PAC Prediction Sets Under Covariate Shift

Sangdon Park [1] **Edgar Dobriban** [2] **Insup Lee** [1] **Osbert Bastani** [1]

Abstract

Quantifying the uncertainty of predictions is an important challenge. A key problem is how to do so under covariate shift. We propose an algorithm for constructing prediction sets that satisfy probably approximately correct (PAC) guarantees under covariate shift, given labeled examples from the original distribution along with importance weights for these examples; we also consider the setting where the importance weights have bounded estimation error. We empirically validate our approach on covariate shifts constructed based on DomainNet and on ImageNet.

1. Introduction

A key challenge in machine learning is to quantify the uncertainty of a model's predictions, which is important for safety-critical applications (e.g., ensuring a robot only navigates when it has high confidence) and for human-in-the-loop settings (e.g., allowing the human to override underconfident predictions). A promising approach is via *prediction sets*, where the model predicts a set of labels instead of a single label. A benefit of this approach is that it can provide probabilistic correctness guarantees—i.e., that the predicted set contains the true label with high probability. In particular, conformal prediction provides such guarantees when the train and test distributions are equal—formally, assuming the observations are exchangeable (Vovk et al., 2005; Papadopoulos et al., 2002; Lei et al., 2015) or i.i.d. (Vovk, 2013; Park et al., 2020a; Bates et al., 2021).[1]

However, the assumption that the train and test distributions are equal often fails to hold in practice due to *covariate shift*—i.e., where the input distribution changes but the label distribution remains the same. These shifts can be both distributional (e.g., changes in color and lighting) (Hendrycks & Dietterich, 2019) or adversarial (e.g., ℓ_∞ attacks) (Szegedy et al., 2014). Under such shifts, the prediction set guarantees may no longer hold.

We consider the setting of unsupervised domain adaptation (Ben-David et al., 2007), where we are given labeled examples from the source domain, but only unlabeled examples from the target—possibly covariate shifted—domain. We propose an algorithm that constructs *probably approximately correct (PAC)* prediction sets under bounded covariate shifts (Valiant, 1984)—i.e., with high probability over the training data ("probably"), the prediction set contains the true label on future test instances ("approximately correct"). When the importance weights (IWs) are known, our algorithm uses rejection sampling (von Neumann, 1951) to construct the prediction sets. In many settings, the IWs must be estimated; thus, we extend our algorithm for when the IWs are only approximately known (i.e., we have a prediction interval around the IW rather than a point estimate) by being robust to the uncertainty in the IWs.

We evaluate our approach in two settings: (i) *rate shift*, where the model is trained on a broad domain, but deployed on a narrow domain—e.g., an autonomous car trained on both day and night images but is currently operating at night, and (ii) where the model is trained using unsupervised domain adaptation (Ganin et al., 2016). In both settings, the learned model can perform well, but we need to account for covariate shift to construct valid prediction sets. We show that our approach constructs valid prediction sets, whereas existing approaches do not.

2. Background on PAC Prediction Sets

We describe PAC prediction sets (Park et al., 2020a).

2.1. PAC Prediction Sets Algorithm

Let $x \in \mathcal{X}$ be covariates and $y \in \mathcal{Y}$ be labels. We consider a source distribution P over $\mathcal{X} \times \mathcal{Y}$ with a density function $p(x, y)$ with respect to a fixed σ-finite measure on $\mathcal{X}$.

Inputs. We are given (i) a held-out calibration set $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ of i.i.d. samples $(x_i, y_i) \sim P$, for $i \in [m] := \{1, \ldots, m\}$, and (ii) a score function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. For example, $f(x, y)$ can be a prediction for the probability

[*]Equal contribution [1]Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, USA [2]Dep.t of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, USA. Correspondence to: Sangdon Park <sangdonp@cis.upenn.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

[1]We give additional related work in Appendix A.

that y is the label for x. The score function can be arbitrary, but score functions assigning higher scores to likely labels yield smaller prediction sets.

Problem. Our goal is to construct a prediction set C that satisfies PAC guarantees. First, C is *approximately correct* if it contains the true label with a given probability—i.e.,

$$L_P(C) \coloneqq \mathop{\mathbb{E}}_{(x,y)\sim P}[\mathbb{1}(y \notin C(x))] = \mathop{\mathbb{P}}_{(x,y)\sim P}[y \notin C(x)] \leq \varepsilon,$$

where $L_P(C)$ is the expected error of C and $\varepsilon \in (0,1)$. Then, given a calibration set $S_m \sim P^m$, a prediction set C_{S_m} constructed using S_m is *probably approximately correct (PAC)* if C_{S_m} is approximately correct with a given probability over S_m—i.e.,

$$\mathop{\mathbb{P}}_{S_m\sim P^m}[L_P(C_{S_m}) \leq \varepsilon] \geq 1-\delta,$$

where $\delta \in (0,1)$. The goal is to devise an algorithm for constructing a PAC prediction set C. A large prediction set such as $C(x) = \mathcal{Y}$ is always PAC; thus, we additionally want to minimize the expected size $\mathbb{E}[S(C(x))]$ of C, where $S : 2^{\mathcal{Y}} \to \mathbb{R}_{\geq 0}$ is a size measure described below.

Algorithm. For constructing C, we first define the search space of possible prediction sets along with the size measure S. We parameterize C by a scalar τ—in particular,

$$C_\tau(x) = \{y \in \mathcal{Y} \mid f(x,y) \geq \tau\}$$

for all $\tau \in \mathcal{T} \coloneqq \mathbb{R}_{\geq 0}$. Thus τ is the threshold on $f(x,y)$ above which we include y in $C(x)$. It is readily verified that if $\tau_1 \leq \tau_2$, then $C_{\tau_2}(x) \subseteq C_{\tau_1}(x)$ for all $x \in \mathcal{X}$, so the prediction set size is monotonically decreasing in τ, as in (Gupta et al., 2021; Park et al., 2020a). Thus the expected error is also monotonically increasing in τ: $L_P(C_{\tau_1}) \leq L_P(C_{\tau_2})$. Moreover, we assume that the size measure is monotonically increasing with respect to inclusion. Thus, if $C_{\tau_2}(x) \subseteq C_{\tau_1}(x)$, then $S(C_{\tau_2}(x)) \leq S(C_{\tau_1}(x))$ for all $x \in \mathcal{X}$; thus the expected size can be minimized by maximizing τ. Then, (Park et al., 2020a) constructs C via

$$\hat{\tau} = \mathop{\arg\max}_{\tau\in\mathcal{T}} \tau \quad \text{subj. to} \quad \bar{L}_{S_m}(C_\tau) \leq k(m,\varepsilon,\delta), \quad (1)$$

where $\bar{L}_{S_m}(C) \coloneqq \sum_{(x,y)\in S_m} \mathbb{1}(y \notin C(x))$ is the unnormalized empirical error on S_m. Also,

$$k(m,\varepsilon,\delta) = \mathop{\arg\max}_{k\in\mathbb{N}\cup\{0\}} k \quad \text{subj. to} \quad F(k;m,\varepsilon) \leq \delta,$$

where $F(k;m,\varepsilon) = \sum_{i=0}^{k} \binom{m}{k}\varepsilon^i(1-\varepsilon)^{m-i}$ is the cumulative distribution function (CDF) of the binomial distribution $\text{Binom}(m,\varepsilon)$ with m trials and success probability ε. Since $\mathbb{1}(y \notin C(x))$ has a Bernoulli$(L_P(C))$ distribution, so $\bar{L}_{S_m}(C)$ has a $\text{Binom}(m, L_P(C))$ distribution. Thus, $k(m,\varepsilon,\delta)$ defines a confidence interval around the expected error $L_P(C)$ such that if $\bar{L}_{S_m}(C) \leq k(m,\varepsilon,\delta)$, then $L_P(C) \leq \varepsilon$ with probability at least $1-\delta$. Below, we formalize this intuition by drawing a connection to the Clopper-Pearson confidence interval.

2.2. Clopper-Pearson Interpretation

We interpret (1) using the Clopper-Pearson (CP) upper bound $\overline{\theta}(k;m,\delta) \in [0,1]$ (Clopper & Pearson, 1934; Cai, 2005; Park et al., 2021), which is an upper bound on the true success probability μ constructed from a sample $k \sim \text{Binom}(m,\mu)$, that holds with probability at least $1-\delta$—i.e., $\mathbb{P}_{k\sim\text{Binom}(m,\mu)}[\mu \leq \overline{\theta}(k;m,\delta)] \geq 1-\delta$. Then,

$$\overline{\theta}(k;m,\delta) \coloneqq \inf\{\theta \in [0,1] \mid F(k;m,\theta) \leq \delta\} \cup \{1\}.$$

Since $\bar{L}_{S_m}(C) \sim \text{Binom}(m, L_P(C))$, we have the following (see Appendix D.1 for the proof):

Theorem 1 *We have $\mathbb{P}_{S_m\sim P^m}[L_P(C_{\hat{\tau}}) \leq \varepsilon] \geq 1-\delta$, where letting $U_{CP}(C,S_m,\delta) \coloneqq \overline{\theta}(\bar{L}_{S_m}(C);m;\delta)$, we have*

$$\hat{\tau} = \mathop{\arg\max}_{\tau\in\mathcal{T}} \tau \quad \textit{subj. to} \quad U_{CP}(C_\tau,S_m,\delta) \leq \varepsilon. \quad (2)$$

The CP bound U_{CP} enjoys certain monotonicity properties that we will need. Intuitively, the CDF decreases as the number of observations m increases while holding the number of successes k fixed, but increases if both m and k are increased by the same amount (i.e., holding the number of failures $m-k$ fixed). In particular, we have the following (see Appendix D.2 for a proof):

Lemma 1 *We have $\overline{\theta}(k;m-1,\delta) \geq \overline{\theta}(k;m,\delta)$ and $\overline{\theta}(k-1;m-1,\delta) \leq \overline{\theta}(k;m,\delta)$.*

3. PAC Prediction Sets Under Covariate Shift

We extend the PAC prediction set algorithm described in Section 2 to the covariate shift setting.

3.1. Problem Formulation

We are given labeled training examples from the source distribution P. We want to construct prediction sets that satisfy the PAC property with respect to a shifted *target distribution* Q given only *unlabeled* examples from Q. We assume that only the covariate distribution is shifted, and the label distributions of P and Q are equal.

Preliminaries and assumptions. Let Q be the target distribution over $\mathcal{X}\times\mathcal{Y}$ with PDF $q(x,y)$, and let Q_X be its covariate distribution over $\mathcal{X}$ with PDF $q(x)$. We denote the likelihood ratio of covariate distributions by $w^*(x) \coloneqq q(x)/p(x)$, also called the *importance weight (IW)* of x. We assume *covariate shift* condition, which says the label

distributions are equal—i.e., $p(y \mid x) = q(y \mid x)$, but the covariate distributions may differ—i.e., $p(x) \neq q(x)$.

Inputs. We assume given a labeled calibration set S_m consisting of i.i.d. samples $(x_i, y_i) \sim P$ (for $i \in [m]$), and (ii) a score function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. For now, we also assume we have the true IWs $w_i \coloneqq w^*(x_i)$ for each $(x_i, y_i) \sim P$, as well as an upper bound $b \geq \max_{x \in \mathcal{X}} w^*(x)$ on the IWs. In Sections 3.3 & Appendix B, we describe how to estimate these quantities in a way that provides guarantees under certain assumptions on the IWs $w^*(x)$.

Problem. Our goal is to construct a prediction set C_{S_m} that is PAC for Q—i.e., $\mathbb{P}_{S_m \sim P^m,}\left[L_Q(C_{S_m}) \leq \varepsilon\right] \geq 1 - \delta$, where $L_Q(C) \coloneqq \mathbb{P}_{(x,y)\sim Q}[y \notin C(x)]$ is the error of C on Q, while minimizing the size of C.

3.2. Rejection Sampling Strategy

Our strategy is to use rejection sampling to convert S_m consisting of i.i.d. samples from P into a labeled calibration set T_n consisting of i.i.d. samples from Q.

Rejection sampling. Rejection sampling (von Neumann, 1951; Owen, 2013) is a technique to generate samples from a target distribution $q(x)$ that is hard to sample from, by leveraging a proposal distribution $p(x)$ that can be sampled from conveniently. It requires the IW $w^*(x)$ and an upper bound $b \geq \max_{x \in \mathcal{X}} w^*(x)$ on the IWs, and constructs a set of i.i.d. samples from q. Our algorithm takes the proposal distribution to be the source covariate distribution P_X, and the target distribution to be our target covariate distribution Q_X. Then, given m samples S_m from the source distribution, rejection sampling outputs the samples

$$T_N(S_m, V, w, b) \coloneqq \left\{(x_i, y_i) \in S_m \;\middle|\; V_i \leq \frac{w_i}{b}\right\}$$

where $U \coloneqq \text{Uniform}([0,1])$, and $V_i \sim U$ i.i.d. for all $i \in [m]$. The expected number of samples $\mathbb{E}[N]$ is m/b; thus, rejection sampling is more effective when the proposal distribution is similar to the target.

Rejection sampling Clopper-Pearson (RSCP) bound. Once T_n has been constructed, we use the CP bound to construct a PAC prediction set C for Q. Let $\sigma_i \coloneqq \mathbb{1}(V_i \leq w_i/b)$ indicate whether example $(x_i, y_i) \in S_m$ is accepted—i.e., $T_N(S_m, V, w, b) = \{(x_i, y_i) \in S_m \mid \sigma_i = 1\}$ and $|T_N(S_m, V, w, b)| = \mathbf{1}^T\sigma$. Given $T_n = T_N(S_m, V, w, b)$, the CP bound $U_{\text{CP}}(C, T_n, \delta)$ bounds the error on Q—i.e., we have the following (see Appendix D.3 for the proof):

Theorem 2 *We have* $\mathbb{P}_{S_m \sim P^m, V \sim U^m}\left[L_Q(C_{\hat{\tau}}) \leq \varepsilon\right] \geq 1 - \delta$, *where*

$$\hat{\tau} = \arg\max_{\tau \in \mathcal{T}} \tau \quad \textit{subj. to} \quad U_{\textit{RSCP}}(C_\tau, S_m, V, w, b, \delta) \leq \varepsilon$$
$$U_{\textit{RSCP}}(C, S_m, V, w, b, \delta) \coloneqq U_{\textit{CP}}(C, T_N(S_m, V, w, b), \delta).$$

3.3. Approximate Importance Weights

So far, we have assumed that the true IWs $w^*(x)$ are known, but it often must be estimated. We relax this assumption to only needing an uncertainty set of possible importance weights, allowing us to handle estimation error.

Assumption. Letting $w^* = (w_1, ..., w_m) \in \mathbb{R}^m$ be the vector of true IWs $w_i \coloneqq w^*(x_i)$ (for $(x_i, y_i) \in S_m$), we assume given an uncertainty set $\mathcal{W} \subseteq \mathbb{R}^m$ that contains w^* with high probability—i.e., $\mathbb{P}_A\left[w^* \in \mathcal{W}\right] \geq 1 - \delta_w$, where $\delta_w \in (0,1)$ and A is the randomness in the dataset and algorithm used to construct $\mathcal{W}$. We assume $\mathcal{W}$ has form

$$\mathcal{W} \coloneqq \left\{w \;\middle|\; \forall i \in [m]\ ,\ \underline{w}_i \leq w_i \leq \overline{w}_i\ ,\ \underline{c} \leq \mathbf{1}^T w \leq \overline{c}\right\},$$

for some $\underline{w}_i$, $\overline{w}_i$, $\underline{c}$, and $\overline{c}$. We can expect an empirical constraint on $\mathbf{1}^T w$ in the confidence set because of the population constraint $\mathbb{E}_{x\sim P}[w^*(x)] = 1$.

Robust Clopper-Pearson bound. To construct a PAC prediction set C for Q, it suffices to bound the worst-case error over IWs $w \in \mathcal{W}$—i.e., we have the following (see Appendix D.4 for the proof):

Theorem 3 *We have* $\mathbb{P}_{S_m \sim P^m, V \sim U^m, A}[L_Q(C_{\hat{\tau}}) \leq \varepsilon] \geq 1 - \delta_C - \delta_w$, *where*

$$\hat{\tau} = \arg\max_{\tau \in \mathcal{T}} \tau$$
$$\textit{subj. to} \ \max_{w \in \mathcal{W}} U_{\textit{RSCP}}(C_\tau, S_m, V, w, b, \delta_C) \leq \varepsilon. \quad (3)$$

A key challenge applying Theorem 3 is that solving the maximum over $w \in \mathcal{W}$ can be intractable. We propose a simple greedy algorithm that upper bounds the maximum.

Greedy algorithm for U_{RSCP}. The RSCP bound U_{RSCP} satisfies certain monotonicity properties that enable us to efficiently compute an upper bound to the maximum in (3). In particular, if C makes an error on (x_i, y_i) (i.e., $y_i \notin C(x_i)$), then U_{RSCP} is monotonically non-decreasing in $w_i = w^*(x_i)$; intuitively, this holds since a larger w_i increases the probability that (x_i, y_i) is included in $T_N(S_m, V, w, b)$, which increases the empirical error $\bar{L}_{T_N(S_m,V,w,b)}(C)$. Conversely, if C does not make an error on (x_i, y_i) (i.e., $y_i \in C(x_i)$), U_{RSCP} is non-increasing in w_i. More formally, we have the following result (see Appendix D.5 for a proof):

Lemma 2 *For any* $i \in [m]$, $U_{\textit{RSCP}}(C, S_m, V, w, b, \delta)$ *is monotonically non-decreasing in* w_i *if* $y_i \notin C(x_i)$, *and monotonically non-increasing in* w_i *if* $y_i \in C(x_i)$.

We leverage the monotonicity of U_{RSCP} for our greedy algorithm. In particular, the choice

$$\hat{w}_i = \begin{cases} \overline{w}_i & \text{if } y_i \notin C(x_i) \\ \underline{w}_i & \text{if } y_i \in C(x_i) \end{cases} \qquad (\forall i \in [m])$$

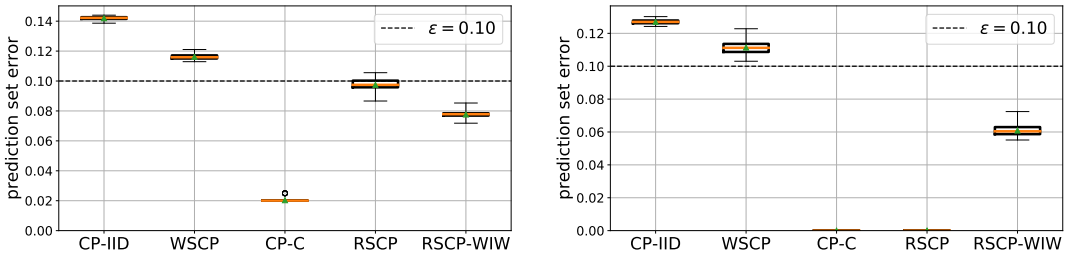


Figure 1: Error under natural rate shift by DomainNet for All → Sketch (left), and ImageNet-C synthetic perturbations to ImageNet (right), over 100 random trials, with $m = 50,000$ (for DomainNet) and $m = 20,000$ (for ImageNet), $\varepsilon = 0.1$, and $\delta = 10^{-5}$.

forms an upper bound on the value of the maximum over $w \in \mathcal{W}$ in (3). More precisely, it is the maximizer if we remove the mass constraint—i.e., letting $\mathcal{W}' := \{w \in \mathbb{R}^m \mid \underline{w}_i \leq w_i \leq \overline{w}_i\}$, then

$$\begin{aligned}\max_{w\in\mathcal{W}} & U_{\text{RSCP}}(C, S_m, V, w, b, \delta) \\ &\leq \max_{w\in\mathcal{W}'} U_{\text{RSCP}}(C, S_m, V, w, b, \delta) \\ &= U_{\text{RSCP}}(C, S_m, V, \hat{w}, b, \delta). \end{aligned} \tag{4}$$

Thus, we have the following, which follows by (4) and the same argument as the proof of Theorem 3:

Theorem 4 *We have* $\mathbb{P}_{S_m\sim P^m, V\sim U^m, A}[L_Q(C_{\hat{\tau}}) \leq \varepsilon] \geq 1 - \delta_C - \delta_w$, *where*

$$\hat{\tau} = \arg\max_{\tau\in\mathcal{T}} \tau \quad \textit{subj. to} \quad U_{RSCP}(C_\tau, S_m, V, \hat{w}, b, \delta_C) \leq \varepsilon.$$

Importance weight estimation. Finally, a number of approaches have been proposed for estimating the importance weights (IWs); our approach is compatible with any of these strategies if they can be modified to provide uncertainty estimates of the IWs. We build on a cluster-based approach (Cortes et al., 2008); see Appendix B for details.

4. Experiments

We demonstrate the efficacy of our proposed approach on natural rate shifts (where the score function f is specialized to a subdomain) and on unsupervised domain adaptation (where the score function is trained using unlabeled data from the target domain). We summarize our results here; see Appendix C for details.

Baselines. We compare our approach in Theorem 4 (**RSCP-WIW**) with Park et al. (2020a) (**CP-IID**), a conservative approach described in Appendix E.3 (**CP-C**), and weighted split conformal prediction (Tibshirani et al., 2019) (**WSCP**).

Metrics. We report the prediction set error on a held-out test set—i.e., the fraction of test instances for which $y \notin C(x)$; we show the prediction set sizes in Appendix F.

Natural rate shift. We consider the setting where the model is trained on data from a variety of domains, but is then deployed on a specific domain; we call such a shift *rate shift*. In this setting, the model should still perform reasonably well since the target domain is a subset of the source domain, but the covariate shift can nevertheless invalidate prediction set guarantees. We evaluate our approach on DomainNet (Peng et al., 2019), where the source distribution is the whole dataset (from 6 domains) and the target distribution is a single domain. We use $\varepsilon = 0.1$ and $\delta = 10^{-5}$.

In Figure 5 (left), we show results for All → Sketch; we give more results in Appendix C. As can be seen, the prediction set error of our approach (RSCP-WIW) does not violate the error constraint $\varepsilon = 0.1$, while all other approaches (except for CP-C) violate it for at least one of the shifts. While CP-C satisfies the desired bound, it is overly conservative. Its prediction sets are significantly larger than necessary, making it less useful for uncertainty quantification.

Unsupervised domain adaptation. Next, we evaluate our approach applied with a score function f trained using unsupervised domain adaptation (Ganin et al., 2016). We consider ImageNet-C (Hendrycks & Dietterich, 2019), which modifies the original ImageNet dataset (Russakovsky et al., 2015) using synthetic perturbations. We use ImageNet as the source, and ImageNet-C as the target. We show results in Figure 3a (right). As can be seen, the error of our approach (RSCP-WIW) is below the desired level. RSCP performs poorly, likely due to the uncalibrated point IW estimation; using Theorem 5 to rescale the importance weights mitigates these issues, though accounting for uncertainty in the IWs is necessary for achieving the desired error rate.

5. Conclusion

We propose a novel algorithm to build a PAC prediction set under covariate shift; it leverages rejection sampling and Clopper-Pearson binomial interval, assuming the true IW is known; this assumption is relaxed by rigorous IW estimation, where the smoothness assumption on the covaraite distributions is required. We demonstrate the efficacy of the proposed approaches over natural, synthetic and adversarial covariate shifts, curated by DomainNet and ImageNet.

References

Balasubramanian, V., Ho, S.-S., and Vovk, V. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference, 2020.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. I. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pp. 81–88. ACM, 2007.

Cai, T. T. One-sided confidence intervals in discrete distributions. *Journal of Statistical planning and inference*, 131(1):63–88, 2005.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Robust validation: Confident predictions even when distributions shift, 2020.

Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pp. 38–53. Springer, 2008.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Ding, G. W., Wang, L., and Jin, X. AdverTorch v0.1: An adversarial robustness toolbox based on pytorch. *arXiv preprint arXiv:1902.07623*, 2019.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL `http://jmlr.org/papers/v17/15-239.html`.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. K. Nested conformal prediction and quantile out-of-bag ensemble methods, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

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

Kanamori, T., Hido, S., and Sugiyama, M. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.

Kuleshov, V. and Liang, P. S. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*, pp. 3474–3482, 2015.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pp. 3792–3803, 2019.

Lei, J., Rinaldo, A., and Wasserman, L. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1):29–43, 2015.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Malik, A., Kuleshov, V., Song, J., Nemer, D., Seymour, H., and Ermon, S. Calibrated model-based deep reinforcement learning. In *International Conference on Machine Learning*, pp. 4314–4323, 2019.

Murphy, A. H. Scalar and vector partitions of the probability score: Part i. two-state situation. *Journal of Applied Meteorology*, 11(2):273–282, 1972.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, Nov 2010. ISSN 1557-9654. doi: 10.1109/tit.2010.2068870.

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

Park, S., Bastani, O., Matni, N., and Lee, I. Pac confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020a.

Park, S., Bastani, O., Weimer, J., and Lee, I. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020b.

Park, S., Li, S., Lee, I., and Bastani, O. PAC confidence predictions for deep neural network classifiers. In *International Conference on Learning Representations*, 2021.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 1999.

Podkopaev, A. and Ramdas, A. Distribution-free uncertainty quantification for classification under label shift. *arXiv preprint arXiv:2103.03323*, 2021.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. J. With malice towards none: Assessing uncertainty via equalized coverage, 2019.

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(3): 211–252, 2015.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32: 2530–2540, 2019.

Valiant, L. G. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

von Neumann, J. Various techniques used in connection with random digits. In Householder, A. S., Forsythe, G. E., and Germond, H. H. (eds.), *Monte Carlo Method*, volume 12 of *National Bureau of Standards Applied Mathematics Series*, chapter 13, pp. 36–38. US Government Printing Office, Washington, DC, 1951.

Vovk, V. Conditional validity of inductive conformal predictors. *Machine learning*, 92(2-3):349–376, 2013.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Wang, X., Long, M., Wang, J., and Jordan, M. Transferable calibration with lower bias and variance in domain adaptation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19212–19223. Curran Associates, Inc., 2020.

Wilks, S. S. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *In Proceedings of the Eighteenth International Conference on Machine Learning*. Citeseer, 2001.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699. ACM, 2002.

A. Related Work

The goal of conformal prediction (Vovk et al., 2005; Balasubramanian et al., 2014) is to construct models that predict sets of labels designed to include the true label with high probability (instead of individual labels). We build on *inductive* (or *split*) conformal prediction (ICP) (Papadopoulos et al., 2002; Lei et al., 2015), where the training set is split into (i) a *proper training set* used to train a traditional predictive model, and (ii) a *calibration set* used to construct the prediction sets (Vovk, 2013). Then, the general approach is (i) to train a model $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that outputs (heuristic) *conformity scores*, and then (ii) to choose a threshold $\tau \in \mathbb{R}$ that satisfies a correctness guarantee, where the corresponding prediction sets are $C(x) = \{y \in \mathcal{Y} \mid f(x, y) \geq \tau\}$.

Several kinds of correctness guarantees have been considered. One possibility is *conditional* (Barber et al., 2020) or *object conditional* (Vovk, 2013) validity, which ensures correctness for *all* future covariates, with high probability only over the conditional label distribution $p(y \mid x)$. This guarantee is very strong and hard to ensure in practice. A weaker notion is *approximate* (Barber et al., 2020) or *group* (Romano et al., 2019) conditional validity, which ensures correctness with high probability over $p(y \mid x)$ as well as some distribution $p(x)$ over a subgroup. Finally, *unconditional validity* ensures only correctness over $p(x, y)$. We focus on unconditional validity, though our approach extends straightforwardly to group conditional validity.

A separate issue is how to condition on the calibration set Z. One approach is to simply include this randomness with the above—e.g., for unconditional validity, the high-probability guarantee is now over $p(x, y, Z)$; we refer to this strategy as *fully unconditional validity*. The guarantee we consider uses a separate confidence level for the training data, which is called *training conditional guarantee* (Vovk, 2013); this correctness notion is equivalent to a PAC correctness guarantee (Park et al., 2020a), and is also related to the notion of tolerance regions from statistics (Wilks, 1941). We build on (Park et al., 2020a), which formulates the problem of choosing τ as learning a binary classifier where the input and parameter spaces are both one-dimensional; thus, the correctness guarantee corresponds to a PAC generalization bound. This approach is widely applicable since it can work with a variety of objectives (Bates et al., 2021).

Recent work has extended ICP to a setting with covariate shift (Tibshirani et al., 2019); similarly, (Podkopaev & Ramdas, 2021) considers conformal prediction under label shift (Lipton et al., 2018)—i.e., assuming the conditional probabilities $p(x \mid y)$ do not change. These approaches assume that the true importance weights (IWs) are known, whereas our approach considers uncertainty in the IWs. In addition, they are focused on unconditional validity, whereas we obtain PAC prediction sets. In addition, (Cauchois et al., 2020) designs confidence sets that are robust to *all* distribution shifts with bounded f-divergence; in contrast, we consider the unsupervised learning setting where we have examples from the target distribution.

Finally, an alternative way to quantify uncertainty is *calibrated prediction* (Murphy, 1972; DeGroot & Fienberg, 1983; Guo et al., 2017), which aims to ensure that among instances with a predicted confidence p, the model is correct a fraction p of the time. Techniques have been proposed to re-scale predicted confidences to improve calibration (Platt, 1999; Guo et al., 2017; Zadrozny & Elkan, 2001; 2002; Kuleshov & Liang, 2015; Kuleshov et al., 2018; Malik et al., 2019); including ones with theoretical guarantees (Kumar et al., 2019; Park et al., 2021) and ones that handle covariate shift (Park et al., 2020b; Wang et al., 2020). These approaches provide a qualitatively different form of uncertainty quantification.

B. Importance Weight Estimation

In general, to estimate the importance weights (IWs), some assumptions on their structure are required. A number of approaches have been proposed, with varying guarantees under different assumptions (Kanamori et al., 2009; Cortes et al., 2008; Nguyen et al., 2010; Lipton et al., 2018). We use a cluster-based approach (Cortes et al., 2008); our approach is compatible with any of these strategies if they can be modified to provide uncertainty estimates of the IWs.

In our approach, given a partition $\mathcal{X} = \bigcup_{j=1}^{K} B_j$ into bins, we can estimate the IWs based on the fractions of source and target samples in each bin. If the partition is sufficiently fine, then we can obtain confidence intervals around the estimated IWs with finite-sample guarantees. However, this strategy requires the number of bins in the partition to be exponential in the dimension of $\mathcal{X}$. Thus, in practice, we use a heuristic to construct the partition. We describe the cluster-based approach and our partition construction heuristic below.

Cluster-based approach. We assume given independent unlabeled calibration sets S_m^X and T_n^X, where S_m^X consists of i.i.d. samples $x_i \sim p(x)$ for $i \in [m]$, and T_n^X consists of i.i.d. samples $x_i \sim q(x)$ for $i \in [n]$, respectively.[2] Roughly speaking,

[2] We can use the same calibration set to construct IWs and prediction sets due to the union bound in Theorem 3.

the cluster-based strategy estimates the average IW in each bin B_j; assuming p and q are roughly constant in each bin, these accurately estimate the true IWs. Let $j(x)$ be the bin containing x (i.e., $x \in B_{j(x)}$), and let

$$p_B(x) \coloneqq p_{j(x)} \text{ s.t. } p_j = \int_{B_j} p(x')\,\mathrm{d}x' \quad \text{and}$$
$$q_B(x) \coloneqq q_{j(x)} \text{ s.t. } q_j = \int_{B_j} q(x')\,\mathrm{d}x'$$

be the (unnormalized) approximations of the densities p and q, respectively, that are constant on each bin. We assume that p_B and q_B are accurate approximations:

Assumption 1 Given $E \in \mathbb{R}_{\geq 0}$, the partition satisfies

$$\int_{B_j} |p(x) - p(x')|\mathrm{d}x' \leq E \quad \text{and}$$
$$\int_{B_j} |q(x) - q(x')|\mathrm{d}x' \leq E \qquad (\forall j \in [K]). \tag{5}$$

Thus, p and q are roughly constant on the partitions. In general, (5) can hold for any $E \in \mathbb{R}_{>0}$ if p and q are Lipschitz continuous and each B_j is sufficiently small. Then, it can be verified that

$$|v(x) \cdot p(x) - p_B(x)| \leq E \qquad \text{and}$$
$$|v(x) \cdot q(x) - q_B(x)| \leq E \qquad (\forall x \in \mathcal{X}), \tag{6}$$

where $v(x) = v_{j(x)}$ and $v_j = \int_{B_j} \mathrm{d}x'$ is the volume of bin B_j. Next, we have the following empirical estimates of p_B and q_B, respectively:

$$\hat{p}_B(x) \coloneqq \hat{p}_{j(x)} \text{ s.t. } \hat{p}_j = \frac{1}{m} \sum_{x' \in S_m^X} \mathbb{1}\left(x' \in B_j\right) \text{ and}$$
$$\hat{q}_B(x) \coloneqq \hat{q}_{j(x)} \text{ s.t. } \hat{q}_j = \frac{1}{n} \sum_{x' \in T_n^X} \mathbb{1}\left(x' \in B_j\right).$$

Now, $\mathbb{1}(x' \in B_j)$ has distribution Bernoulli(p_j) when $x \sim P$, so $m \cdot \hat{p}_j$ has distribution Binom(m, p_j). Thus, p_j is contained in a Clopper-Pearson interval around $\hat{p}_j$ with high probability—i.e.,

$$\underline{\theta}(m \cdot \hat{p}_j; m, \delta') \leq p_j \leq \overline{\theta}(m \cdot \hat{p}_j; m, \delta') \tag{7}$$

with probability at least $1 - \delta'$ with respect to the samples S_m^X. Here, $\underline{\theta}$ is the Clopper-Pearson lower bound corresponding to the Clopper-Pearson upper bound defined in Section 2.2—i.e., $\mathbb{P}_{k \sim \text{Binom}(m,\mu)}[\mu \geq \underline{\theta}(k; m, \delta)] \geq 1 - \delta$. Combining (6) and (7), we have the following result:

Theorem 5 *Letting $\delta' = \delta_w/(2K)$ and $[v]^+ \coloneqq \max\{0, v\}$ for all $v \in \mathbb{R}$, we have*

$$\frac{\left[\underline{\theta}\left(n \cdot \hat{q}_B(x); n, \delta'\right) - E\right]^+}{\overline{\theta}\left(m \cdot \hat{p}_B(x); m, \delta'\right) + E} \leq w^*(x)$$
$$\leq \frac{\overline{\theta}\left(n \cdot \hat{q}_B(x), n, \delta'\right) + E}{\left[\underline{\theta}\left(m \cdot \hat{p}_B(x), m, \delta'\right) - E\right]^+} \qquad (\forall x \in \mathcal{X})$$

with probability at least $1 - \delta_w$ over S_m^X and T_n^X.

We use these upper and lower bounds on the IWs as the inputs $\underline{w}_i$ and $\overline{w}_i$ to our algorithm, and use $b = \max_i \overline{w}_i$ as the maximum IW.

Partition construction heuristic. In general, exponentially many bins are needed to guarantee Assumption 1. Instead, we consider an intuitive heuristic for constructing these bins, so that the importance weights $w(x)$—rather than the density

functions $p(x)$ and $q(x)$ individually—are roughly constant on each bin. A standard heuristic for estimating IWs is to train a probabilistic classifier $\hat{s}(y \mid x)$ to distinguish source and target examples, and then use these probabilities to construct the IWs. In particular, define the distribution

$$s^*(x, y) = \frac{1}{2}p(x) \cdot \mathbb{1}(y = 1) + \frac{1}{2}q(x) \cdot \mathbb{1}(y = 0).$$

Then, letting $s^*(y \mid x)$ be the marginal distribution, we have $w^*(x) = 1/s^*(y = 1 \mid x) - 1$ (Bickel et al., 2007). Thus, we train $\hat{s}(y \mid x) \approx s^*(y \mid x)$ and construct bins according to $\hat{w}(x) = 1/\hat{s}(y = 1 \mid x) - 1$—i.e.,

$$B_j = \{x \in \mathcal{X} \mid \hat{w}(x) \in [w_j, w_{j+1})\},$$

where $0 = w_1 \leq w_2 \leq ... \leq w_{K+1} = \infty$. Finally, we describe how to train $\hat{s}$. For simplicity, assume that $m = n$; then, the set

$$R^X_{2m} = \{(x, 1) \mid x \in S^X_m\} \cup \{(x, 0) \mid x \in T^X_n\}$$

consists of i.i.d. samples $(x, y) \sim s^*(x, y)$. Thus, we can train $\hat{s}$ on R^X_m using supervised learning. In practice, the corresponding IW estimates $\hat{w}(x)$ can be inaccurate partly since $\hat{w}$ is likely overfit to R^X_{2m}, which is why re-estimating the IWs in each bin according to Theorem 5 remains necessary.

C. Experiments

We demonstrate the efficacy of our proposed approach on natural rate shifts (where the score function f is specialized to a subdomain) and on unsupervised domain adaptation (where the score function is trained using unlabeled data from the target domain) [3].

C.1. Experimental Setup

Models. For each source-target distribution pair, we split the labeled source data into train and calibration sets, and split the unlabeled target data into train and calibration sets; we use a separate labeled test set of target data for evaluation. For each shift from source to target, we use a deep neural network (DNN) score function f based on ResNet101 (He et al., 2016), which we train using unsupervised domain adaptation based on the source and target training sets. See Appendix E for more details.

Prediction set construction. To construct our prediction sets, we first estimate IWs by training a probabilistic classifier $\hat{s}$ using the source and target training sets. Next, we use $\hat{s}$ to construct heuristic IWs $\hat{w}(x) = 1/\hat{s}(y = 1|x) - 1$. Then, we estimate the lower and upper bound of the true IWs by using Theorem 5 with $E = 0.01$ and $K = 10$ bins (chosen to contain equal numbers of heuristic IWs), where we compute the lower and upper Clopper-Pearson interval using the source and target calibration sets. Furthermore, given a confidence level δ, we use $\delta_C = \delta_w = \delta/2$. Finally, we evaluate the prediction set error on the labeled target test set.

Baselines. We compare the proposed method in Theorem 4 (**RSCP-WIW**) with the following:

- **CP-IID:** The algorithm in Theorem 1 based on (Park et al., 2020a), which makes the i.i.d. assumption.
- **CP-C:** A Clopper-Pearson method accounting for covariate shift by a conservative upper bound on the empirical loss (see Appendix E.3 for details), resulting in the algorithm

$$\hat{\tau} = \underset{\tau \in \mathcal{T}}{\arg\max}\, \tau \quad \text{subj. to} \quad U_{\text{CP}}(C_\tau, S_m, \delta) \leq \varepsilon/b.$$

- **WSCP:** Weighted split conformal prediction, proposed in (Tibshirani et al., 2019). This approach provides correctness guarantees only for a single future test instance—i.e., it includes an ε confidence level (denoted α in their paper) but not δ. Furthermore, it assumes the true IWs are known; following their experiments, we use the heuristic IWs constructed using $\hat{s}$.
- **RSCP:** The RSCP prediction set in Theorem 2 with the heuristic IWs constructed using $\hat{s}$.

Metrics. We measure performance via the prediction set error on the held-out labeled test set—i.e., the fraction of test instances for which $y \notin C(x)$; we discuss the prediction set sizes in Appendix F.

[3] We use the thrid-part code (Hendrycks & Dietterich, 2019) under Apache-2.0 license and (Ding et al., 2019) under LGPL.

C.2. Natural Rate Shift via DomainNet

First, we consider the setting where the model is trained on data from a variety of domains, but is then deployed on a specific domain; we call such a shift *rate shift*. For instance, a self-driving car may be trained on both day and night images, but tested during the night time. In this setting, the model should still perform reasonably well since the target domain is a subset of the source domain, but the covariate shift can nevertheless invalidate prediction set guarantees.

Datasets. DomainNet (Peng et al., 2019) consists of images of 345 classes from six domains (sketch, clipart, painting, quickdraw, real, and infograph). We split the dataset into 409,832 training, 88,371 calibration, and 88,372 test images; we use the entire dataset as the source and each domain as a target.

Results. We set the desired error $\varepsilon = 0.1$ and confidence level $\delta = 10^{-5}$. As can be seen in Figure 5, the prediction set error of our approach (RSCP-WIW) does not violate the error constraint $\varepsilon = 0.1$, while all other approaches (except for CP-C) violate it for at least one of the shifts. While CP-C satisfies the desired bound, it is overly conservative. Its prediction sets are significantly larger than necessary, making it less useful for uncertainty quantification. In Figure 5f, most methods output conservative prediction sets; we believe the reason is that in this case, labeled examples with smaller $f(x, y)$ have larger IWs; see Appendix F.3 for discussion.

Ablation on IW uncertainties. We consider a variant RSCP-MIW of RSCP-WIW that ignores the worst-case IWs as in Theorem 3; instead, it rescales the importance weights in each bin via a point estimate—i.e., Theorem 5 without the Clopper-Pearson interval (i.e., $m = n = \infty$) and with $E = 0$. In Figure 2g, we show results on the shift to the sketch domain. our approach satisfies the PAC guarantee but this version does not—in fact, it performs even worse than RSCP, which uses the non-rescaled importance weights from the probabilistic classifier $\hat{s}$.

Varying E. The parameter E in Assumption 1 bounds the quality of our estimates of $p(x)$ and $q(x)$; since these errors cannot be conveniently measured, we have chosen it heuristically as a hyperparameter. In Figure 2h, we show the error of RSCP-WIW as a function of E for the shift to the sketch domain. As E becomes larger, the prediction sets become more conservative while still satisfying the PAC guarantee. Intriguingly, the optimal case is $E = 0$, since RSCP-WIW still satisfies the PAC guarantee, suggesting that our importance weight estimates may not be too inaccurate.

C.3. Unsupervised Domain Adaptation via Synthethic and Adversarial Shifts on ImageNet

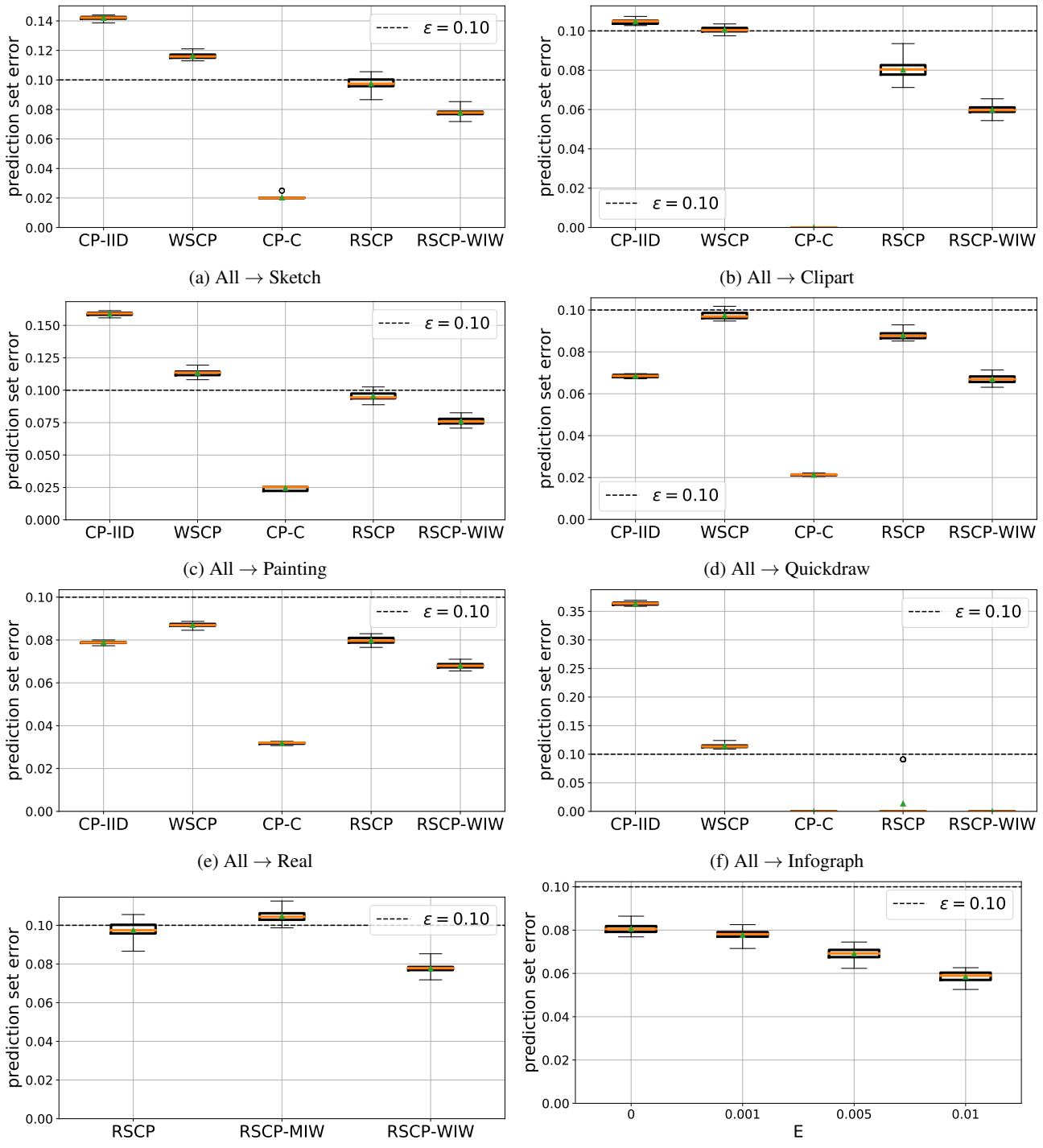
Next, we evaluate our approach applied with a score function f trained using unsupervised domain adaptation (Ganin et al., 2016) on synthetic and adversarial distribution shifts.

Datasets. For synthetic shifts, we consider ImageNet-C (Hendrycks & Dietterich, 2019), which modifies the original ImageNet dataset (Russakovsky et al., 2015) using 15 synthetic perturbations with 5 severity levels. We use 13 perturbations, omitting "snow" and "glass blur", which are computationally expensive to run. We consider the original ImageNet dataset as the source, and the all synthetic perturbations on all of ImageNet (denoted ImageNet-C13) as the target. We split ImageNet into 1.2M training, 25K calibration, and 25K test images, and ImageNet-C13 into 83M training, 1.6M calibration, and 1.6M test images.

For adversarial shifts, we generate adversarial examples for ImageNet using the PGD attack (Madry et al., 2017) with 0.01 ℓ_∞-norm perturbations with respect to a pretrained ResNet101. We consider the original ImageNet as the source and the adversarially perturbed ImageNet as the target.

Results. We show results for the synthetic shift in Figure 3a, and for the adversarial shift in Figure 3b. As can be seen, the error of our approach (RSCP-WIW) is below the desired level. RSCP performs poorly, likely due to the uncalibrated point IW estimation—we find that using Theorem 5 to rescale the importance weights mitigates these issues, though accounting for uncertainty in the IWs is necessary for achieving the desired error rate; see Appendix F.4 for details.

Interpretation. As an uncertainty quantifier, prediction sets provide a guarantee for interpreting label predictions. For adversarial shifts, the target classification error of the source-trained ResNet101 and the domain-adapted ResNet101 is 99.97% and 28.05%, respectively. Domain adaptation could significantly decrease average error rate, but label predictions do not have guarantees. However, our prediction set controls the prediction set error rate as specified by ε. As shown in Figure 3b, the our prediction set function outputs a prediction set for a given example that includes the true label at least 90% of the time. Thus, downstream modules can rely on this guarantee for further decision-making.



(a) All → Sketch

(b) All → Clipart

(c) All → Painting

(d) All → Quickdraw

(e) All → Real

(f) All → Infograph

(g) Ablation study

(h) Varying E

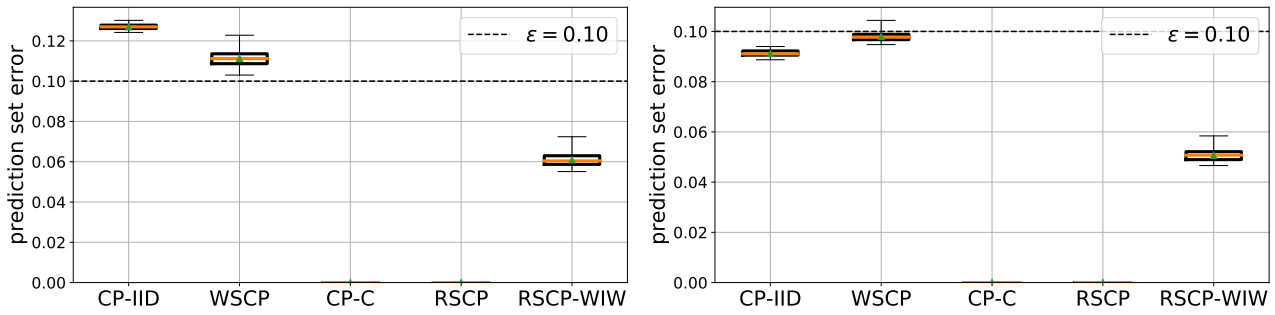
Figure 2: Error under natural rate shift by DomainNet (over 100 random trials). (a-f) We consider six different natural shifts. (g) Ablation study on the effect of heuristic and rigorous calibration. (h) Error by varying E. Parameters are $m = 50,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$.

D. Proofs

D.1. Proof of Theorem 1

First, note that the constraint in (1) implies $F(\bar{L}_{S_m}(C_\tau); m, \varepsilon) \leq \delta$; conversely, any value of τ satisfying $F(\bar{L}_{S_m}(C_\tau); m, \varepsilon) \leq \delta$ also satisfies $\bar{L}_{S_m}(C_\tau) \leq k(m, \varepsilon, \delta)$. Thus, we can rewrite (1) as

$$\hat{\tau} = \underset{\tau \in \mathcal{T}}{\arg\max}\ \tau \quad \text{subj. to} \quad F(\bar{L}_{S_m}(C_\tau); m, \varepsilon) \leq \delta.$$



(a) ImageNet-C13

(b) ImageNet-PGD

Figure 3: Error on ImageNet shifted by synthetic and adversarial perturbations (over 100 random trials). Parameters are $m = 20,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$.

On the other hand, if τ satisfies (2), then by definition of $U_{\text{CP}}(C_\tau, S_m, \delta) = \overline{\theta}(k; m, \delta)$, we have

$$\inf\{\theta \in [0,1] \mid F(\bar{L}_{S_m}(C_\tau); m, \theta) \leq \delta\} \cup \{1\} \leq \varepsilon,$$

which implies that $F(\bar{L}_{S_m}(C_\tau); m, \varepsilon) \leq \delta$ (since the infimum is obtained within the set since the binomial CDF F is continuous in ε). Conversely, any value of τ satisfying $F(\bar{L}_{S_m}(C_\tau); m, \varepsilon) \leq \delta$ also satisfies $U_{\text{CP}}(C_\tau, S_m, \delta) \leq \varepsilon$. Thus, we can rewrite (2) as

$$\hat{\tau} = \arg\max_{\tau \in \mathcal{T}} \tau \quad \text{subj. to} \quad F(\bar{L}_{S_m}(C_\tau); m, \varepsilon) \leq \delta.$$

Thus, (1) and (2) are equal. Thus, the claim follows from Theorem 1 of (Park et al., 2020a). □

D.2. Proof of Lemma 1

Recall that $F(k; m, \theta)$ is the cumulative distribution function of a binomial distribution Binom (m, θ), or equivalently of the random variable $\sum_{i=1}^{m} X_i$, where $X_i \sim$ Bernoulli(θ) are i.i.d.

Decreasing case. If $k \leq m - 1$, then we have

$$\sum_{i=1}^{m} X_i \leq k \Rightarrow \sum_{i=1}^{m-1} X_i \leq k,$$

hence

$$\mathbb{P}\left[\sum_{i=1}^{m} X_i \leq k\right] \subseteq \mathbb{P}\left[\sum_{i=1}^{m-1} X_i \leq k\right],$$

so $F(k; m, \theta) \leq F(k; m-1, \theta)$.

Then, we have

$$\begin{aligned}\overline{\theta}(k; m, \delta) &:= \inf\{\theta \in [0,1] \mid F(k; m, \theta) \leq \delta\} \cup \{1\} \\ &\leq \inf\{\theta \in [0,1] \mid F(k; m-1, \theta) \leq \delta\} \cup \{1\} \\ &=: \overline{\theta}(k; m-1, \delta),\end{aligned}$$

thus $\overline{\theta}$ is monotonically non-increasing in m.

Increasing case. We have

$$\sum_{i=1}^{m-1} X_i \leq k - 1 \Rightarrow \sum_{i=1}^{m} X_i \leq k,$$

hence

$$\mathbb{P}\left[\sum_{i=1}^{m-1} X_i \leq k-1\right] \subseteq \mathbb{P}\left[\sum_{i=1}^{m} X_i \leq k\right],$$

so $F(k-1; m-1, \theta) \leq F(k; m, \theta)$.

Then, we have

$$\begin{aligned}\overline{\theta}(k; m, \delta) &\coloneqq \inf\{\theta \in [0,1] \mid F(k; m, \theta) \leq \delta\} \cup \{1\} \\ &\geq \inf\{\theta \in [0,1] \mid F(k-1; m-1, \theta) \leq \delta\} \cup \{1\} \\ &=: \overline{\theta}(k-1; m-1, \delta),\end{aligned}$$

thus $\overline{\theta}$ is monotonically jointly non-decreasing in (m, k).

D.3. Proof of Theorem 2

The rejection sampling prediction set consists of two steps: (i) generate target samples, using source samples S_m, importance weights w, and an upper bound on their maximum value b, and (ii) construct the Clopper-Pearson prediction set using the generated target samples.

From rejection sampling, we choose $N \coloneqq \sum_{i=1}^{m} \sigma_i$ samples from S_m, denoting them by T_N; here, $N \sim \text{Binom}\,(m, 1/b)$, and $1/b$ is the acceptance probability (von Neumann, 1951)—i.e.,

$$\mathbb{P}\left[V' \leq \frac{w(X)}{b}\right] = \frac{1}{b},$$

where $V' \sim \text{Uniform}([0,1])$. The samples in T_N are independent and identically distributed, conditionally on the random number N of samples being equal to any fixed value n. The reason is that one can view the rejection sampling algorithm proceeding in stages, iterating through the samples one by one. The first stage starts at the very beginning, and then each stage ends when a datapoint is accepted, followed by starting a new stage at the next datapoint. The last stage ends at the last datapoint.

Based only on the source samples observed in one stage, rejection sampling produces a sample from the target distribution. Thus, within each stage, we produce one sample from the target distribution, and because each stage is independent of all the other ones, conditionally on any number of stages reached, our produced target samples are iid. Thus, we can use the Clopper-Pearson bound conditionally on each $N = n$.

To this end, let $\hat{\tau}(S_m, V) = \hat{\tau}$ to explicitly denote the dependence on S_m and V, and let

$$\tilde{\tau}(T_n) = \operatorname*{arg\,max}_{\tau \in \mathcal{T}} \tau \quad \text{subj. to} \quad U_{\text{RSCP}}(C_\tau, T_n, \delta) \leq \varepsilon.$$

Note that conditioned on obtaining n samples using rejection sampling (i.e., $|T_n(S_m, V, w, b)| = n$), we have $\hat{\tau}(S_m, V) \overset{D}{=} \tilde{\tau}(T_n)$, where $\overset{D}{=}$ denotes equality in distribution. Then, we have

$$\begin{aligned}\mathbb{P}_{S_m \sim P^m, V \sim U^m}\left[L_Q(C_{\hat{\tau}(S_m,V)}) \leq \varepsilon\right] &= \sum_{n=0}^{m} \mathbb{P}_{S_m \sim P^m, V \sim U^m}[L_Q(C_{\hat{\tau}(S_m,V)}) \leq \varepsilon \mid N = n] \cdot \mathbb{P}[N = n] \\ &= \sum_{n=0}^{m} \mathbb{P}_{T_n \sim Q^n}[L_Q(C_{\tilde{\tau}(T_n)}) \leq \varepsilon] \cdot \mathbb{P}[N = n] \\ &\geq \sum_{n=0}^{m} (1-\delta) \cdot \mathbb{P}[N = n] \\ &= 1 - \delta,\end{aligned}$$

where the inequality follows by Theorem 1. The claim follows. □

D.4. Proof of Theorem 3

First, let

$$\tilde{\tau} = \arg\max_{\tau \in \mathcal{T}} \tau \quad \text{subj. to} \quad U_{\text{RSCP}}(C_\tau, S_m, V, w^*, b, \delta_C) \leq \varepsilon, \tag{8}$$

which satisfies $\mathbb{P}_{S_m \sim P^m, V \sim U^m}\left[L_Q(C_{\tilde{\tau}}) \leq \varepsilon\right] \geq 1 - \delta_C$ by Theorem 2. Now, with probability at least $1 - \delta_w$, we have $w^* \in \mathcal{W}$. Under this event, we have

$$U_{\text{RSCP}}(C_\tau, S_m, V, w^*, b, \delta_C) \leq \max_{w \in \mathcal{W}} U_{\text{RSCP}}(C_\tau, S_m, V, w, b, \delta_C),$$

so $\hat{\tau}$ satisfies the constraint in (8). Thus, we must have $\hat{\tau} \leq \tilde{\tau}$. By monotonicity of $L_Q(C_\tau)$ in τ, we have $L_Q(C_{\hat{\tau}}) \leq L_Q(C_{\tilde{\tau}})$, which implies that

$$\mathbb{P}_{S_m \sim P^m, V \sim U^m, A}[L_Q(C_{\hat{\tau}}) \leq \varepsilon] \geq \mathbb{P}_{S_m \sim P^m, V \sim U^m, A}[L_Q(C_{\tilde{\tau}}) \leq \varepsilon] \geq 1 - \delta_C,$$

where the last step follows by Theorem 2. The claim follows by a union bound, since $w^* \in \mathcal{W}$ with probability at least $1 - \delta_w$. $\square$

D.5. Proof of Lemma 2

Let w and v be IWs where $w(x_i) \geq v(x_i)$ and $w(x_j) = v(x_j)$ for $j \neq i$. Additionally, we use the following shorthands:

$$\begin{aligned}
n_w &\coloneqq \sum_{i=1}^{m} \mathbb{1}\left(V_i \leq \frac{w(x_i)}{b}\right), \\
T_{n_w} &\coloneqq \left\{(x_i, y_i) \in S_m \;\middle|\; V_i \leq \frac{w(x_i)}{b}\right\}, \\
k_w &\coloneqq \sum_{(x,y) \in T_{n_w}} \mathbb{1}\left(y \notin C(x)\right), \\
n_v &\coloneqq \sum_{i=1}^{m} \mathbb{1}\left(V_i \leq \frac{v(x_i)}{b}\right), \\
T_{n_v} &\coloneqq \left\{(x_i, y_i) \in S_m \;\middle|\; V_i \leq \frac{v(x_i)}{b}\right\}, \text{ and} \\
k_v &\coloneqq \sum_{(x,y) \in T_{n_v}} \mathbb{1}\left(y \notin C(x)\right).
\end{aligned}$$

Here, $n_w \geq n_v$ since $w(x_i) \geq v(x_i)$. Finally, recall that $F(k; m, \theta)$ be the cumulative distribution function of a binomial random variable $\sum_{i=1}^{m} X_i$, where $X_i \sim \text{Bern}(\theta)$.

Non-decreasing case. If $y_i \notin C(x_i)$, there are two cases to consider:

1. If $\frac{v(x_i)}{b} < V_i \leq \frac{w(x_i)}{b}$, then we can verify that $n_w = n_v + 1$ and $k_w = k_v + 1$.
2. Otherwise, we can verify that $n_w = n_v$ and $k_w = k_v$.

In both cases, $k_w \geq n_v$ and $n_w \geq n_v$. Since $\overline{\theta}$ is monotonically jointly non-decreasing in (m, k) as in Lemma 1, we have

$$\begin{aligned}
U_{\text{RSCP}}(C, S_m, V, w, b, \delta) &\coloneqq U_{\text{CP}}(\bar{L}_{T_{n_w}}(C), \delta) \\
&\coloneqq \overline{\theta}(k_w; n_w, \delta) \\
&\geq \overline{\theta}(k_v; n_v, \delta) \\
&=: U_{\text{CP}}(\bar{L}_{T_{n_v}}(C), \delta) \\
&=: U_{\text{RSCP}}(C, S_m, V, v, b, \delta),
\end{aligned}$$

thus U_{RSCP} is monotonically non-decreasing in $w(x_i)$.

Non-increasing case. If $y_i \in C(x_i)$, then $k_w = k_v$. Since $\overline{\theta}$ is monotonically non-increasing in m as in Lemma 1, we have

$$\begin{aligned} U_{\text{RSCP}}(C, S_m, V, w, b, \delta) &\coloneqq U_{\text{CP}}(\bar{L}_{T_{n_w}}(C), \delta) \\ &\coloneqq \overline{\theta}(k_w; n_w, \delta) \\ &\leq \overline{\theta}(k_v; n_v, \delta) \\ &=: U_{\text{CP}}(\bar{L}_{T_{n_v}}(C), \delta) \\ &=: U_{\text{RSCP}}(C, S_m, V, v, b, \delta), \end{aligned}$$

thus U_{RSCP} is monotonically non-increasing in $w(x_i)$.

D.6. Proof of Theorem 5

Recall that

$$\begin{aligned} \hat{p}_B(x) &\coloneqq \sum_{j=1}^{K} \mathbb{1}\left(x \in B_j\right) \left[\frac{1}{m} \sum_{x' \in S_m^X} \mathbb{1}\left(x' \in B_j\right)\right], \\ \hat{q}_B(x) &\coloneqq \sum_{j=1}^{K} \mathbb{1}\left(x \in B_j\right) \left[\frac{1}{n} \sum_{x' \in T_n^X} \mathbb{1}\left(x_i \in B_j\right)\right], \\ p_B(x) &\coloneqq \sum_{j=1}^{K} \mathbb{1}\left(x \in B_j\right) \int_{B_j} p(x')\, \mathrm{d}x', \text{ and} \\ q_B(x) &\coloneqq \sum_{j=1}^{K} \mathbb{1}\left(x \in B_j\right) \int_{B_j} q(x')\, \mathrm{d}x'. \end{aligned}$$

Due to the assumption of (5), $|v(x) \cdot p(x) - p_K(x)|$ is bounded for any $x \in B_j$ as follows:

$$\begin{aligned} |v(x) \cdot p(x) - p_B(x)| &= \left| \int_{B_j} p(x)\, \mathrm{d}x' - \int_{B_j} p(x')\, \mathrm{d}x' \right| \\ &= \left| \int_{B_j} p(x) - p(x') \mathrm{d}x' \right| \\ &\leq \int_{B_j} |p(x) - p(x')|\, \mathrm{d}x' \\ &= E. \end{aligned} \tag{9}$$

Similarly,

$$|v(x) \cdot q(x) - q_B(x)| \leq E. \tag{10}$$

Observe that $m\hat{p}(x) \sim \text{Binom}\left(m, \int_{B_j} p(x')\mathrm{d}x'\right)$ for any $x \in B_j$; thus p_K is bounded with probability at least $1 - \delta'$ as follows due to the Clopper-Pearson interval $(\underline{\theta}, \overline{\theta})$:

$$\underline{\theta}(m\hat{p}(x); m, \delta') \leq p_K(x) \leq \overline{\theta}(m\hat{p}(x); m, \delta'). \tag{11}$$

Similarly,

$$\underline{\theta}(n\hat{q}(x); n, \delta') \leq q_K(x) \leq \overline{\theta}(n\hat{q}(x); n, \delta'). \tag{12}$$

From (9), (10), (11), and (12), the following holds:

$$\underline{\theta}(m\hat{p}(x); m, \delta') - E \leq v(x) \cdot p(x) \leq \overline{\theta}(m\hat{p}(x); m, \delta') + E \text{ and}$$
$$\underline{\theta}(n\hat{q}(x); n, \delta') - E \leq v(x) \cdot q(x) \leq \overline{\theta}(n\hat{q}(x); n, \delta') + E.$$

Therefore, for any $x \in B_j$, $w^*(x)$ is bounded as follows:

$$\frac{\underline{\theta}(n\hat{q}(x); n, \delta') - E}{\overline{\theta}(m\hat{p}(x); m, \delta') + E} \leq w^*(x) = \frac{q(x)}{p(x)} \leq \frac{\overline{\theta}(n\hat{q}(x); n, \delta') + E}{\underline{\theta}(m\hat{p}(x); m, \delta') - E}.$$

Since we apply the Clopper-Pearson interval for K partitions for both source and target, the claim holds due to the union bound.

E. Experiment Details

E.1. Domain Adaptation

We use a fully-connected network (with two hidden layers, where each layers has 500 neurons followed by ReLU activations and a 0.5-dropout layer) as the domain classifier (recall that the input of this domain classifier is the last hidden layer of ResNet101). We use the last hidden layer of the model as example space $\mathcal{X}$, where its dimension is 2048. For neural network training, we run stochastic gradient descent (SGD) for 100 epochs with an initial learning rate of 0.1, decaying it by half once every 20 epochs. The domain adaptation regularizer is gradually increased as in (Ganin et al., 2016). We use the same hyperparameters for all experiments.

E.2. ImageNetC-13

To train a model using domain adaptation, due to the large size of the target training set, we subsample the target training set to be the same size as the source training set on each training epoch.

E.3. CP-C Method Description

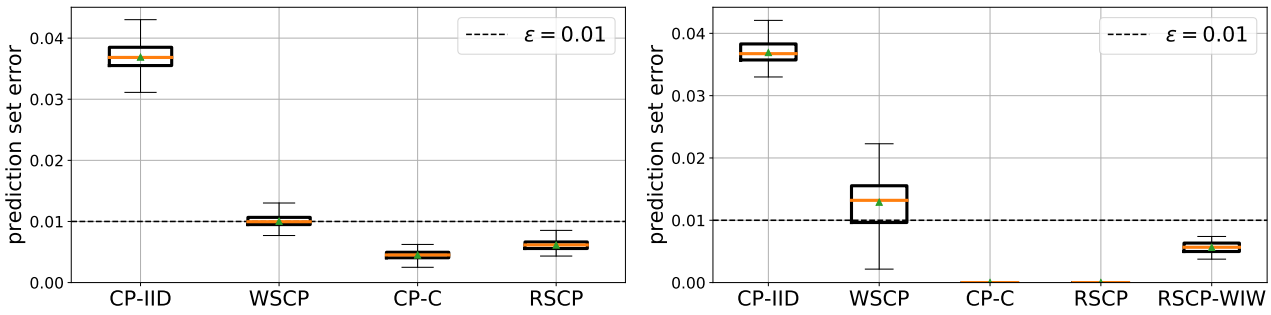
We describe the CP-C algorithm, which uses a conservative upper bound on the CP interval. Let

$$L_P(C) \coloneqq \mathop{\mathbb{E}}_{(x,y)\sim P}[\mathbb{1}(y \notin C(x))]$$
$$L_Q(C) \coloneqq \mathop{\mathbb{E}}_{(x,y)\sim Q}[\mathbb{1}(y \notin C(x))]$$
$$w^*(x) \coloneqq \frac{q(x)}{p(x)}$$
$$b \coloneqq \max_{x\in\mathcal{X}} w^*(x).$$

Then, we have

$$\begin{aligned} L_Q(C) &= \mathop{\mathbb{E}}_{(x,y)\sim Q}[\mathbb{1}(y \notin C(x))] \\ &= \mathop{\mathbb{E}}_{(x,y)\sim P}[w^*(x)\mathbb{1}(y \notin C(x))] \\ &\leq \mathop{\mathbb{E}}_{(x,y)\sim P}[b \cdot \mathbb{1}(y \notin C(x))] \\ &= b \mathop{\mathbb{E}}_{(x,y)\sim P}[\mathbb{1}(y \notin C(x))] \\ &= b \cdot L_P(C). \end{aligned}$$

Thus, $L_Q(C) \leq \varepsilon$ if $b \cdot L_P(C) \leq \varepsilon$. Equivalently, $L_Q(C) \leq \varepsilon$ if $L_P(C) \leq \varepsilon/b$. As a consequence, we can choose C based on the CP bound for the i.i.d. case (i.e., Theorem 1), except using the desired error of ε/b (instead of ε).



(a) With the known true IW

(b) With an estimated IW

Figure 4: Error under the rate shift by Two Gaussians (over 100 random trials). Parameters are $m = 50,000$, $\varepsilon = 0.01$, and $\delta = 10^{-5}$.

F. Additional Results

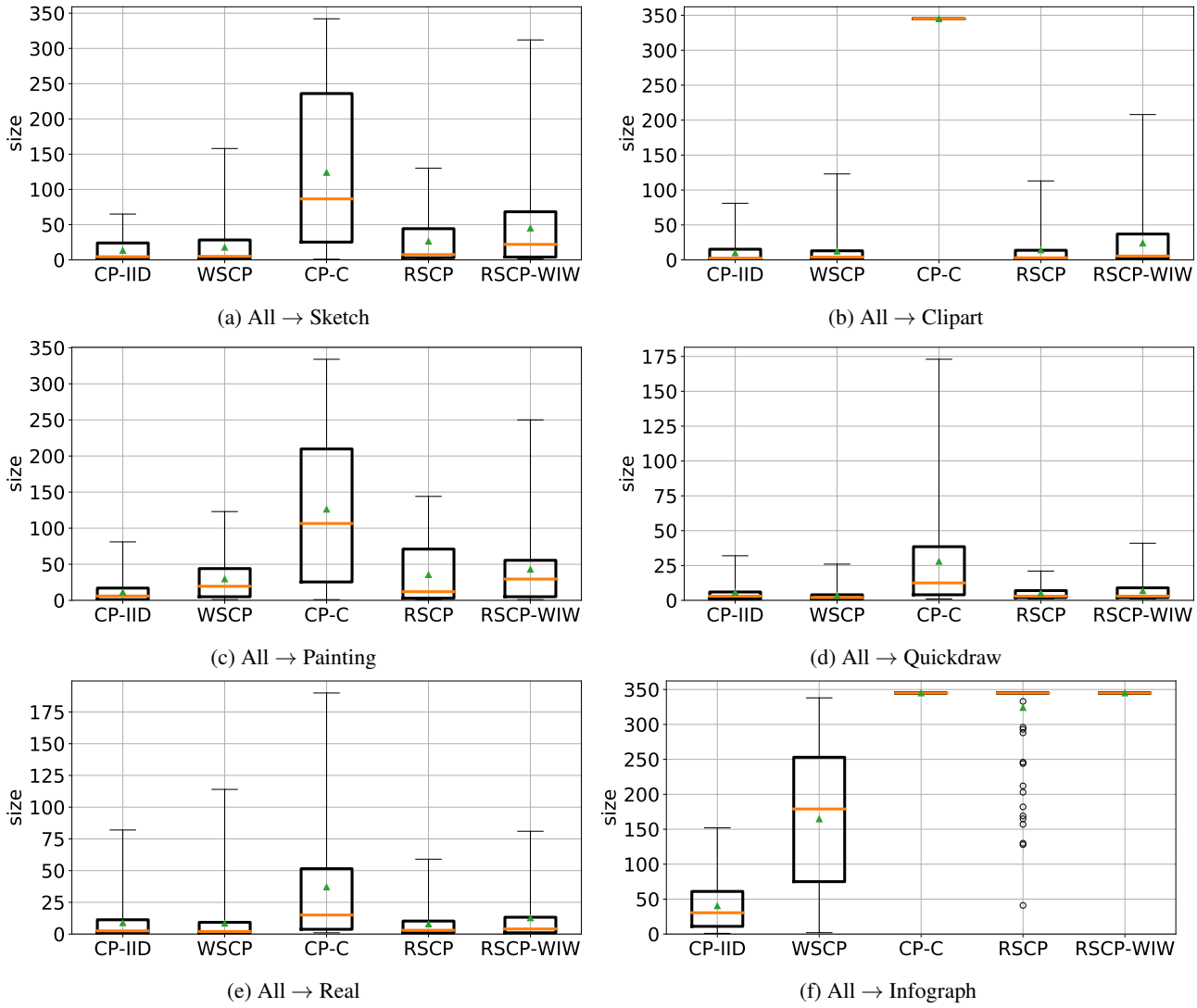
F.1. Synthetic Rate Shift by Two Gaussians

We demonstrate the efficacy of the proposed approaches (i.e., RSCP and RSCP-WIW) using a synthetic dataset consisting of samples from two Gaussian distributions.

Dataset. We consider two Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu, \Sigma')$ over 2048-dimensional covariate space $\mathcal{X}$. Here, $\mu = \mathbf{0}$; Σ and Σ' are diagonal where $\Sigma_{1,1} = 5^2$, $\Sigma_{i,i} = 10^{-1}$, $\Sigma'_{1,1} = 1$, and $\Sigma'_{i,i} = 10^{-1}$ for $i \in \{2, \ldots, 2048\}$. We consider the "flat" Gaussian $\mathcal{N}(\mu, \Sigma)$ as the source and the "tall" Gaussian $\mathcal{N}(\mu, \Sigma')$ as the target. Intuitively, there is a rate shift from the source to the target—i.e., the target examples are a subset of the source, but occur with higher frequencies. We use the following labeling function: $p(y \mid x) = \sigma(5x_1)$, where σ is the sigmoid function. Finally, we generate 50,000 labeled examples for each training, calibration, and test.

Results. We consider two different setups: 1) the true IW is known, and 2) the true IW is unknown. In Figure 4a, we demonstrate the prediction set errors given the true IW. As expected rejection sampling-based approach (RSCP) satisfies the PAC guarantee—i.e., the error is below ε. However, as shown in Figure 4b, when we need to estimate IWs, using just the point-estimate of the IW results in RSCP performing poorly in terms of prediction set error; it still satisfies the ε constraint, but the error is close to zero, indicating that the prediction set size is too large to be useful for an uncertainty quantifier. In contrast, RSCP-WIW (i.e., rejection sampling based on interval estimates of IWs) produces a larger, more reasonable error rate while still satisfying the PAC condition. These experiments demonstrate that RSCP works well when given the true IW, but accounting for IW uncertainty is important when using estimated IWs.

F.2. Prediction Set Size



(a) All → Sketch

(b) All → Clipart

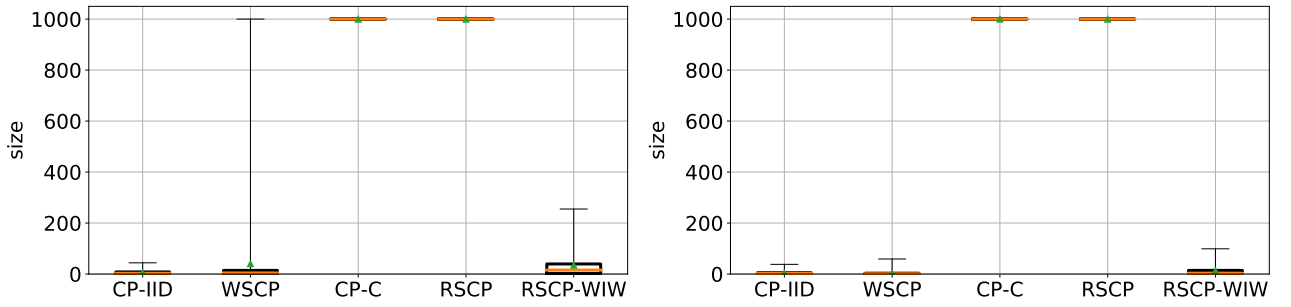
(c) All → Painting

(d) All → Quickdraw

(e) All → Real

(f) All → Infograph

Figure 5: Size distribution for the natural rate shifts on DomainNet. (a-f) We consider six different natural rate shifts. Parameters are $m = 50,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$.

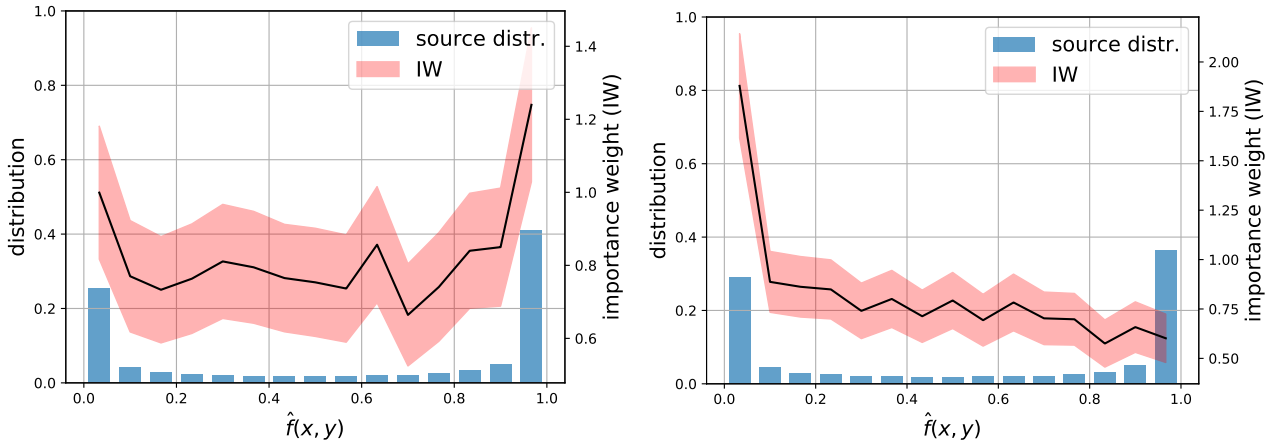


(a) ImageNet-C13

(b) ImageNet-PGD

Figure 6: Size distribution for the synthetic and adversarial shifts on ImageNet. Parameters are $m = 20,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$.

F.3. IW Distribution

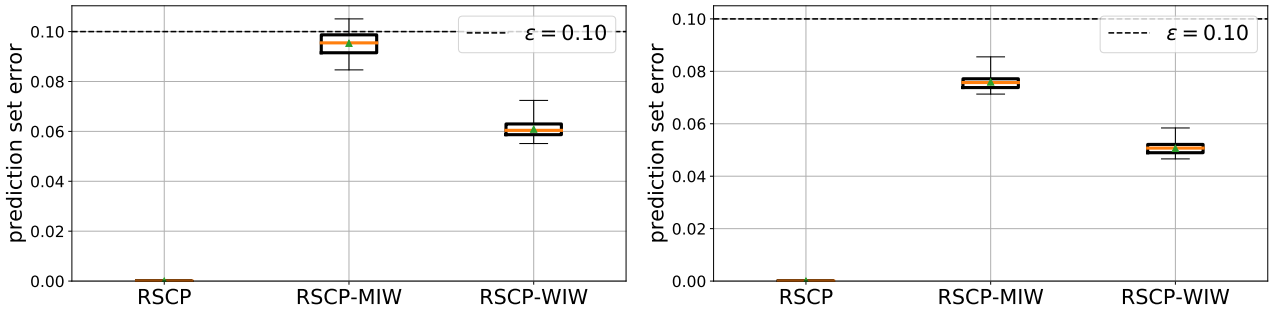


(a) All → Clipart

(b) All → Infograph

Figure 7: IW distribution over the score $f(x, y)$ axis. Each $f(x, y)$ has an IW interval $[\underline{w}(x), \overline{w}(x)]$. For visualization purposes, we quantize the $f(x, y)$ line for the histogram plot, and compute the average of $\underline{w}(x)$ and $\overline{w}(x)$ for each bin (thus, the averaged $\overline{w}(x)$ can be smaller than the original one). The trend on the IW distribution suggests that the IWs of the smaller $f(x, y)$ is higher for the shift from All to Infograph. Thus, the IWs when the prediction set $C_T(x)$ makes error is high. As a consequence, our algorithm needs to choose a very conservative T to satisfy the ε constraint, which results in zero prediction set error but very large prediction sets.

F.4. Ablation Study on Calibration



(a) ImageNet-C13

(b) ImageNet-PGD

Figure 8: Ablation study on calibration (over 100 random trial). Parameters are $m = 20,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$.

F.5. Prediction Set Visualization

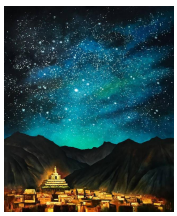
Example x	$\hat{C}_{\text{CP-IID}}(x)$	$\hat{C}_{\text{RSCP-WIW}}(x)$	Example x	$\hat{C}_{\text{CP-IID}}(x)$	$\hat{C}_{\text{RSCP-WIW}}(x)$
	{raccoon}	{owl, raccoon}		{angel, harp}	{angel, cello, harp, microphone, piano, violin}
	{wine bottle}	{bread, grapes, wine bottle, wine glass}		{shark, snorkel}	{dolphin, shark, snorkel, submarine, whale}
	{campfire}	{campfire, ocean, star, tent}		{coffee cup, cup}	{coffee cup, cup, mug, teapot}
	{ocean}	{hurricane, ocean, square, tornado}		{brain}	{brain, fish, lion, lollipop, sea turtle}
	{penguin}	{fire hydrant, foot, penguin, telephone}		{hat, hot tub}	{bed, belt, birthday cake, guitar, hat, hot tub, tiny paint can, pillow, shoe, table}
	{asparagus}	{asparagus, basket, bread, carrot, harp, lobster, toothbrush}		{baseball, onion}	{baseball, baseball bat, bread, light bulb, onion, potato}

Figure 9: Prediction sets of the DomainNet shift from All to Paint. Parameters are $m = 50,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$. A label in green and label with hat mean true label and predicted label, respectively. Note that in this shift, if RSCP-WIW is incorrect, then CP-IID is incorrect as well since the prediction set sizes are monotone in τ. Thus, we choose examples that contrast two approaches.

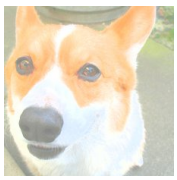
Example x	$\hat{C}_{\text{CP-IID}}(x)$	$\hat{C}_{\text{RSCP-WIW}}(x)$	Example x	$\hat{C}_{\text{CP-IID}}(x)$	$\hat{C}_{\text{RSCP-WIW}}(x)$
	{photocopier}	{printer, photocopier}		{timber wolf, red wolf, dingo}	{timber wolf, white wolf, red wolf, coyote, dingo}
	{airliner, airship, warplane}	{aircraft carrier, airliner, airship, container ship, stopwatch, parachute, tank, warplane, wing}		{forklift, golfcart, harvester, lawn mower, snowplow, tracktor}	{forklift, gokart, golfcart, harvester, lawn mower, pickup, snowplow, thresher, tracktor}
	{leopard, snow leopard, cheetah}	{leopard, snow leopard, jaguar, cheetah}		{diamondback, sidewinder}	{hognose snake, diamondback, sidewinder}
	{television}	{ent. center, monitor, screen, television}		{analog clock, barometer, odometer, stopwatch, wall clock}	{analog clock, barometer, digital watch, mag. compass, odometer, stopwatch, wall clock}
	{chameleon, green lizard}	{chameleon, green lizard, green snake, waling stick, mantis}		{barbell, dumbbell, lens cap, puck}	{barbell, barometer, car wheel, dumbbell, lens cap, power drill, puck, stethoscope}
	{Pembroke, Cardigan}	{Pembroke, Cardigan}		{face powder, lipstick, paintbrush}	{ballpoint, face powde, lipstick, paintbrush, perfume, sunscreen}

Figure 10: Prediction sets of the shift from ImageNet to ImageNet-C13. Parameters are $m = 20,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$. A label in green and label with hat mean true label and predicted label, respectively. Note that in this shift, if RSCP-WIW is incorrect, then CP-IID is incorrect as well since the prediction set sizes are monotone in τ. Thus, we choose examples that contrast two approaches.

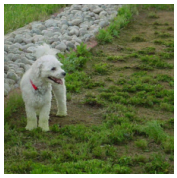
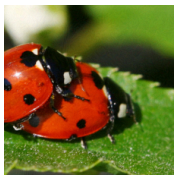
Example x	$\hat{C}_{\text{CP-IID}}(x)$	$\hat{C}_{\text{RSCP-WIW}}(x)$	Example x	$\hat{C}_{\text{CP-IID}}(x)$	$\hat{C}_{\text{RSCP-WIW}}(x)$
	{box turtle}	{terrapin, box turtle}		{brain coral, starfish, sea urchin}	{brain coral, chiton, starfish, sea urchin, sea cucumber, coral reef, stinkhorn}
	{white terrier, kuvasz, komondor}	{Maltese dog, white terrier, kuvasz, komondor, Samoyed}		{kit fox}	{red fox, kit fox}
	{ladybug}	{leaf beetle, ladybug}		{tusker}	{tusker, Afri. elephant}
	{banjo, elec. guitar, stage}	{aco. guitar, banjo, elec. guitar, stage}		{beaker, pop bottle, water bottle, wine bottle}	{beaker, beer bottle, perfume, pop bottle, water bottle, wine bottle}
	{cuirass}	{breastplate, cuirass}		{convertible, sports car}	{car wheel, convertible, sports car}
	{confectionery}	{confectionery}		{crash helmet}	{bonnet, crash helmet, football helmet}

Figure 11: Prediction sets of the shift from ImageNet to ImageNet-PGD. Parameters are $m = 20,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$. A label in green and label with hat mean true label and predicted label, respectively. Note that in this shift, if RSCP-WIW is incorrect, then CP-IID is incorrect as well since the prediction set sizes are monotone in τ. Thus, we choose examples that contrast two approaches.