive meaningful bounds for specific values of α in the absence of labeled data. Nevertheless, if the goal is to enhance coverage for a particular α, one can directly optimize the bound in Theorem A.11. Following prior work [4, 7, 39], this can be achieved by introducing differentiable relaxations in the computation of quantiles and empirical CDFs, thereby enabling gradient-based optimization with respect to the weights in $s_\sharp \hat{P}_n^{\boldsymbol{w}}$. However, this approach has proven less effective than optimizing upper bounds on the total coverage gap.

 https://doi.org/10.52202/085713-5278

B Applicability, Limitations, and Extensions

Prior-Knowledge-Based Sandwiching Design. Our bounds in Theorem 3.3 rely on auxiliary distributions $(s_\sharp Q_\downarrow, s_\sharp Q_\uparrow)$ that stochastically dominate the unknown test score distribution $s_\sharp Q$. In the absence of prior knowledge, we use uninformed constructions such as $(\min, \max)$ or (f, U), which work well empirically but may yield loose bounds. A natural extension is to exploit domain-specific or structural prior information to design tighter sandwiching distributions.

For example, if labels are organized in a hierarchy (e.g., ImageNet superclasses), side information can constrain the feasible set of candidate labels for each test point. This allows constructing $s_\sharp Q_\downarrow$ and $s_\sharp Q_\uparrow$ by selecting the most plausible and least plausible labels within that subset, leading to provably tighter bounds. Similarly, if the distribution shift is known to be bounded (e.g., perturbations within an ℓ_∞ ball), optimization procedures can identify feasible score ranges that respect these constraints. The smaller the feasible set, the closer the auxiliary distributions approximate $s_\sharp Q$, improving both theoretical guarantees and empirical performance.

Exploring these strategies (hierarchical constraints, perturbation models, or other structured priors) represents a promising direction for future work, as it bridges the gap between general-purpose bounds and application-specific robustness.

Robustness to Misspecification. Our bounds in Theorem 3.3 assume that the auxiliary distributions $(s_\sharp Q_\downarrow, s_\sharp Q_\uparrow)$ satisfy the stochastic dominance relationship $s_\sharp Q_\uparrow \succeq s_\sharp Q \succeq s_\sharp Q_\downarrow$. This condition is guaranteed for the $(\min, \max)$ construction and was observed to hold in many experiments for (f, U), which uses scores derived from the model's predicted distribution $Q_f(Y|X)$ and from a uniform distribution over labels. However, (f, U) does not always satisfy this assumption.

Coverage improvements in these cases can be explained by the fact that the optimization objective remains effective whenever the auxiliary distributions help move the calibration score distribution closer to the test score distribution. This alignment, even if imperfect, can still reduce the coverage gap. However, this is an important caveat: if the auxiliary distributions fail to capture the nature of the shift, optimization may bias the calibration distribution in the wrong direction and worsen coverage. Handling this risk requires care.

Future work should systematically study these failure modes and develop safeguards. Promising directions include diagnostics to detect dominance violations, adaptive refinement of auxiliary distributions based on empirical checks, and regularization strategies to prevent extreme deviations when auxiliary distributions are poorly aligned with the test distribution.

Ambiguous Ground Truth. Ambiguous ground truth arises in settings where each instance may correspond to multiple plausible labels with associated probabilities, such as in fine-grained classification or scenarios with inherent uncertainty. This problem has recently attracted attention in the conformal prediction literature [5, 40].

Our method operates directly on nonconformity scores, which are typically unidimensional, without imposing any assumptions on how the scores are constructed. This property makes it naturally compatible with most CP techniques, including scenarios involving ambiguous ground truth. For example, following Stutz et al. [40], one can define a score function as a weighted average of class-specific scores under a plausibility vector $\lambda \in \Delta_K$, i.e.,

$$s'(x, \lambda) := \sum_{k=1}^{K} \lambda_k \, s(x, y_k).$$

Once such a score is defined, our approach can learn weights over the calibration data and compute a threshold that adapts under distribution shift, just as in the standard setting.

The main challenge lies in constructing auxiliary distributions for ambiguous ground truth. Instead of working with a discrete set of labels, we must consider distributions over the simplex, which complicates the design of $(s_\sharp Q_\downarrow, s_\sharp Q_\uparrow)$. While a min–max construction remains possible, it may be overly conservative, as the resulting auxiliary distributions could be too far apart to yield tight bounds. Future work could explore more informative strategies for building auxiliary distributions in this setting, potentially leveraging prior knowledge or structural constraints on the plausibility vectors.

C Extra experimental details and results

In this section, we present additional details about our experimental setup and supplementary results. We begin by outlining how our methods fit within the split conformal prediction framework in Section C.1 and specifically in Algorithm 2. Next, we detail the baseline methods in Section C.3, followed by a description of the datasets used in Section C.4. Finally, in Section C.5, we discuss key design choices and ablation studies that may offer valuable insights for future research.

Before proceeding, we comment on a few technical details. We note that the code was implemented in Python 3 using PyTorch [30] and all experiments were conducted on a single commercial NVIDIA GPU with 12 GB of memory.

Algorithm 2 End-to-end Non-exchangeable CP with Optimal Transport (Split CP)

Input:
  n labeled samples $\{(x_i, y_i)\}_{i=1}^n$ from P
  m unlabeled samples $\{x_j\}_{j=n+1}^{n+m}$ from Q
  target miscoverage $\alpha \in (0, 1)$
  score function s
Initialize unnormalized weights $\tilde{\boldsymbol{w}} = \{\tilde{w}_i\}_{i=1}^n$ or weight function w_θ
Compute scores $\{s(x_i, y_i)\}_{i=1}^n$
Compute score vectors $\{\boldsymbol{s}(x_j)\}_{j=n+1}^{n+m}$ // $\boldsymbol{s}(x) = \{s(x, y) : y \in \mathcal{Y}\}$
repeat
  Construct $s_\sharp \hat{Q}^\downarrow$ and $s_\sharp \hat{Q}^\uparrow$ from $\{\boldsymbol{s}(x_j)\}_{j=n+1}^{n+m}$ // e.g., $(\min, \max)$ or (f, U)
  Fit KDE to $\{s(x_i, y_i)\}_{i=1}^n$ with weights $\boldsymbol{w}$
  Update $\tilde{\boldsymbol{w}}$ or w_θ to minimize either (7) or (8) // weighted-CDF or 1-Wasserstein bound
until convergence or max steps
Weighted normalization: Compute normalized weights $\boldsymbol{w}$ // accounting for weight of test point
Weighted threshold: $q_{1-\alpha} \leftarrow Q_\alpha^{\boldsymbol{w}}\big(\{s(x_i, y_i)\}_{i=1}^n\big)$
Prediction sets: for each x_t, set $C(x_t) \leftarrow \{\, y \in \mathcal{Y} \;:\; s(x_t, y) \leq q_{1-\alpha} \,\}$
Output: $\{C(x_t)\}_{t=1}^T$ (if $\{x_t\}$ provided) and learned weights $\boldsymbol{w}$

C.1 Split Conformal Prediction Procedure

We follow a standard split conformal prediction framework, with the main difference being that we introduce weights over the calibration nonconformity scores to better align their empirical distribution with that of test scores. Concretely, we construct prediction sets by thresholding on the nonconformity scores given by one minus the model-assigned probabilities. However, our methods are agnostic to the choice of score function and could be applied to other approaches like APS [35]. We outline the complete split CP algorithm we use, including weight optimization with our bounds, in Algorithm 2.

C.1.1 Weighting of Test Samples

In standard split conformal prediction, a test data point is implicitly assigned a weight of $1/_{n+1}$, preserving symmetry with the calibration set. Extensions to non-exchangeable settings like [2, 42] also address this issue explicitly. Tibshirani et al. [42] propose assigning the test point a weight proportional to its likelihood ratio under the shifted distribution (see details in Section C.3.2), while Barber et al. [2] fix the unnormalized weight of the test point to 1, reflecting the fact that it already comes from the target distribution.

Free-form weights We adopt a similar convention of [2] for free-form weights, assigning test points the unnormalized weight $\tilde{w}_{n+1} = 1$. Intuitively, importance weights are meant to correct for distribution mismatch, and the test point is already drawn from the target distribution, which justifies unit weights. In that case, we have the following weighted empirical distribution $s_\sharp P_n^{\boldsymbol{w}}$

$$s_\sharp \hat{P}_n^{\boldsymbol{w}} = \sum_{i=1}^n w_i \delta_{s(x_i, y_i)} \quad \text{with normalized weights} \quad w_i = \frac{\tilde{w}_i}{1 + \sum_{j=1}^n \tilde{w}_j}.$$

https://doi.org/10.52202/085713-5278

Weight Function When using a weight function, we adopt a similar strategy to that of Tibshirani et al. [42]. Unlike their setting, where the weight function maps inputs x to weights, our function maps nonconformity scores to weights. Consequently, for each candidate label $y \in \mathcal{Y}$, we obtain a distinct test weight $w_\theta(s(x_{n+1}, y))$. Computing all these weights can be expensive when $|\mathcal{Y}|$ is large. To mitigate this cost, we approximate the test weight by taking the maximum over the score range $[s_{\min}, s_{\max}]$ to get a conservative upper bound for the conformal threshold:

$$\tilde{w}_{n+1} = \max_{s \in [s_{\min}, s_{\max}]} w_\theta(s).$$

In our classification experiments, we define nonconformity scores as one minus the probabilities assigned by the model, with $[s_{\min}, s_{\max}] = [0, 1]$. For calibration points, we compute unnormalized weights directly as $\tilde{w}_i = w_\theta(s(x_i, y_i))$ for $i \in \{0, \ldots, n\}$. Finally, the weighted empirical distribution $s_\sharp P_n^{\boldsymbol{w}}$ is given by

$$s_\sharp \hat{P}_n^{\boldsymbol{w}} = \sum_{i=1}^{n} w_i \delta_{s(x_i, y_i)} \quad \text{with normalized weights} \quad w_i = \frac{\tilde{w}_i}{\sum_{j=1}^{n+1} \tilde{w}_j}.$$

C.2 Other Implementation Details

In all experiments, free-form weights are randomly initialized from a uniform distribution in $[0, 1]$ but mapped to log space for stability. When learning a weight function, we implement it as a small multi-layer perceptron (MLP) applied directly to scalars representing nonconformity scores. In all cases, the architecture is a simple MLP with shape $1 \to 256 \to 16 \to 8 \to 1$ and ReLU activations followed a tempered $\tanh$ output to bound log-weights in $[-20, 20]$. Optimization follows the exact same procedure for both parametrizations. In particular, for the image classification tasks, we use Adam with learning rate 10^{-3} for all datasets, varying the number of steps from 1000 to 5000 steps depending on shift severity; see Section C.4 for exact details.

C.3 Baselines

C.3.1 "Oracle"

We use the term "oracle" to describe the marginal coverage and expected prediction set size achieved when both calibration and testing are performed on samples from Q. This setup guarantees the desired coverage and, in our context where the test scores remain fixed, represents the best possible outcome. We only show these results as a reference for what we would get if we knew the true $s_\sharp Q$, highlighting that the prediction set sizes are considerably larger only because the model is less accurate, as it was trained on samples from P and not Q.

C.3.2 Likelihood Ratios

Tibshirani et al. [42] also proposed to reweight calibration points from P and apply split CP using a weighted distribution of calibration scores $s_\sharp \hat{P}_n^{\boldsymbol{w}} = \frac{1}{n} \sum_{i=1}^{n} w_i \delta_{s(x_i, y_i)}$. Since they only address covariate shifts, it is easy to show the optimal unnormalized weights are given by likelihood ratios of the form $dQ(x_i)/dP(x_i)$. Unfortunately, learning accurate likelihood ratios is known to be challenging, especially when the two distributions are far apart. In our experiments, we applied the telescopic density ratio estimation approach of [34], which we found useful in the context of more severe shifts. In all cases, learning likelihood ratios directly on the input space $\mathcal{X}$ proved challenging and we found more success when operating on the space of scores, i.e., fitting the density ratio estimator to map vector of scores $\boldsymbol{s}(x_i)$ to (approximate) ratios $\tilde{w}_i \approx dQ(x_i)/dP(x_i)$. The weights w_i are then recovered by normalizing the likelihood ratios over the calibration set and the test point in question. That is, at test time, we must first evaluate the density ratio estimator to get $\tilde{w}_{n+1} \approx dQ(x_{n+1})/dP(x_{n+1})$ and then compute normalized weights as

$$w_i = \frac{\tilde{w}_i}{\sum_{j=1}^{n+1} \tilde{w}_j}.$$

For the large datasets, ImageNet-C and iWildcam, our density ratio estimator was given by a neural network with two hidden layers and ten bridges. In this context, each bridge predicts the density

ratio between two intermediary distributions defined by a mixture of samples from P and Q. In our experiments, we constructed the intermediary distributions via linear combinations as detailed in [34]. We train the density ratio estimator for ten epochs with a learning rate of $1e^{-3}$ and weight decay of $1e^{-3}$ to avoid overfitting. In the regression task, we also used an MLP with two hidden layers and the same learning rate of $1e^{-3}$ but forewent the telescopic approach (no bridges) as it did not prove useful. Regarding the architectures, we used ReLU activations and kept the hidden size constant and equal to the input size, i.e., 1000 for ImageNet-C, 182 for iWildCam, and 4 for the regression task.

C.3.3 Entropy scaled Conformal Prediction

We also compare our methods to Entropy scaled Conformal Prediction (ECP) proposed by [21], which constructs prediction sets as

$$u_{\mathcal{D}_Q^{(2)}} = \mathbb{Q}_\alpha\left(\{h(\boldsymbol{f}(X_i))\}_{i=1}^n\right)$$
$$\mathcal{C}_{Kasa}(X_t) = \left\{y \in \mathcal{Y} : s(X_t, y) \cdot \max(1, u_{\mathcal{D}_Q^{(2)}}) \leq \mathbb{Q}_\alpha\left(\mathcal{S}_c\right)\right\},$$

where $\boldsymbol{f}(X_i) = \{f(X_i, y') : y' \in \mathcal{Y}\}$ is the set of probabilities assigned to each class by the underlying predictor f, and $h(\boldsymbol{f}(X_i)) = -\sum_{y\in\mathcal{Y}} f(X_i, y)\log(f(X_i, y))$ is the entropy of this set of probabilities. Their method is designed to work with test-time adaption methods [26, 41], which adapt the model f in a stream of test samples. However, it can also be applied to our setting where the model is kept fixed and a set of test samples $\mathcal{D}_Q^{(2)}$ is observed all at once, as formulated above.

C.3.4 Conformal Prediction With Conditional Guarantees

Gibbs et al. [15] proposed a method that, under a prespecified function class of covariate shifts, guarantees conditional coverage, i.e., ensuring the prediction set contains the true label for every test point X_t

$$\mathbb{P}(Y_t \in \mathcal{C}(X_t)|X_t) \geq 1 - \alpha.$$

Their approach essentially changes the conformal threshold for each new test point by learning a function $\hat{g}_{s(X_t,y)}$ as follows

$$\hat{g}_S = \arg\min_{g\in\mathcal{F}} \frac{1}{n+1}\sum_{i=1}^n \ell_\alpha(g(X_i), S_i) + \frac{1}{n+1}\ell_\alpha(g(X_t), S)$$
$$\mathcal{C}_{Gibbs}(X_t) = \left\{y \in \mathcal{Y} : s(X_t, y) \leq \hat{g}_{s(X_t,y)}(X_t)\right\}, \tag{24}$$

where $\mathcal{F}$ is the function class of distribution shifts of interest, ℓ_α is the pinball loss with target quantile level α, and $S \in \mathbb{R}$ refers to the unknown nonconformity score of X_t. By computing a new threshold per test point and providing conditional guarantees, their method is inherently more powerful than ours. However, this extra power comes with extra limitations:

- **Function class $\mathcal{F}$ is typically unknown**. Conditional guarantees are impossible in the most general case of an arbitrary infinite dimensional class [11, 44]. Therefore, we must constrain ourselves to a prespecified class of functions, which requires precise knowledge about the types of distribution shift we expect in practice. In experiments with large datasets like ImageNet-C or iWildCam, it is not clear how to define $\mathcal{F}$ effectively.
- **Computational cost at test time**. Gibbs et al. [15] propose an efficient algorithm to optimize (24) that leverages the monotonicity of quantile regression to avoid evaluating each possible test score for X_t. Yet, this procedure still significantly increases latency at test time, and in our hardware, it took 1.5 seconds per test point when using 300 calibration points, and 30 seconds when using 1000 calibration points. Due to this extra computational cost, we only used 300 calibration points when applying the method of [15].
- **Prediction set size**. In practice, we observed that prediction sets produced by the method of [15] to be significantly larger than optimal. This is in part due to the stronger conditional guarantee but might reduce the usefulness of the prediction sets in practice.

In our experiments, we used the official implementation available at github.com/jjcherian/conditional-conformal. Similarly to the likelihood ratio baseline described above, we applied their method

 https://doi.org/10.52202/085713-5278

to the vector of scores instead of the input space. Since the function class corresponding to the type of distribution shifts observed in ImageNet-C and iWildCam are hard to define in practice, we applied the most general approach using radial basis function (RBF) kernels with hyperparameters $\gamma = 12.5$ and $\lambda = 0.005$ (see the official implementation for details) for ImageNet-C and iWildCam experiments, since we found this to work best in preliminary experiments.

C.4 Datasets

C.4.1 Regression - Synthetic Data

For the toy regression task we adopt a setting similar to the one proposed in [49], where we have a regression problem with 4-dimensional input variable $X \sim \mathcal{N}(0, I_4)$, and target variable given by $Y = 210 + 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4 + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 1)$. We will refer to this distribution on $\mathcal{X} \times \mathcal{Y}$ as the unshifted distribution P. We then induce a covariate shift to get a new distribution Q via exponential tilting by resampling the data with weights $w_{\text{tilt}}(x) = \exp(-1x_1 + 0.5x_2 - 0.25x_3 - 0.1x_4)$. This automatically gives us ground-truth likelihood ratios, since $w_{\text{tilt}}(x) = {}^{dQ(x)}/_{dP(x)}$ by design. Note that these tilting weights should still be normalized over the calibration set to define a proper empirical distribution we can apply conformal prediction over.

For the experiment depicted in Figure 3, we repeat 500 simulations, each time sampling new datasets $\mathcal{D}_P^{(1)}$, $\mathcal{D}_P^{(2)}$, $\mathcal{D}_Q^{(1)}$, and $\mathcal{D}_Q^{(2)}$, each with 1000 samples. We use the regression-as-classification method of Guha et al. [16], splitting the output space into 50 equally spaced bins. For the predictor, we train a multilayer perceptron (MLP) with a single hidden layer of 256 units and ReLU activations. For the density ratio estimator, we use an MLP with one hidden layer with 32 neurons, which we train to distinguish samples from $\mathcal{D}_P^{(2)}$ and $\mathcal{D}_Q^{(1)}$ by minimizing the logistic loss as common in density estimation tasks. For our methods, we only considered the free-form parametrization for this experiment. We repeat the optimization process in Algorithm 1 for 10 steps, each time updating all the weights using Adam [22] with a learning rate of 0.1.

C.4.2 ImageNet-C

In all cases, the underlying classifier is the pretrained ResNet-50 from torchvision [27]. To simulate label shift, as in previous work [12], we do so by resampling data points (without replacement) according to a new label marginal $Q(Y) \sim \text{Dir}(\boldsymbol{c})$ where the concentration parameters $\boldsymbol{c}$ are given by $c_k = P(Y = k) * \gamma$. The parameter γ controls the intensity of the label shift, with lower values of γ producing more skewed label marginals. However, since we sample without replacement, low values of γ lead to small sample sizes (< 50) and, consequently, noisy results. For that reason, we set $\gamma = 10$ which yields a significant label shift while producing sample sizes of around 300 samples.

ImageNet-C [17] comprises 15 different types of corruption applied on top of the original validation dataset of ImageNet [36], which contains 50K samples. We repeat each experiment with 10 random seeds. As explained in Section 6.1, each time we randomly split the corrupted data $\mathcal{D}_Q$ into two: $\mathcal{D}_Q^{(1)}$ used to learn the weights or density ratio model, and $\mathcal{D}_Q^{(2)}$ for testing. Since we use a pretrained model as classifier, we do not need to reserve a subset of the clean data $\mathcal{D}_P$ (the original validation dataset) for training the model. However, one thing to note is that in ImageNet-C $\mathcal{D}_Q$ is constructed using the same images in $\mathcal{D}_P$. Thus, we have to ensure the calibration dataset, which we denote $\mathcal{D}_P^{(2)}$ to be consistent with Section 6.1, does not contain any of the images in either $\mathcal{D}_Q^{(1)}$ or $\mathcal{D}_Q^{(2)}$. This will affect the sizes of each of these sets, as described in the following sections. In all ImageNet-C experiments, unless explicitly stated otherwise, we construct calibration and test sets with the following sizes: calibration sets with $|\mathcal{D}_P^{(2)}| = 300$ and $|\mathcal{D}_Q^{(1)}| = 300$, and test sets with $|\mathcal{D}_Q^{(2)}| = 30000$.

The underlying classifier was a pretrained ResNet-50 available in Torchvision package [27], which was kept fixed in all experiments. We only learn weighting scheme for the calibration data points, and the model as well as the nonconformity score function remain unchanged in all cases. In each optimization step and for both parametrizations, we backprop through all weights (no batching) using Adam with a learning rate of $1\text{e}{-3}$ and $\beta = (0.9, 0.999)$. These hyperparameters were the same across all runs, with only the number of optimization steps varying: 1000 steps for distribution shifts of severity 1, 3000 steps for severity 3, and 5000 steps for severity 5. We observed our methods to be

fairly robust to these hyperparameters. One should only keep in mind that the more severe the shift, the longer the optimization or the larger the learning rate should be, as demonstrated in our approach.

In Table 6, we present the empirical coverage obtained for each method and corruption in ImageNet-C, giving a more complete picture of the ImageNet-C results reported in Table 1.

C.4.3 iWildCam

iWildCam involves images of animals from different camera traps that aim to monitor biodiversity loss. The distribution shift arises from the differences in the characteristics of the environment of each camera trap (e.g., changes in illumination, camera angle, background, etc.). We use different subsets of camera traps for training, validation and testing, which induces a distribution shift. For the classifier, we train a ResNet-50 model from scratch on the training set. As for ImageNet-C, we repeat each experiment with 10 random seeds.

We experiment with two different settings of distribution shift: the natural one (i.e., differences in camera-traps between validation and testing) that already exists in the data, and a combination of the natural shift with a label shift induced by changing the marginal over the labels via a Dirichlet distribution in the same way as for the ImageNet-C dataset, also with $\gamma = 10$. For this experiment, we also used Adam with a learning rate of $1\mathrm{e}{-3}$ and $\beta = (0.9, 0.999)$, running optimization for 1000 steps. As for the ImageNet-C dataset, we use calibration sets with $|\mathcal{D}_P^{(2)}| = 300$ and $|\mathcal{D}_Q^{(1)}| = 300$ for the experiments reported in the main paper, with the test set composed of 10000 samples, i.e., $|\mathcal{D}_Q^{(2)}| = 10000$. However, we also use the iWildCam dataset to study the effect of variations in sample sizes in our methods, as explained in the next section.

C.5 Further Discussion

In this section, we examine key design choices, including the impact of sample size and the performance differences arising from the selection of bound and auxiliary distributions. For clarity, the empirical results presented in this section focus exclusively on the free-form parametrization, i.e., we directly optimize the weights over the calibration scores.

C.5.1 Influence of the number of samples from calibration and test distributions

In Tables 2 and 3, we illustrate how the final total coverage gap $\Delta_{P,Q}$ varies with $|\mathcal{D}_P^{(2)}|$, the number of calibration samples from P, and $|\mathcal{D}_Q^{(1)}|$, the number of unlabeled samples from Q. As anticipated, the method's performance improves with an increase in the number of available samples. Interestingly, the number of calibration samples from P appears to be more crucial for performance. This is encouraging, as it suggests that collecting or waiting for a large number of unlabeled samples from the test distribution is unnecessary, with no significant gains observed beyond 1000 samples.

We observed a significant reduction in the total coverage gap in all cases where the number of calibration samples from P was 100 or more. For smaller calibration samples, the observed change in coverage was minimal or even slightly detrimental, as in the case of $(s_\sharp \hat{Q}_m^{\min},\ s_\sharp \hat{Q}_m^{\max})$ with $|\mathcal{D}_P^{(2)}| = 30$ and $|\mathcal{D}_Q^{(1)}| = 100$. This is likely because we do not have enough samples from P to represent $s_\sharp Q$ well enough via a weighted empirical distribution of the form $s_\sharp \hat{P}_n^{\boldsymbol{w}}$. More broadly, this underscores one of the limitations of our approach. The possibility of our method hurting coverage in some cases is not surprising, since we tackle the most general distribution shift case, with no prior information about the shift mechanics. In that setting, there is always a risk that optimizing our methods could negatively impact coverage. However, the results show a positive trend, indicating that these risks tend to diminish as the number of available samples increases.

C.5.2 Choice of Bound

Before applying our methods, two key decisions must be made. The first is whether to use the weighted CDF formulation in (7) or the 1-Wasserstein distance formulation in (8). The second involves selecting the appropriate pair of auxiliary distributions. To guide these choices, we evaluate the total coverage gap achieved after optimization under each configuration. The results for the image classification datasets are summarized in Table 4.

 https://doi.org/10.52202/085713-5278

Table 2: Total coverage gap on iWildCam with ResNet-50 for varying number of calibration samples from P ($|\mathcal{D}_P^{(2)}|$) and Q ($|\mathcal{D}_Q^{(1)}|$) for our method with $(s_\sharp Q^{\min}, s_\sharp Q^{\max})$. We highlight in blue, the cases where total coverage improved by more than one standard deviation. We report mean and standard deviation across 10 random seeds. Lower is better.

		Number of unlabeled samples from Q				
# samples from P	Uncorrected	$\lvert\mathcal{D}_Q^{(1)}\rvert = 30$	$\lvert\mathcal{D}_Q^{(1)}\rvert = 100$	$\lvert\mathcal{D}_Q^{(1)}\rvert = 300$	$\lvert\mathcal{D}_Q^{(1)}\rvert = 1000$	$\lvert\mathcal{D}_Q^{(1)}\rvert = 3000$
$\lvert\mathcal{D}_P^{(2)}\rvert = 30$	$0.119_{\pm 0.046}$	$0.117_{\pm 0.048}$	$0.122_{\pm 0.044}$	$0.115_{\pm 0.035}$	$0.118_{\pm 0.039}$	$0.119_{\pm 0.045}$
$\lvert\mathcal{D}_P^{(2)}\rvert = 100$	$0.143_{\pm 0.022}$	$0.107_{\pm 0.019}$	$0.102_{\pm 0.017}$	$0.099_{\pm 0.015}$	$0.096_{\pm 0.016}$	$0.101_{\pm 0.018}$
$\lvert\mathcal{D}_P^{(2)}\rvert = 300$	$0.145_{\pm 0.014}$	$0.091_{\pm 0.017}$	$0.085_{\pm 0.009}$	$0.084_{\pm 0.010}$	$0.083_{\pm 0.010}$	$0.085_{\pm 0.010}$
$\lvert\mathcal{D}_P^{(2)}\rvert = 1000$	$0.139_{\pm 0.010}$	$0.085_{\pm 0.015}$	$0.078_{\pm 0.006}$	$0.072_{\pm 0.005}$	$0.072_{\pm 0.004}$	$0.073_{\pm 0.006}$
$\lvert\mathcal{D}_P^{(2)}\rvert = 3000$	$0.139_{\pm 0.005}$	$0.082_{\pm 0.015}$	$0.075_{\pm 0.008}$	$0.070_{\pm 0.007}$	$0.069_{\pm 0.005}$	$0.070_{\pm 0.006}$

Table 3: Total coverage gap on iWildCam with ResNet-50 for varying number of calibration samples from P ($|\mathcal{D}_P^{(2)}|$) and Q ($|\mathcal{D}_Q^{(1)}|$) for our method with $(s_\sharp Q^f, s_\sharp Q^U)$. We report mean and standard deviation across 10 random seeds. Lower is better.

		Number of unlabeled samples from Q				
# samples from P	Uncorrected	$\lvert\mathcal{D}_Q^{(1)}\rvert = 30$	$\lvert\mathcal{D}_Q^{(1)}\rvert = 100$	$\lvert\mathcal{D}_Q^{(1)}\rvert = 300$	$\lvert\mathcal{D}_Q^{(1)}\rvert = 1000$	$\lvert\mathcal{D}_Q^{(1)}\rvert = 3000$
$\lvert\mathcal{D}_P^{(2)}\rvert = 30$	$0.119_{\pm 0.046}$	$0.112_{\pm 0.044}$	$0.117_{\pm 0.042}$	$0.109_{\pm 0.033}$	$0.112_{\pm 0.037}$	$0.113_{\pm 0.044}$
$\lvert\mathcal{D}_P^{(2)}\rvert = 100$	$0.143_{\pm 0.022}$	$0.097_{\pm 0.021}$	$0.094_{\pm 0.018}$	$0.090_{\pm 0.014}$	$0.086_{\pm 0.014}$	$0.091_{\pm 0.018}$
$\lvert\mathcal{D}_P^{(2)}\rvert = 300$	$0.145_{\pm 0.014}$	$0.083_{\pm 0.018}$	$0.076_{\pm 0.010}$	$0.073_{\pm 0.009}$	$0.073_{\pm 0.010}$	$0.073_{\pm 0.011}$
$\lvert\mathcal{D}_P^{(2)}\rvert = 1000$	$0.139_{\pm 0.010}$	$0.072_{\pm 0.018}$	$0.067_{\pm 0.008}$	$0.061_{\pm 0.006}$	$0.059_{\pm 0.004}$	$0.060_{\pm 0.005}$
$\lvert\mathcal{D}_P^{(2)}\rvert = 3000$	$0.139_{\pm 0.005}$	$0.072_{\pm 0.017}$	$0.064_{\pm 0.009}$	$0.058_{\pm 0.007}$	$0.057_{\pm 0.004}$	$0.058_{\pm 0.005}$

Weighted CDF or 1-Wasserstein

It is clear from the theoretical results that the weighted CDF version of the bound is provably tighter than the 1-Wasserstein distance. Therefore, one should expect (7) to produce better results, and this seems to be the case for most image classification datasets, with the exception of ImageNet-C with severity level 1. In contrast, for regression tasks, we observed the opposite trend: the 1-Wasserstein distance outperformed the weighted CDF bound. As shown in Figure 4, the weighted CDF bound led to overcoverage when paired with $(s_\sharp \hat{Q}_m^f, s_\sharp \hat{Q}_m^U)$ and produced less consistent results when used with $(s_\sharp \hat{Q}_m^{\min}, s_\sharp \hat{Q}_m^{\max})$. We conjecture that this discrepancy arises from the increased complexity of optimizing the weighted CDF bound, which may account for the divergent empirical outcomes. Indeed, the task of finding the density of scores needed for the weighted CDF distance is more intricate and precision-sensitive than simply identifying its maximum. As a result, in certain cases, the weighted CDF bound may underperform relative to its 1-Wasserstein counterpart.

Choice of auxiliary distributions

The tightness and practical utility of our bounds are influenced by the choice of auxiliary distributions $(s_\sharp Q^\downarrow, s_\sharp Q^\uparrow)$: the closer these are to the true distribution $s_\sharp Q$, the tighter the resulting bounds. However, our bounds have proven effective for learning the weights of $s_\sharp P^{\boldsymbol{w}}$ even when using uninformed and widely applicable auxiliary pairs such as $(s_\sharp Q^{\min}, s_\sharp Q^{\max})$ and $(s_\sharp Q^f, s_\sharp Q^U)$, which in general do not bound $s_\sharp Q$ tightly.

The performance of the two auxiliary distribution pairs was comparable in most cases, with $(s_\sharp Q^f, s_\sharp Q^U)$ generally achieving better coverage, albeit with a tendency to overcover. As mentioned in the main paper, $(s_\sharp Q^f, s_\sharp Q^U)$ is motivated by the observation that $s_\sharp Q^U$ tends to produce nonconformity scores higher than those from the true distribution $s_\sharp Q$—it corresponds to an uninformative model in which the correct label is independent of the model output—while $s_\sharp Q^f$ tends to yield lower nonconformity scores, reflecting a perfect model where the true class is sampled according to the model-assigned probabilities.

Beyond this theoretical motivation, there is also a practical reason for the better performance of $(s_\sharp Q^f, s_\sharp Q^U)$, which relates to the specific form of the nonconformity scores used—namely, one

Table 4: Total coverage gap on iWildcam and ImageNet-C with severity levels 1, 3 and 5 comparing optimization via the weighted-CDF (7) and the 1-Wasserstein (8) bounds with free-form parametrization. The classifier is given by a ResNet-50 and we consider both pairs $(s_\sharp Q^{\min}, s_\sharp Q^{\max})$ and $(s_\sharp Q^f, s_\sharp Q^U)$. For ImageNet-C we report the average across all 15 corruptions. Lower is better.

		iWildCam	ImageNet-C Sev. 1	ImageNet-C Sev. 3	ImageNet-C Sev. 5
	Uncorrected	$0.132_{\pm 0.016}$	$0.141_{\pm 0.048}$	$0.267_{\pm 0.081}$	$0.388_{\pm 0.076}$
weighted CDF	$(\min, \max)$	$0.084_{\pm 0.010}$	$0.098_{\pm 0.033}$	$0.171_{\pm 0.060}$	$0.281_{\pm 0.076}$
	(f, U)	$\mathbf{0.073}_{\pm 0.009}$	$0.059_{\pm 0.023}$	$\mathbf{0.102}_{\pm 0.036}$	$\mathbf{0.173}_{\pm 0.071}$
1-Wasserstein	$(\min, \max)$	$0.125_{\pm 0.005}$	$0.069_{\pm 0.023}$	$0.196_{\pm 0.062}$	$0.337_{\pm 0.093}$
	(f, U)	$0.125_{\pm 0.005}$	$\mathbf{0.044}_{\pm 0.020}$	$0.139_{\pm 0.066}$	$0.310_{\pm 0.108}$

minus the model-assigned probability for each class. Under this scoring scheme, the main difference between the two pairs arises from the contrast between $s_\sharp Q^f$ and $s_\sharp Q^{\min}$. Since most classes tend to receive relatively high nonconformity scores, $s_\sharp Q^U$ and $s_\sharp Q^{\max}$ are typically quite similar. On the other hand, unless the model is highly confident, $s_\sharp Q^f$ and $s_\sharp Q^{\min}$ can differ substantially. This explains why $s_\sharp Q^f$ may be more effective in practice. In particular, if we expect the model to have low accuracy under Q, $s_\sharp Q^{\min}$ becomes overly conservative and diverges significantly from the true $s_\sharp Q$, a pattern clearly illustrated in Figure 1. Therefore, we can improve performance by biasing $s_\sharp Q^\downarrow$ towards lower values, for instance, by sampling from the model, potentially with a high temperature.

Computing the Bounds

We compute the 1-Wasserstein version of the bound by estimating $\max_{s_c \in \mathbb{R}} p_{s_\sharp P}(s_c)$ with a Gaussian KDE and computing $W_1(s_\sharp \hat{P}_n, s_\sharp \hat{Q}_m)$ analytically. However, for the weighted CDF version of the bound we have a couple of options. The first is to treat the bound as a expectation under $s_\sharp P$, which can be approximated via the n samples $(X_i, Y_i)_{i=1}^n$ we have from P

$$\Delta_{P,Q} \leq \int p_{s_\sharp P}(s_c) \left| F_{s_\sharp \hat{P}_n}(s_c) - F_{s_\sharp \hat{Q}_m}(s_c) \right| = \mathbb{E}_{s_c \sim s_\sharp P} \left[\left| F_{s_\sharp \hat{P}_n}(s_c) - F_{s_\sharp \hat{Q}_m}(s_c) \right| \right] \tag{25}$$
$$\approx \sum_{i=1}^n \frac{1}{n} \left| F_{s_\sharp \hat{P}_n}(s(X_i, Y_i)) - F_{s_\sharp \hat{Q}_m}(s(X_i, Y_i)) \right|.$$

This gives tight estimates and is computationally cheap but did not prove useful as an optimization objective for learning a weighting scheme. Alternatively, we could compute the upper bound by numerical integration, which works well for unidimensional data. Since our nonconformity scores are bounded in $[0, 1]$, we use a grid of equally spaced K points s_k to get the following estimate

$$\Delta_{P,Q} \leq \int p_{s_\sharp P}(s_c) \left| F_{s_\sharp \hat{P}_n}(s_c) - F_{s_\sharp \hat{Q}_m}(s_c) \right| \approx \sum_{k=1}^K \frac{p_{s_\sharp P}(s_k)}{K} \left| F_{s_\sharp \hat{P}_n}(s_k) - F_{s_\sharp \hat{Q}_m}(s_k) \right|. \tag{26}$$

In this case, we have to estimate the probability $s_\sharp P(s_k)$ for each of the points in the grid. We do that with a Gaussian KDE, using reflection [20] to deal with the boundaries in $[0, 1]$. This proved a better optimization objective, facilitating the learning of the weighted distribution $s_\sharp \hat{P}_n^{\boldsymbol{w}}$. Thus, when computing the weighted-CDF version of the bound we use the numerical integration method as in (26) for both training and evaluation; see Table 5 for an analysis of the tightness of the bound (26) in ImageNet-C with severity 5. Note that in all cases, we can replace $s_\sharp \hat{P}_n$ with $s_\sharp \hat{P}_n^{\boldsymbol{w}}$.

In terms of complexity, the bounds require computing either weighted CDF or 1-Wasserstein distances, which are tractable for unidimensional variables like nonconformity scores. In both cases, we need to compute the difference between the empirical CDFs, which has overall time complexity $\mathcal{O}((m+n)\log(m+n))$. The complexity here is dominated by the sorting operation needed to compute the difference between the empirical CDFs, but fortunately this operation is applied only to the score values and not to the weights, and thus we need to compute it only once during optimization. Finally, with the exception of the bound in (25), we also need to estimate the density $p_{s_\sharp P}$. We do so via a Gaussian KDE defined on the n samples from $s_\sharp P$, which has cost $\mathcal{O}(k \cdot n)$, where k is the number of points the KDE is evaluated on, e.g. the grid size in (26). We set the KDE bandwidth using Scott's rule [37] but scale it by a factor of 0.1 in classification tasks to improve resolution in the tails.

 https://doi.org/10.52202/085713-5278

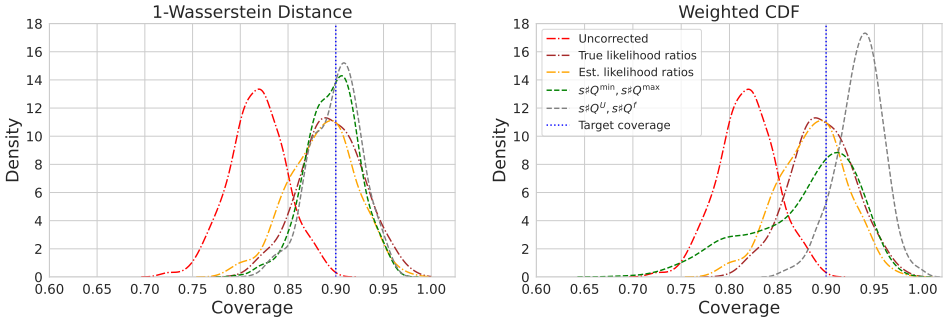

Figure 4: Distribution of coverage for the synthetic regression task across 500 simulations and target coverage rate of 90% (blue vertical line). Results with the 1-Wasserstein distance formulation on the left and with the weighted CDF formulation on the right. The baselines remain the same in both plots. For ease of visualization, we plot the density estimated with a KDE fit to the 500 observations.

C.5.3 Tightness of the Bounds

Our upper bounds to the total coverage gap include terms that depend only on the auxiliary distributions, such as $\int_{\mathbb{R}} F_{s_\sharp Q^\downarrow}(s_c) - F_{s_\sharp Q^\uparrow}(s_c) ds_c$ in (7) and $\mathbb{E}_{s_\sharp Q^\uparrow}[S] - \mathbb{E}_{s_\sharp Q^\downarrow}[S]$ in (8). Therefore, we cannot hope to have tight bounds, unless $s_\sharp Q^\uparrow$ and $s_\sharp Q^\downarrow$ are close to each other and sandwich $s_\sharp Q$, i.e., satisfy the stochastic dominance relation $s_\sharp Q^\uparrow \succcurlyeq s_\sharp Q \succcurlyeq s_\sharp Q^\downarrow$. To illustrate this point, we evaluate our upper bounds on ImageNet-C with severity level 5 in Table 5, where we can see the upper bounds constructed with the auxiliary distributions are relatively loose, as expected.

As discussed throughout the paper and demonstrated in the experiments, these bounds, although not tight, are still effective in mitigating the coverage gap by serving as a practical optimization objective for learning $s_\sharp P^{\boldsymbol{w}}$. Nevertheless, we conjecture that there is still room to improve the tightness of these bounds. We leave further improvements for future work, but one promising direction is to learn a transformation of the scores under Q jointly with the weights of $s_\sharp P^{\boldsymbol{w}}$. Although this approach would still require auxiliary distributions to evaluate our bounds, it could yield tighter estimates, for instance, by reducing the gap between $s_\sharp Q^{\min}$ and $s_\sharp Q^{\max}$.

Table 5: Upper bounds to the total coverage gap for ImageNet-C with severity level 5. For each pair of auxiliary distributions we consider, we have the total coverage gap $\Delta_{P^{\boldsymbol{w}},Q}$ after optimization, and the weighted-CDF upper bound computed with unlabeled samples and no DKW correction. We report mean and standard deviation across 10 random seeds. Lower is better.

Corruption	$\Delta_{P^{\boldsymbol{w}},Q}$	$(s_\sharp Q^{\min}, s_\sharp Q^{\max})$	$\Delta_{P^{\boldsymbol{w}},Q}$	$(s_\sharp Q^f, s_\sharp Q^U)$
Gauss	$0.377_{\pm 0.026}$	$0.445_{\pm 0.014}$	$0.264_{\pm 0.033}$	$0.291_{\pm 0.019}$
Shot	$0.378_{\pm 0.024}$	$0.441_{\pm 0.017}$	$0.291_{\pm 0.028}$	$0.319_{\pm 0.015}$
Impul	$0.378_{\pm 0.026}$	$0.450_{\pm 0.013}$	$0.274_{\pm 0.027}$	$0.306_{\pm 0.012}$
Defoc	$0.258_{\pm 0.030}$	$0.328_{\pm 0.022}$	$0.094_{\pm 0.013}$	$0.113_{\pm 0.018}$
Glass	$0.316_{\pm 0.030}$	$0.389_{\pm 0.019}$	$0.167_{\pm 0.032}$	$0.208_{\pm 0.021}$
Motion	$0.291_{\pm 0.028}$	$0.354_{\pm 0.019}$	$0.168_{\pm 0.023}$	$0.209_{\pm 0.013}$
Zoom	$0.252_{\pm 0.027}$	$0.306_{\pm 0.024}$	$0.153_{\pm 0.027}$	$0.190_{\pm 0.030}$
Snow	$0.289_{\pm 0.026}$	$0.333_{\pm 0.027}$	$0.201_{\pm 0.026}$	$0.226_{\pm 0.025}$
Frost	$0.254_{\pm 0.026}$	$0.298_{\pm 0.025}$	$0.162_{\pm 0.023}$	$0.189_{\pm 0.021}$
Fog	$0.246_{\pm 0.028}$	$0.296_{\pm 0.025}$	$0.166_{\pm 0.029}$	$0.202_{\pm 0.028}$
Bright	$0.095_{\pm 0.009}$	$0.098_{\pm 0.011}$	$0.056_{\pm 0.009}$	$0.060_{\pm 0.009}$
Contr	$0.327_{\pm 0.034}$	$0.413_{\pm 0.015}$	$0.114_{\pm 0.034}$	$0.134_{\pm 0.016}$
Elastic	$0.298_{\pm 0.025}$	$0.354_{\pm 0.026}$	$0.232_{\pm 0.023}$	$0.275_{\pm 0.024}$
Pixel	$0.257_{\pm 0.027}$	$0.314_{\pm 0.022}$	$0.147_{\pm 0.021}$	$0.184_{\pm 0.009}$
Jpeg	$0.196_{\pm 0.023}$	$0.230_{\pm 0.030}$	$0.106_{\pm 0.013}$	$0.119_{\pm 0.017}$

Table 6: **Coverage on ImageNet-C with ResNet-50 with and without label shift (see Section 6.2)**. Results for uncorrected distributions, calibrating and testing on samples from Q (Oracle), estimated likelihood ratios (LR), the methods of Kasa et al. and Gibbs et al., and weights learned via our weighted CDF objective (7), including $(\min, \max)$ and (f, U) variants with free-form (FF) and weight function (WF) parametrizations. Target coverage of 90%. We report mean and standard deviation across 10 random seeds.

		Noise			Blur				Weather					Digital				
		Gauss	Shot	Impul	Defoc	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr	Elastic	Pixel	Jpeg	Avg. Cov.	Avg. Size
Severity 1 no label shift	Uncorrected	$78.0_{\pm 3.1}$	$76.8_{\pm 3.2}$	$68.2_{\pm 3.8}$	$78.4_{\pm 3.5}$	$73.2_{\pm 3.6}$	$82.2_{\pm 2.8}$	$71.8_{\pm 3.7}$	$73.2_{\pm 3.3}$	$78.8_{\pm 3.0}$	$80.3_{\pm 3.2}$	$88.5_{\pm 1.8}$	$82.6_{\pm 2.9}$	$83.5_{\pm 2.6}$	$81.7_{\pm 2.7}$	$83.4_{\pm 2.6}$	$78.7_{\pm 6.0}$	$2.7_{\pm 0.7}$
	Oracle	$89.7_{\pm 1.7}$	$89.2_{\pm 1.9}$	$89.7_{\pm 1.6}$	$89.7_{\pm 2.1}$	$89.8_{\pm 1.5}$	$90.5_{\pm 1.6}$	$90.1_{\pm 1.4}$	$89.2_{\pm 0.8}$	$89.7_{\pm 2.1}$	$89.4_{\pm 1.6}$	$90.1_{\pm 1.2}$	$90.6_{\pm 1.5}$	$90.4_{\pm 1.3}$	$89.5_{\pm 1.3}$	$89.7_{\pm 1.7}$	$89.8_{\pm 1.6}$	$10.7_{\pm 7.2}$
	LR	$82.2_{\pm 3.2}$	$82.6_{\pm 4.2}$	$77.3_{\pm 5.7}$	$85.2_{\pm 3.9}$	$81.2_{\pm 4.3}$	$86.0_{\pm 3.6}$	$81.3_{\pm 4.3}$	$79.8_{\pm 4.0}$	$84.0_{\pm 4.2}$	$86.2_{\pm 4.1}$	$89.5_{\pm 2.2}$	$86.8_{\pm 3.4}$	$87.3_{\pm 3.5}$	$85.4_{\pm 3.0}$	$86.5_{\pm 3.2}$	$84.1_{\pm 4.9}$	$5.2_{\pm 3.8}$
	Kasa et al.	$96.2_{\pm 0.8}$	$96.1_{\pm 0.8}$	$95.2_{\pm 1.2}$	$97.5_{\pm 0.7}$	$96.4_{\pm 0.9}$	$97.3_{\pm 0.7}$	$95.6_{\pm 1.0}$	$95.3_{\pm 1.0}$	$96.6_{\pm 0.8}$	$97.1_{\pm 0.7}$	$97.6_{\pm 0.5}$	$97.4_{\pm 0.6}$	$97.5_{\pm 0.6}$	$96.6_{\pm 0.7}$	$97.0_{\pm 0.7}$	$96.6_{\pm 1.1}$	$42.4_{\pm 19.2}$
	Gibbs et al.	$88.1_{\pm 2.0}$	$87.8_{\pm 1.8}$	$86.1_{\pm 2.0}$	$89.8_{\pm 1.8}$	$87.4_{\pm 2.1}$	$90.6_{\pm 2.3}$	$87.0_{\pm 2.8}$	$85.4_{\pm 3.0}$	$89.1_{\pm 2.3}$	$89.1_{\pm 1.9}$	$92.4_{\pm 2.1}$	$89.7_{\pm 2.6}$	$91.3_{\pm 2.4}$	$89.0_{\pm 1.7}$	$89.9_{\pm 2.2}$	$88.9_{\pm 2.8}$	$548.5_{\pm 36.7}$
	FF $(\min, \max)$	$91.1_{\pm 3.8}$	$89.5_{\pm 5.1}$	$88.0_{\pm 3.6}$	$92.9_{\pm 3.0}$	$89.0_{\pm 4.6}$	$92.9_{\pm 2.9}$	$89.5_{\pm 4.0}$	$88.9_{\pm 3.4}$	$91.2_{\pm 3.9}$	$92.8_{\pm 3.7}$	$93.9_{\pm 2.9}$	$93.5_{\pm 2.3}$	$92.8_{\pm 4.0}$	$91.7_{\pm 3.4}$	$93.0_{\pm 2.6}$	$91.4_{\pm 3.9}$	$15.1_{\pm 9.7}$
	WF $(\min, \max)$	$93.7_{\pm 4.4}$	$92.5_{\pm 9.2}$	$92.6_{\pm 3.4}$	$96.3_{\pm 2.0}$	$94.5_{\pm 2.4}$	$96.8_{\pm 1.5}$	$93.6_{\pm 2.9}$	$93.7_{\pm 2.6}$	$95.7_{\pm 1.9}$	$88.5_{\pm 19.6}$	$95.4_{\pm 5.6}$	$96.8_{\pm 1.4}$	$96.9_{\pm 1.5}$	$95.9_{\pm 2.0}$	$96.5_{\pm 1.5}$	$94.6_{\pm 6.3}$	$34.7_{\pm 20.6}$
	FF (f, U)	$93.1_{\pm 2.5}$	$91.1_{\pm 3.0}$	$88.6_{\pm 3.8}$	$94.2_{\pm 2.2}$	$91.4_{\pm 1.7}$	$94.5_{\pm 2.6}$	$91.7_{\pm 3.1}$	$90.6_{\pm 3.2}$	$93.4_{\pm 2.7}$	$94.3_{\pm 2.1}$	$95.6_{\pm 1.4}$	$94.4_{\pm 1.6}$	$95.0_{\pm 1.9}$	$93.6_{\pm 2.0}$	$93.9_{\pm 2.2}$	$93.0_{\pm 3.0}$	$19.2_{\pm 11.4}$
	WF (f, U)	$95.3_{\pm 2.0}$	$95.1_{\pm 2.0}$	$93.2_{\pm 2.6}$	$96.4_{\pm 1.9}$	$94.7_{\pm 2.5}$	$96.7_{\pm 1.6}$	$94.3_{\pm 2.2}$	$94.2_{\pm 2.1}$	$95.8_{\pm 1.9}$	$96.0_{\pm 2.0}$	$97.5_{\pm 0.9}$	$96.8_{\pm 1.5}$	$96.9_{\pm 1.5}$	$96.0_{\pm 1.6}$	$96.6_{\pm 1.6}$	$95.7_{\pm 2.2}$	$36.6_{\pm 20.1}$
Severity 1 with label shift	Uncorrected	$78.8_{\pm 5.3}$	$77.4_{\pm 6.2}$	$69.2_{\pm 8.0}$	$78.9_{\pm 7.3}$	$76.7_{\pm 6.0}$	$83.8_{\pm 5.6}$	$72.1_{\pm 7.8}$	$72.5_{\pm 9.5}$	$79.2_{\pm 7.3}$	$78.7_{\pm 4.5}$	$89.5_{\pm 2.4}$	$82.9_{\pm 8.0}$	$84.6_{\pm 6.0}$	$80.0_{\pm 7.2}$	$83.3_{\pm 5.5}$	$79.2_{\pm 8.2}$	$2.7_{\pm 0.8}$
	Oracle	$89.0_{\pm 3.7}$	$87.1_{\pm 5.5}$	$91.7_{\pm 4.7}$	$93.7_{\pm 1.1}$	$90.5_{\pm 4.2}$	$90.8_{\pm 4.7}$	$90.4_{\pm 3.3}$	$90.9_{\pm 6.7}$	$89.0_{\pm 3.3}$	$87.2_{\pm 6.3}$	$89.9_{\pm 2.7}$	$93.2_{\pm 3.2}$	$90.1_{\pm 5.5}$	$90.1_{\pm 3.5}$	$90.4_{\pm 3.8}$	$90.3_{\pm 4.5}$	$15.0_{\pm 19.9}$
	LR	$85.2_{\pm 5.5}$	$84.8_{\pm 4.4}$	$78.2_{\pm 6.9}$	$82.9_{\pm 8.7}$	$80.1_{\pm 6.4}$	$88.0_{\pm 6.6}$	$82.1_{\pm 5.1}$	$79.2_{\pm 9.8}$	$85.3_{\pm 7.7}$	$85.8_{\pm 5.8}$	$89.0_{\pm 7.2}$	$89.3_{\pm 3.7}$	$85.0_{\pm 6.4}$	$84.6_{\pm 7.5}$	$87.4_{\pm 5.0}$	$84.5_{\pm 7.1}$	$6.1_{\pm 5.5}$
	Kasa et al.	$78.8_{\pm 5.3}$	$77.4_{\pm 6.2}$	$69.3_{\pm 7.9}$	$78.9_{\pm 7.3}$	$77.0_{\pm 5.9}$	$83.8_{\pm 5.6}$	$72.1_{\pm 7.8}$	$72.5_{\pm 9.5}$	$79.2_{\pm 7.3}$	$78.7_{\pm 4.5}$	$89.5_{\pm 2.4}$	$82.9_{\pm 8.0}$	$84.6_{\pm 6.0}$	$80.0_{\pm 7.2}$	$83.3_{\pm 5.5}$	$79.2_{\pm 8.1}$	$2.7_{\pm 0.8}$
	Gibbs et al.	$87.8_{\pm 2.6}$	$87.8_{\pm 4.3}$	$84.6_{\pm 2.2}$	$89.0_{\pm 2.1}$	$87.9_{\pm 3.2}$	$88.9_{\pm 2.4}$	$85.7_{\pm 4.2}$	$86.7_{\pm 5.7}$	$89.4_{\pm 4.3}$	$89.3_{\pm 4.7}$	$91.9_{\pm 2.9}$	$90.4_{\pm 2.2}$	$89.2_{\pm 3.4}$	$89.3_{\pm 1.9}$	$90.9_{\pm 4.3}$	$88.6_{\pm 3.8}$	$552.3_{\pm 46.0}$
	FF $(\min, \max)$	$90.1_{\pm 3.6}$	$89.5_{\pm 4.1}$	$85.7_{\pm 9.5}$	$93.7_{\pm 3.8}$	$87.9_{\pm 7.2}$	$93.9_{\pm 4.1}$	$87.8_{\pm 5.8}$	$88.1_{\pm 6.2}$	$91.0_{\pm 5.3}$	$91.1_{\pm 4.7}$	$94.1_{\pm 3.3}$	$94.8_{\pm 3.8}$	$94.5_{\pm 3.3}$	$90.5_{\pm 5.1}$	$91.8_{\pm 5.8}$	$91.0_{\pm 5.7}$	$14.3_{\pm 10.2}$
	WF $(\min, \max)$	$93.9_{\pm 2.9}$	$94.8_{\pm 2.6}$	$93.1_{\pm 2.4}$	$95.1_{\pm 3.2}$	$94.7_{\pm 3.5}$	$96.7_{\pm 1.6}$	$92.0_{\pm 5.2}$	$92.8_{\pm 4.6}$	$94.9_{\pm 3.2}$	$92.4_{\pm 11.5}$	$85.1_{\pm 24.4}$	$90.2_{\pm 21.0}$	$96.7_{\pm 2.7}$	$96.0_{\pm 2.1}$	$96.6_{\pm 1.8}$	$93.7_{\pm 9.3}$	$33.1_{\pm 21.5}$
	FF (f, U)	$92.3_{\pm 3.2}$	$92.5_{\pm 3.5}$	$89.9_{\pm 4.8}$	$95.6_{\pm 2.6}$	$92.9_{\pm 3.6}$	$94.4_{\pm 3.6}$	$91.1_{\pm 5.2}$	$93.2_{\pm 2.8}$	$92.2_{\pm 2.5}$	$94.5_{\pm 2.4}$	$95.3_{\pm 3.1}$	$95.1_{\pm 3.0}$	$96.1_{\pm 2.0}$	$92.8_{\pm 4.4}$	$94.4_{\pm 2.4}$	$93.5_{\pm 3.6}$	$20.7_{\pm 14.6}$
	WF (f, U)	$93.9_{\pm 3.1}$	$95.8_{\pm 2.5}$	$93.5_{\pm 3.5}$	$96.9_{\pm 3.9}$	$95.4_{\pm 2.6}$	$97.0_{\pm 2.4}$	$94.2_{\pm 3.9}$	$93.6_{\pm 2.9}$	$95.7_{\pm 3.7}$	$95.8_{\pm 3.0}$	$89.4_{\pm 25.3}$	$96.3_{\pm 4.0}$	$97.4_{\pm 1.8}$	$95.5_{\pm 2.7}$	$96.4_{\pm 2.4}$	$95.1_{\pm 7.1}$	$36.3_{\pm 21.8}$
Severity 3 no label shift	Uncorrected	$50.3_{\pm 4.6}$	$46.7_{\pm 4.4}$	$46.9_{\pm 4.7}$	$57.3_{\pm 5.4}$	$30.3_{\pm 4.2}$	$56.6_{\pm 4.7}$	$54.2_{\pm 4.6}$	$53.0_{\pm 4.1}$	$48.9_{\pm 4.3}$	$67.0_{\pm 4.2}$	$85.7_{\pm 2.3}$	$66.3_{\pm 4.3}$	$73.4_{\pm 3.2}$	$66.1_{\pm 3.9}$	$78.3_{\pm 3.1}$	$58.7_{\pm 14.3}$	$3.3_{\pm 1.2}$
	Oracle	$89.5_{\pm 1.4}$	$89.5_{\pm 2.4}$	$90.0_{\pm 2.0}$	$90.0_{\pm 1.8}$	$89.5_{\pm 2.3}$	$90.7_{\pm 1.7}$	$90.4_{\pm 1.6}$	$89.3_{\pm 0.9}$	$89.5_{\pm 2.1}$	$89.6_{\pm 1.3}$	$90.0_{\pm 1.6}$	$89.8_{\pm 1.5}$	$90.4_{\pm 1.7}$	$89.7_{\pm 1.3}$	$89.9_{\pm 1.4}$	$89.9_{\pm 1.7}$	$80.1_{\pm 71.4}$
	LR	$64.2_{\pm 7.4}$	$62.8_{\pm 8.3}$	$63.7_{\pm 9.1}$	$75.3_{\pm 8.5}$	$48.2_{\pm 10.3}$	$71.5_{\pm 8.6}$	$69.8_{\pm 7.6}$	$65.7_{\pm 7.0}$	$63.8_{\pm 8.1}$	$78.2_{\pm 5.9}$	$87.4_{\pm 2.7}$	$78.3_{\pm 5.6}$	$80.8_{\pm 4.7}$	$76.2_{\pm 5.3}$	$82.9_{\pm 4.2}$	$71.3_{\pm 11.9}$	$12.3_{\pm 13.1}$
	Kasa et al.	$92.4_{\pm 1.9}$	$91.6_{\pm 2.1}$	$92.0_{\pm 2.0}$	$97.2_{\pm 0.9}$	$87.6_{\pm 3.2}$	$94.4_{\pm 1.5}$	$92.9_{\pm 1.7}$	$90.2_{\pm 2.1}$	$90.8_{\pm 2.2}$	$95.4_{\pm 1.2}$	$97.4_{\pm 0.5}$	$96.3_{\pm 1.1}$	$95.7_{\pm 1.0}$	$94.5_{\pm 1.3}$	$96.5_{\pm 0.9}$	$93.7_{\pm 3.2}$	$119.3_{\pm 65.4}$
	Gibbs et al.	$83.1_{\pm 1.8}$	$81.7_{\pm 2.3}$	$83.2_{\pm 2.7}$	$89.3_{\pm 1.8}$	$83.5_{\pm 2.0}$	$87.3_{\pm 2.5}$	$83.6_{\pm 2.5}$	$82.2_{\pm 3.0}$	$83.0_{\pm 2.1}$	$85.7_{\pm 2.1}$	$91.0_{\pm 1.6}$	$86.3_{\pm 2.7}$	$89.1_{\pm 2.2}$	$85.5_{\pm 2.1}$	$88.8_{\pm 2.2}$	$85.6_{\pm 3.6}$	$635.7_{\pm 66.3}$
	FF $(\min, \max)$	$85.1_{\pm 5.7}$	$83.3_{\pm 6.7}$	$84.2_{\pm 6.3}$	$92.5_{\pm 4.3}$	$73.5_{\pm 9.9}$	$89.3_{\pm 4.5}$	$86.6_{\pm 4.7}$	$83.6_{\pm 4.8}$	$83.1_{\pm 6.2}$	$92.3_{\pm 3.2}$	$96.8_{\pm 1.2}$	$92.9_{\pm 3.2}$	$93.4_{\pm 2.6}$	$91.2_{\pm 3.6}$	$95.1_{\pm 2.0}$	$88.2_{\pm 7.7}$	$63.9_{\pm 41.3}$
	WF $(\min, \max)$	$85.7_{\pm 6.4}$	$83.6_{\pm 7.1}$	$84.6_{\pm 6.8}$	$92.7_{\pm 4.7}$	$74.6_{\pm 9.5}$	$89.8_{\pm 5.0}$	$87.3_{\pm 5.5}$	$84.4_{\pm 5.9}$	$83.4_{\pm 6.6}$	$93.2_{\pm 2.9}$	$97.2_{\pm 1.2}$	$84.6_{\pm 28.8}$	$94.1_{\pm 2.6}$	$82.9_{\pm 27.9}$	$86.7_{\pm 27.8}$	$87.0_{\pm 14.1}$	$69.7_{\pm 46.6}$
	FF (f, U)	$87.9_{\pm 4.3}$	$86.1_{\pm 4.9}$	$86.9_{\pm 4.6}$	$95.0_{\pm 2.6}$	$80.0_{\pm 6.3}$	$91.3_{\pm 3.3}$	$89.0_{\pm 3.8}$	$86.0_{\pm 4.1}$	$85.6_{\pm 4.8}$	$93.4_{\pm 2.7}$	$97.0_{\pm 1.2}$	$94.1_{\pm 2.7}$	$94.2_{\pm 2.5}$	$92.1_{\pm 3.1}$	$95.6_{\pm 1.9}$	$90.3_{\pm 5.9}$	$79.8_{\pm 48.9}$
	WF (f, U)	$87.4_{\pm 5.1}$	$85.3_{\pm 5.5}$	$86.4_{\pm 5.4}$	$94.7_{\pm 2.7}$	$79.7_{\pm 6.4}$	$91.2_{\pm 3.8}$	$88.8_{\pm 4.4}$	$85.8_{\pm 4.7}$	$85.3_{\pm 5.4}$	$93.4_{\pm 2.7}$	$97.1_{\pm 1.4}$	$94.1_{\pm 2.7}$	$94.5_{\pm 2.2}$	$92.3_{\pm 2.9}$	$95.6_{\pm 2.0}$	$90.1_{\pm 6.2}$	$79.7_{\pm 50.1}$
Severity 3 with label shift	Uncorrected	$51.4_{\pm 8.6}$	$48.3_{\pm 11.1}$	$47.5_{\pm 8.6}$	$57.0_{\pm 8.0}$	$31.4_{\pm 8.2}$	$58.6_{\pm 7.3}$	$53.5_{\pm 7.7}$	$54.1_{\pm 6.9}$	$49.8_{\pm 9.9}$	$67.3_{\pm 5.8}$	$86.6_{\pm 4.2}$	$65.4_{\pm 10.8}$	$75.4_{\pm 8.0}$	$64.8_{\pm 11.6}$	$76.5_{\pm 7.1}$	$59.2_{\pm 15.7}$	$3.2_{\pm 1.2}$
	Oracle	$89.3_{\pm 4.0}$	$89.5_{\pm 3.9}$	$92.9_{\pm 2.6}$	$91.8_{\pm 3.0}$	$89.1_{\pm 3.5}$	$89.9_{\pm 4.0}$	$89.1_{\pm 4.3}$	$90.1_{\pm 3.4}$	$89.9_{\pm 2.9}$	$88.0_{\pm 5.2}$	$88.8_{\pm 3.3}$	$93.0_{\pm 4.2}$	$91.2_{\pm 4.3}$	$90.7_{\pm 2.6}$	$90.4_{\pm 4.2}$	$90.2_{\pm 3.8}$	$88.1_{\pm 87.4}$
	LR	$69.2_{\pm 8.8}$	$64.3_{\pm 12.0}$	$64.6_{\pm 11.4}$	$78.0_{\pm 8.6}$	$49.9_{\pm 12.9}$	$70.4_{\pm 9.5}$	$72.9_{\pm 5.2}$	$67.6_{\pm 12.5}$	$63.3_{\pm 13.4}$	$78.5_{\pm 6.5}$	$88.7_{\pm 6.6}$	$81.2_{\pm 4.8}$	$77.2_{\pm 6.6}$	$77.3_{\pm 5.3}$	$84.1_{\pm 6.0}$	$72.5_{\pm 12.9}$	$14.3_{\pm 15.8}$
	Kasa et al.	$51.9_{\pm 8.6}$	$48.3_{\pm 11.1}$	$48.9_{\pm 8.0}$	$57.6_{\pm 8.5}$	$34.5_{\pm 7.0}$	$58.9_{\pm 7.6}$	$54.1_{\pm 8.0}$	$54.3_{\pm 6.9}$	$50.0_{\pm 10.0}$	$67.3_{\pm 5.8}$	$86.6_{\pm 4.2}$	$66.4_{\pm 10.6}$	$75.4_{\pm 8.0}$	$65.1_{\pm 11.3}$	$76.5_{\pm 7.1}$	$59.7_{\pm 15.2}$	$3.4_{\pm 1.4}$
	Gibbs et al.	$82.7_{\pm 2.8}$	$84.6_{\pm 2.8}$	$80.7_{\pm 3.0}$	$89.4_{\pm 2.6}$	$82.9_{\pm 3.4}$	$86.5_{\pm 3.0}$	$82.9_{\pm 3.9}$	$81.5_{\pm 4.8}$	$84.4_{\pm 3.4}$	$86.0_{\pm 4.6}$	$91.0_{\pm 2.7}$	$87.8_{\pm 4.7}$	$88.3_{\pm 3.2}$	$86.0_{\pm 4.6}$	$87.9_{\pm 4.0}$	$85.5_{\pm 4.6}$	$638.6_{\pm 70.2}$
	FF $(\min, \max)$	$85.6_{\pm 4.7}$	$84.9_{\pm 8.5}$	$83.5_{\pm 9.2}$	$92.9_{\pm 4.1}$	$74.9_{\pm 13.3}$	$88.9_{\pm 5.9}$	$86.1_{\pm 7.2}$	$83.8_{\pm 6.1}$	$82.9_{\pm 8.3}$	$91.7_{\pm 4.6}$	$97.2_{\pm 2.1}$	$93.8_{\pm 4.0}$	$94.2_{\pm 3.4}$	$90.1_{\pm 4.7}$	$95.3_{\pm 3.8}$	$88.4_{\pm 8.6}$	$62.6_{\pm 41.4}$
	WF $(\min, \max)$	$84.8_{\pm 6.8}$	$85.1_{\pm 8.1}$	$86.2_{\pm 6.7}$	$92.7_{\pm 4.7}$	$74.6_{\pm 11.2}$	$90.0_{\pm 6.8}$	$85.9_{\pm 12.0}$	$81.5_{\pm 7.7}$	$82.4_{\pm 6.0}$	$92.8_{\pm 5.4}$	$88.1_{\pm 27.5}$	$93.0_{\pm 5.6}$	$94.3_{\pm 3.3}$	$82.4_{\pm 27.9}$	$95.5_{\pm 2.6}$	$87.3_{\pm 12.8}$	$69.4_{\pm 47.3}$
	FF (f, U)	$87.5_{\pm 3.3}$	$85.6_{\pm 5.9}$	$87.7_{\pm 7.4}$	$95.4_{\pm 3.4}$	$77.3_{\pm 8.1}$	$91.7_{\pm 4.2}$	$88.0_{\pm 5.1}$	$88.3_{\pm 4.1}$	$84.7_{\pm 4.6}$	$93.3_{\pm 4.4}$	$97.7_{\pm 1.8}$	$93.8_{\pm 3.9}$	$95.6_{\pm 1.7}$	$92.4_{\pm 3.8}$	$95.6_{\pm 1.8}$	$90.3_{\pm 6.8}$	$78.9_{\pm 51.7}$
	WF (f, U)	$86.4_{\pm 5.6}$	$87.7_{\pm 5.5}$	$87.8_{\pm 6.5}$	$93.8_{\pm 8.9}$	$81.1_{\pm 9.9}$	$92.4_{\pm 3.9}$	$88.8_{\pm 5.7}$	$86.9_{\pm 6.1}$	$86.5_{\pm 8.0}$	$93.6_{\pm 3.4}$	$96.9_{\pm 3.0}$	$92.6_{\pm 7.2}$	$94.7_{\pm 3.3}$	$92.0_{\pm 2.7}$	$95.6_{\pm 2.2}$	$90.4_{\pm 7.0}$	$80.0_{\pm 51.4}$
Severity 5 no label shift	Uncorrected	$6.9_{\pm 1.8}$	$8.3_{\pm 1.9}$	$6.5_{\pm 1.8}$	$30.2_{\pm 5.4}$	$18.4_{\pm 3.7}$	$26.8_{\pm 4.3}$	$38.0_{\pm 4.5}$	$29.4_{\pm 3.7}$	$37.6_{\pm 4.1}$	$40.5_{\pm 4.3}$	$78.0_{\pm 3.2}$	$9.6_{\pm 2.6}$	$30.3_{\pm 3.8}$	$36.2_{\pm 4.4}$	$50.9_{\pm 4.7}$	$29.8_{\pm 18.9}$	$3.3_{\pm 1.5}$
	Oracle	$89.5_{\pm 2.3}$	$89.2_{\pm 2.1}$	$89.5_{\pm 2.6}$	$89.9_{\pm 2.0}$	$89.6_{\pm 2.5}$	$90.4_{\pm 1.0}$	$89.9_{\pm 1.7}$	$89.3_{\pm 1.6}$	$89.2_{\pm 1.5}$	$90.3_{\pm 2.4}$	$89.7_{\pm 1.5}$	$89.0_{\pm 1.4}$	$90.4_{\pm 2.4}$	$89.9_{\pm 1.6}$	$90.3_{\pm 1.5}$	$89.7_{\pm 1.9}$	$338.6_{\pm 190.9}$
	LR	$21.5_{\pm 11.4}$	$23.1_{\pm 13.0}$	$20.9_{\pm 12.6}$	$57.7_{\pm 15.6}$	$40.1_{\pm 13.6}$	$48.4_{\pm 13.9}$	$57.3_{\pm 10.2}$	$47.1_{\pm 10.5}$	$54.3_{\pm 9.7}$	$57.5_{\pm 9.1}$	$82.9_{\pm 4.3}$	$32.8_{\pm 16.5}$	$45.9_{\pm 8.1}$	$55.6_{\pm 10.7}$	$66.5_{\pm 7.7}$	$47.4_{\pm 20.2}$	$27.7_{\pm 37.0}$
	Kasa et al.	$76.1_{\pm 6.1}$	$71.1_{\pm 6.1}$	$75.9_{\pm 6.2}$	$96.1_{\pm 1.6}$	$87.4_{\pm 3.8}$	$87.2_{\pm 3.0}$	$89.0_{\pm 2.7}$	$82.6_{\pm 3.7}$	$86.5_{\pm 2.9}$	$88.2_{\pm 2.8}$	$97.0_{\pm 0.7}$	$93.4_{\pm 3.3}$	$78.3_{\pm 3.6}$	$89.2_{\pm 2.8}$	$93.2_{\pm 1.9}$	$86.1_{\pm 8.4}$	$282.0_{\pm 159.1}$
	Gibbs et al.	$86.9_{\pm 2.2}$	$83.4_{\pm 2.9}$	$86.3_{\pm 1.9}$	$92.8_{\pm 1.7}$	$88.6_{\pm 1.7}$	$85.8_{\pm 1.9}$	$82.2_{\pm 2.8}$	$78.6_{\pm 2.4}$	$81.1_{\pm 1.9}$	$79.3_{\pm 2.6}$	$87.5_{\pm 2.4}$	$95.2_{\pm 1.1}$	$75.1_{\pm 3.4}$	$85.5_{\pm 3.5}$	$85.4_{\pm 1.7}$	$84.9_{\pm 5.6}$	$754.4_{\pm 102.7}$
	FF $(\min, \max)$	$48.7_{\pm 15.2}$	$46.9_{\pm 13.5}$	$48.2_{\pm 15.1}$	$84.3_{\pm 9.7}$	$68.5_{\pm 12.8}$	$73.9_{\pm 10.1}$	$78.8_{\pm 7.8}$	$70.1_{\pm 9.6}$	$76.2_{\pm 7.9}$	$79.1_{\pm 7.9}$	$95.6_{\pm 1.9}$	$69.0_{\pm 13.6}$	$67.2_{\pm 8.4}$	$76.8_{\pm 8.6}$	$86.1_{\pm 6.0}$	$71.3_{\pm 17.0}$	$128.5_{\pm 93.0}$
	WF $(\min, \max)$	$49.8_{\pm 14.8}$	$46.9_{\pm 14.1}$	$48.3_{\pm 15.8}$	$86.1_{\pm 7.5}$	$69.2_{\pm 12.3}$	$73.6_{\pm 10.5}$	$78.8_{\pm 8.2}$	$70.0_{\pm 10.1}$	$76.9_{\pm 7.7}$	$78.9_{\pm 8.3}$	$95.7_{\pm 2.1}$	$67.9_{\pm 14.8}$	$67.1_{\pm 8.8}$	$76.8_{\pm 9.0}$	$85.9_{\pm 6.3}$	$71.5_{\pm 17.1}$	$130.5_{\pm 94.1}$
	FF (f, U)	$66.2_{\pm 10.8}$	$61.7_{\pm 11.2}$	$65.8_{\pm 10.8}$	$92.9_{\pm 4.5}$	$81.2_{\pm 8.1}$	$79.9_{\pm 5.6}$	$81.8_{\pm 5.7}$	$74.9_{\pm 6.6}$	$80.3_{\pm 5.4}$	$81.5_{\pm 6.1}$	$96.0_{\pm 1.9}$	$88.5_{\pm 10.9}$	$69.8_{\pm 6.7}$	$83.0_{\pm 4.9}$	$88.2_{\pm 4.3}$	$79.5_{\pm 12.1}$	$205.7_{\pm 164.4}$
	WF (f, U)	$66.4_{\pm 10.8}$	$57.2_{\pm 11.3}$	$63.1_{\pm 11.8}$	$93.3_{\pm 3.9}$	$79.9_{\pm 8.2}$	$79.6_{\pm 7.1}$	$82.3_{\pm 5.4}$	$73.3_{\pm 7.9}$	$79.2_{\pm 6.1}$	$81.1_{\pm 6.8}$	$96.2_{\pm 1.6}$	$90.6_{\pm 9.7}$	$69.5_{\pm 7.5}$	$80.7_{\pm 6.5}$	$87.7_{\pm 5.0}$	$78.7_{\pm 13.1}$	$200.4_{\pm 164.2}$
Severity 5 with label shift	Uncorrected	$5.0_{\pm 2.5}$	$8.6_{\pm 6.0}$	$5.7_{\pm 2.3}$	$29.2_{\pm 6.5}$	$19.8_{\pm 6.5}$	$26.9_{\pm 7.4}$	$36.5_{\pm 8.6}$	$33.1_{\pm 7.8}$	$37.9_{\pm 11.5}$	$42.9_{\pm 7.8}$	$79.4_{\pm 5.4}$	$10.4_{\pm 6.3}$	$29.0_{\pm 9.6}$	$35.5_{\pm 11.1}$	$48.1_{\pm 8.4}$	$29.8_{\pm 20.1}$	$3.3_{\pm 1.5}$
	Oracle	$87.4_{\pm 4.0}$	$93.0_{\pm 2.8}$	$91.0_{\pm 3.5}$	$92.6_{\pm 2.9}$	$88.8_{\pm 6.2}$	$91.3_{\pm 3.4}$	$89.3_{\pm 4.2}$	$90.6_{\pm 2.7}$	$89.3_{\pm 4.7}$	$88.6_{\pm 3.8}$	$91.2_{\pm 2.8}$	$90.7_{\pm 3.5}$	$90.7_{\pm 5.3}$	$92.1_{\pm 2.8}$	$91.0_{\pm 3.5}$	$90.5_{\pm 4.0}$	$349.0_{\pm 220.0}$
	LR	$18.1_{\pm 9.1}$	$23.1_{\pm 11.8}$	$21.8_{\pm 14.8}$	$58.9_{\pm 14.6}$	$39.1_{\pm 16.2}$	$46.2_{\pm 15.7}$	$61.7_{\pm 6.9}$	$49.2_{\pm 14.3}$	$55.0_{\pm 14.5}$	$58.8_{\pm 8.9}$	$83.0_{\pm 8.5}$	$33.3_{\pm 16.8}$	$41.7_{\pm 10.0}$	$53.6_{\pm 10.1}$	$66.3_{\pm 9.4}$	$47.3_{\pm 21.2}$	$27.7_{\pm 33.5}$
	Kasa et al.	$6.2_{\pm 4.0}$	$9.0_{\pm 6.1}$	$7.4_{\pm 3.2}$	$32.3_{\pm 8.4}$	$23.3_{\pm 7.4}$	$28.4_{\pm 9.1}$	$38.6_{\pm 9.0}$	$33.6_{\pm 7.7}$	$38.3_{\pm 11.6}$	$43.2_{\pm 7.6}$	$79.4_{\pm 5.4}$	$13.4_{\pm 6.8}$	$29.1_{\pm 9.4}$	$37.4_{\pm 10.7}$	$49.1_{\pm 8.4}$	$31.2_{\pm 19.9}$	$4.1_{\pm 2.3}$
	Gibbs et al.	$87.0_{\pm 2.8}$	$83.8_{\pm 3.8}$	$85.8_{\pm 3.9}$	$91.8_{\pm 1.7}$	$88.1_{\pm 2.4}$	$85.5_{\pm 3.5}$	$81.0_{\pm 3.9}$	$77.3_{\pm 5.0}$	$82.0_{\pm 4.1}$	$81.2_{\pm 4.9}$	$88.3_{\pm 3.4}$	$93.8_{\pm 2.4}$	$77.0_{\pm 3.3}$	$84.2_{\pm 6.1}$	$85.0_{\pm 3.2}$	$84.8_{\pm 5.8}$	$753.5_{\pm 100.4}$
	FF $(\min, \max)$	$47.6_{\pm 18.1}$	$50.1_{\pm 19.2}$	$48.0_{\pm 23.8}$	$86.3_{\pm 8.2}$	$71.4_{\pm 15.6}$	$74.9_{\pm 9.8}$	$78.1_{\pm 9.0}$	$73.0_{\pm 11.2}$	$76.4_{\pm 12.3}$	$81.0_{\pm 8.2}$	$95.5_{\pm 2.7}$	$68.4_{\pm 16.7}$	$64.7_{\pm 13.9}$	$79.2_{\pm 9.1}$	$84.7_{\pm 5.8}$	$72.0_{\pm 18.8}$	$127.9_{\pm 94.4}$
	WF $(\min, \max)$	$48.8_{\pm 16.0}$	$47.0_{\pm 18.0}$	$51.5_{\pm 16.9}$	$81.9_{\pm 14.7}$	$70.5_{\pm 14.4}$	$72.8_{\pm 13.4}$	$81.4_{\pm 10.6}$	$64.1_{\pm 13.2}$	$76.1_{\pm 8.7}$	$81.2_{\pm 10.7}$	$87.3_{\pm 27.2}$	$66.4_{\pm 16.6}$	$63.6_{\pm