Conformal Calibration from Unlabelled Pools

Kianoosh Ashouritaklimi[1]

[1]Department of Statistics, University of Oxford
Correspondence to kianoosh.ashouritaklimi@stats.ox.ac.uk

Abstract

We consider the problem of conformalising a fixed, pretrained predictor for deployment: conformal prediction can provide finite–sample marginal coverage, but only if we have a labelled calibration set that is exchangeable with the test distribution. In many applications, labels are expensive, so calibration labels must be acquired selectively from a large unlabelled pool. This creates a fundamental tension: adaptive curation policies typically induce dependencies that violate exchangeability, while naive random sampling is only appropriate when the pool matches the test distribution and can be highly inefficient (or even miscalibrated) under real–world scenarios such as covariate shift. We study calibration–set acquisition under a fixed label budget and propose a simple strategy based on density–ratio rejection sampling. Using only unlabelled covariates, we estimate the target–to–pool density ratio and accept pool points for labelling with probability proportional to this ratio, yielding labelled calibration pairs that are i.i.d. from the target distribution. This allows for the use of standard conformal prediction while avoiding the loss of effective–sample–size that can arise when correcting shift via weighted conformal methods. Synthetic regression experiments under covariate–shifted pools demonstrate valid coverage and substantially tighter prediction intervals across a range of labelling budgets.

1 Introduction

Conformal prediction (CP) (Vovk et al., 2005) is a distribution–free framework for constructing prediction sets $C(X) \subseteq \mathcal{Y}$ around the output of an arbitrary black–box predictor. Assuming that the calibration data $\{(X_i, Y_i)\}_{i=1}^n$ and test point (X_{n+1}, Y_{n+1}) are exchangeable, CP achieves finite–sample marginal coverage:

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha, \tag{1}$$

for a user–specified error rate $\alpha \in (0, 1)$. In this work, we focus on conformalising a fixed, pretrained predictor[1] $f : \mathcal{X} \to \mathcal{Y}$. This requires a held–out calibration set and a nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ which quantifies the extent to which a candidate label y fails to conform to the prediction $f(x)$. Writing $S_i = s(X_i, Y_i)$ for the calibration scores, CP computes the $(1-\alpha)(1+1/n)$–th empirical quantile $\hat{q}$ of $\{S_1, \ldots, S_n\}$ and returns the prediction set:

$$C(x) = \{y \in \mathcal{Y} : s(x, y) \leq \hat{q}\}. \tag{2}$$

In this context, CP relies on a separate, held–out calibration set, which the literature typically assumes as given. However, in many real–world applications—particularly those where labelling is

1. Note that while this is similar to the *split* conformal setting (Papadopoulos et al., 2002; Lei et al., 2015), we do *not* assume that f is trained on samples from the same distribution as our test. For example, f could be a large language model used to answer multiple–choice questions via prompting (Brown et al., 2020), or a pretrained vision–language model such as CLIP whose outputs can be used to produce class probabilities for the labels of interest (Radford et al., 2021).

costly, such as medical imaging or materials science—this assumption does not hold. This creates a need for the *curation* of a calibration set.

We study the problem of curating a calibration set from a large unlabelled pool under a fixed labelling budget. This setting introduces a key tension: standard data curation approaches (Settles, 2009) induce dependencies among selected samples, breaking the exchangeability required for valid coverage. Conversely, simple random sampling preserves exchangeability when the pool matches the test distribution, but can be highly inefficient with real pools containing irrelevant subpopulations or exhibiting covariate shift. In such cases, random sampling wastes the labelling budget on examples that are uninformative for the target test distribution and can yield unnecessarily wide prediction intervals (see Figure 1). Moreover, while it is possible to combine random sampling with *weighted conformal prediction* (Tibshirani et al., 2019) to deal with the shift induced in the calibration set, this can significantly undermine the effective sample size of the calibration set (see Figure 2), leading to a loss of statistical efficiency.

Motivated by these limitations, we propose a *label–efficient rejection sampling strategy* for acquiring calibration data from an unlabelled pool that is effective both when the pool matches the test distribution and when covariate shift is present. Our method uses a proposal density based on the Radon–Nikodym derivative (importance weight) of the test density with respect to the pool density, $w(x) = dP_{\text{test}}/dP_{\text{pool}}$, where w is estimated *purely from unlabelled data* via classifier–based density ratio estimation (Sugiyama et al., 2012). This preserves the entire labelling budget for points that are most relevant to the test distribution, and by restoring the target covariate distribution within the acquired calibration set, it enables the use of standard (unweighted) CP without the loss of effective sample size that can arise in weighted CP.

We validate the proposed approach on a synthetic regression dataset across various labelling budgets and levels of covariate shift. Initial results show that our approach provides significant gains over random sampling and weighted CP in terms of average interval width when there exists significant covariate shift between the pool and test distribution, and remains competitive otherwise.

2 Curation of Calibration Data for Conformal Prediction

2.1 Problem Setting

Let P_{pool} and P_{test} be probability measures on $\mathcal{X} \times \mathcal{Y}$, with covariate marginals $P_{\text{pool},X}$ and $P_{\text{test},X}$. We consider the setting where we are given a pretrained predictor $f : \mathcal{X} \to \mathcal{Y}$ and a nonconformity score function $s(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that measures the conformity between a prediction and the true label. Our goal is to *conformalise* f to produce prediction sets $C(X)$ satisfying the marginal coverage guarantee in (1) with respect to a target distribution P_{test}.

We work under a fixed labelling budget constraint and assume access to a large pool of unlabelled data $\mathcal{D}_{\text{pool}} = \{X_1, \ldots, X_M\}$ drawn independently and identically distributed (i.i.d.) from a source distribution P_{pool}. We are allowed to query an oracle (e.g., a human annotator) to obtain ground–truth labels for a budget of B instances, where $B \ll M$. Let $\mathcal{S} \subset \{1, \ldots, M\}$ be the set of selected indices with $|\mathcal{S}| = B$. The resulting labelled calibration set is $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i \in \mathcal{S}}$.

We assume that the pool distribution P_{pool} and the target distribution P_{test} may differ, but satisfy the standard *covariate shift* assumption (Sugiyama et al., 2007; Yamazaki and Watanabe, 2008) . Specifically, the marginal distributions of the inputs may differ ($P_{\text{pool}}(X) \neq P_{\text{test}}(X)$), but the conditional distribution of the labels given the inputs remains invariant:

$$P_{\text{pool}}(Y|X) = P_{\text{test}}(Y|X). \tag{3}$$

We further assume absolute continuity, i.e. $P_{\text{test},X} \ll P_{\text{pool},X}$, so the Radon–Nikodym derivative $w(x) = dP_{\text{test},X}/dP_{\text{pool},X}$ exists, and a bounded overlap condition $w(x) \leq W_{\max}$ for stability.

2.2 Sampling Strategies

Random sampling. A natural strategy is to construct the calibration set by *uniform random sampling* from the unlabelled pool. When $P_{\text{pool}} = P_{\text{test}}$, the calibration set is i.i.d. from P_{test}, so the calibration scores are exchangeable with the test score and we obtain the standard CP guarantee. Moreover, under $P_{\text{pool}} = P_{\text{test}}$ and a fixed labelling budget B, random sampling maximises the effective calibration sample size under the target distribution and hence minimises the variance of empirical quantile estimates of the nonconformity scores among other i.i.d. sampling strategies from P_{pool}. In contrast, under covariate shift, random sampling produces calibration covariates from $P_{\text{pool},X}$ rather than $P_{\text{test},X}$, so the induced score distribution generally differs between calibration and testing. Consequently, the required exchangeability condition fails, and using the empirical quantile of $\{S_b^{\text{cal}}\}_{b=1}^B$ can lead to systematic miscalibration (under– or over–coverage) on P_{test} (see Figure 1).

Random sampling and weighted conformal prediction. An alternative approach is to accept the induced shift produced by random sampling and correct for it using *weighted conformal prediction* (Tibshirani et al., 2019). Define the importance weights:

$$w(x) = \frac{dP_{\text{test},X}}{dP_{\text{pool},X}}(x). \tag{4}$$

Weighted CP instead uses a *weighted* conformal rank that assigns weights $w(X_i)$ to calibration points and also includes the test covariate via weight $w(x)$ (see Appendix F for more details). With known w, this yields the target–marginal guarantee $\mathbb{P}_{(X,Y)\sim P_{\text{test}}}\{Y \in C(X)\} \geq 1-\alpha$ under covariate shift. However, this approach can be statistically inefficient: under covariate shift, random sampling spends labels on regions where $w(X_i)$ is small (i.e. unrepresentative of $P_{\text{test},X}$), and these points contribute little after reweighting. A useful proxy is the effective sample size (ESS, Kish (1965); Gretton et al. (2009)):

$$B_{\text{eff}} \approx \frac{\left(\sum_{i=1}^B w(X_i)\right)^2}{\sum_{i=1}^B w(X_i)^2}, \tag{5}$$

which can be much smaller than B when w is highly variable, leading to noisier quantile estimates and wider intervals despite using B labels. We observe these effects in Figures 1(d), 2. (When $P_{\text{pool}} = P_{\text{test}}$, w is constant and weighted CP reduces to standard CP.)

3 Density–Based Rejection Sampling

We propose a simple strategy that aims to (i) retain the *unweighted* split conformal guarantee by ensuring exchangeability with the target distribution, while (ii) spending the labelling budget primarily on covariates that are likely under $P_{\text{test},X}$. We describe our procedure below and provide the full algorithm in Algorithm 1.

Assume that the weights $w(x)$ in (4) are essentially bounded, i.e. there exists a finite constant $W_{\max} > 0$ such that

$$0 \leq w(x) \leq W_{\max}, \qquad \text{for } P_{\text{pool},X}\text{–a.e. } x. \tag{6}$$

We sample calibration points by sampling $X \sim P_{\text{pool},X}$ and accepting with probability $w(X)/M$; upon acceptance, we query labels Y from the oracle. This is repeated until B covariates have been accepted and guarantees that the accepted covariates and labelled pairs satisfy $X_i \overset{\text{iid}}{\sim} P_{\text{test},X}$ and $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{\text{test}}$ respectively. Consequently, we can use the standard, unweighted CP procedure on the collected B calibration points and obtain our desired coverage guarantee, as stated below and proved in Appendix C.

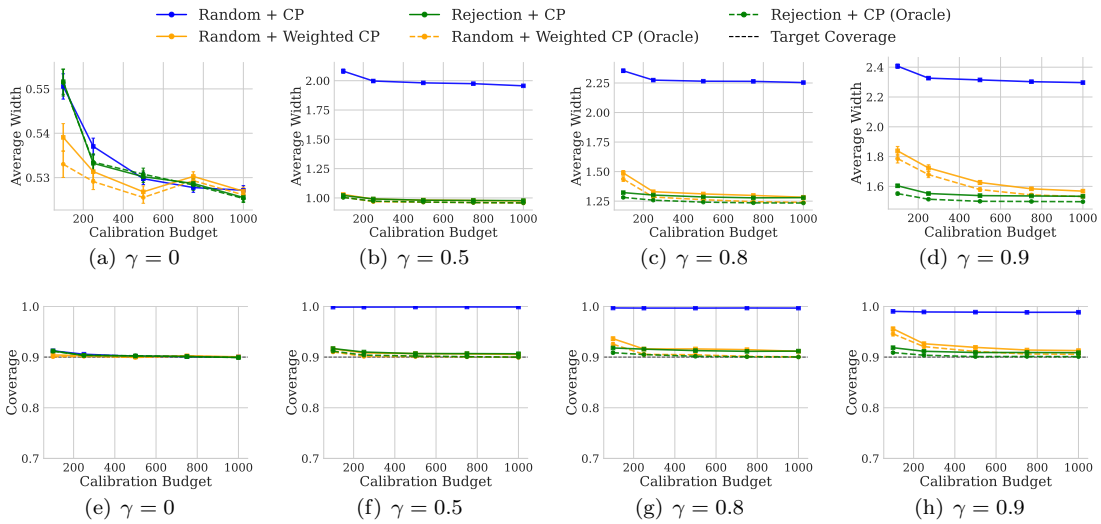


Figure 1. Average interval width and coverage results for different labelling budgets and levels of covariate shift for our synthetic dataset. Results show the mean ± standard err. over 250 trials.

Theorem 1 (Finite–sample, marginal coverage, informal). *Assume the setup in Section 2.1 and that (6) holds for some $W_{\max} < \infty$. Construct a labelled calibration set $\{(X_i, Y_i)\}_{i=1}^{B}$ by the rejection sampling procedure above and let $(X_{B+1}, Y_{B+1}) \sim P_{test}$ be an independent test point. Then, applying standard conformal prediction with any nonconformity score s to this calibration set yields the desired finite–sample target–marginal coverage guarantee:*

$$\mathbb{P}\{Y_{B+1} \in C(X_{B+1})\} \;\geq\; 1 - \alpha. \tag{7}$$

Theorem 1 is stated under two idealisations: (a) we can draw proposals $X \sim P_{\text{pool},X}$, and (b) we know the true $W_{\max}$ and ratio w. In practice, we instead draw proposals by resampling from the finite pool, and we replace $W_{\max}$ and w with estimates $\hat{w}$ and $\hat{W}_{\max}$ obtained from unlabelled data via classifier–based density ratio estimation, i.e. we train a probabilistic classifier to distinguish unlabelled covariates[2] drawn from $P_{\text{test},X}$ versus $P_{\text{pool},X}$. Writing $\hat{\eta}(x) \approx \mathbb{P}(D = 1 \mid X = x)$ for the fitted probability that x comes from the target sample ($D = 1$) rather than the pool ($D = 0$), the density ratio is recovered via the standard likelihood–ratio identity

$$w(x) \;=\; \frac{dP_{\text{test},X}}{dP_{\text{pool},X}}(x) \;\approx\; \hat{w}(x) \;:=\; \frac{1-\pi}{\pi}\,\frac{\hat{\eta}(x)}{1-\hat{\eta}(x)}, \tag{8}$$

where $\pi = \mathbb{P}(D = 1)$ is the class prior. Subsequently, the rejection sampling envelope is estimated as $\hat{W}_{\max} = \max_{x \in \mathcal{D}_{\text{pool}}} \hat{w}(x)$. These approximations induce deviations from exact P_{test} exchangeability; we discuss the errors that they induce in Appendix D and find empirically that they do not materially hurt the coverage of our approach.

4 Experiments

We evaluate our approach on a synthetic 1D regression problem under covariate shift induced by contamination. We compare (i) random sampling with standard CP (**Random**), (ii) random sampling with weighted CP (**Random + Weighted CP**), and (iii) our density–based rejection sampling with standard CP (**Rejection + CP**). For the weighted methods, we report results using both oracle weights and estimated weights.

2. Note that assuming access to unlabelled covariates from our test distribution is common in various works focusing on data curation, e.g. Bickford Smith et al. (2023, 2024).

Experimental Setup. We simulate covariate shift where the test distribution $P_{\text{test},X}$ is Gaussian and the unlabelled pool $P_{\text{pool},X}$ is a mixture of $P_{\text{test},X}$ and a corrupted Gaussian $P_{\text{corr},X}$ with mixing parameter γ (with $\gamma = 1$ corresponding to using only $P_{\text{corr},X}$; details in Appendix B). Labels are generated via $Y = g(X) + \varepsilon$ with heteroscedastic noise, and we train a fixed random forest predictor f on a small labelled set from P_{test}. We use the nonconformity score $s(x, y) = |y - f(x)|$.

We estimate importance weights $w(x) = dP_{\text{test}}/dP_{\text{pool}}$ by training a logistic regression classifier to distinguish test samples from pool samples, using the odds ratio $\hat{w}(x) \propto \hat{\eta}(x)/(1 - \hat{\eta}(x))$ normalised over the pool. We estimate the rejection sampling envelope as $\hat{W}_{\max} = \max_{x \in \mathcal{D}_{\text{pool}}} \hat{w}(x)$. All experiments are repeated for 250 trials with nominal error rate of $\alpha = 0.1$; see Appendix B for full details.

Results. Figure 1 shows that our approach yields the largest reductions in average interval width when covariate shift is severe and the labelling budget is small, while remaining competitive across all other regimes. This is consistent with our intuition: under shift, random sampling wastes labels on irrelevant (corrupted) covariates, whereas weighted CP can suffer from substantially reduced effective sample size (see Figure 2).

We also find that **Random + WCP** significantly improves over **Random + CP**, showing the benefit of correcting for the covariate shift. Moreover, our approach attains the desired coverage level across all settings, suggesting that the practical approximations (finite pool resampling and estimated weights) do not materially degrade performance in this setting. Finally, **Random + CP** and **Rejection + CP** behave similarly under estimated versus oracle weights, indicating that the density ratio is well–approximated in this synthetic setup.

5 Conclusions

In this work, we introduced a label–efficient approach for curating conformal calibration data from an unlabelled pool under covariate shift. By estimating a target–to–pool density ratio from unlabelled covariates and using it to rejection–sample which pool points to label, we concentrate the labelling budget on target–relevant inputs, yielding an effectively unweighted calibration set with (approximately) full effective sample size. Under idealised assumptions, this enables CP to retain the desired finite–sample target–marginal coverage guarantee, and our synthetic experiments indicate that the practical approximations do not materially harm coverage while improving interval width when shift is large and budgets are small.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175, May 2010. ISSN 0885-6125. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7331–7348. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/bickfordsmith23a.html.

Freddie Bickford Smith, Adam Foster, and Tom Rainforth. Making better use of unlabelled data in Bayesian active learning. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 847–855. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/bickford-smith24a.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Imre Csiszár. i-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975. URL https://api.semanticscholar.org/CorpusID:18053591.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M. Borgwardt, Bernhard Scholkopf, Quiñonero Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. Covariate shift by kernel mean matching. In *Neural Information Processing Systems*, 2009. URL https://api.semanticscholar.org/CorpusID:108301245.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

Leslie Kish. *Survey sampling.* Wiley, New York, 1965.

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

Zachary Chase Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. *ArXiv*, abs/1802.03916, 2018. URL https://api.semanticscholar.org/CorpusID:3608424.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, 2002. URL https://api.semanticscholar.org/CorpusID:42084298.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Christian P. Robert and George Casella. Monte carlo statistical methods. In *Springer Texts in Statistics*, 2004. URL https://api.semanticscholar.org/CorpusID:33807429.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14:21–41, 2002. URL https://api.semanticscholar.org/CorpusID:18254013.

Burr Settles. Active learning literature survey. 2009. URL https://api.semanticscholar.org/CorpusID:324600.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:985–1005, 2007. URL https://api.semanticscholar.org/CorpusID:17547265.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

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

Keisuke Yamazaki and Sumio Watanabe. Experimental bayesian generalization error of non-regular models under covariate shift. In Masumi Ishikawa, Kenji Doya, Hiroyuki Miyamoto, and Takeshi Yamakawa, editors, *Neural Information Processing*, pages 466–476, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-69158-7.

Appendix Contents

A Further Discussions 8

B Experimental Setup 10

C Proofs 11

D Approximation Errors 12

D.1 Estimated Weights 12

D.2 Finite–Pool Approximations 15

E Additional Results 17

F Additional Details 18

G Algorithm Block 19

Appendix A. Further Discussions

Below, we describe the limitations and potential extensions of our approach.

Beyond covariate shift: messier pools. Our approach relies on the covariate shift assumption $P_{\text{pool}}(Y \mid X) = P_{\text{test}}(Y \mid X)$, which ensures that once we restore the target covariate distribution $P_{\text{test},X}$ via rejection sampling, the accepted labelled pairs are i.i.d. from P_{test}. Real pools may violate this, and it is useful to distinguish several cases.

(i) Label shift. If $P_{\text{pool}}(X \mid Y) = P_{\text{test}}(X \mid Y)$ but $P_{\text{pool}}(Y) \neq P_{\text{test}}(Y)$ (Saerens et al., 2002; Lipton et al., 2018), then

$$\frac{\mathrm{d}P_{\text{test}}}{\mathrm{d}P_{\text{pool}}}(x, y) = \frac{P_{\text{test}}(y)}{P_{\text{pool}}(y)} =: \rho(y), \qquad \text{(depends only on } y\text{)}.$$

In this setting, targeting $P_{\text{test},X}$ is not sufficient: even with $X \sim P_{\text{test},X}$, the conditional $Y \mid X$ differs unless $\rho(y) \equiv 1$. A natural extension is therefore to use *joint* importance weights $\rho(Y_i)$ inside a weighted conformal procedure.

(ii) Mixture pools and subpopulation shift. Large pools often contain multiple latent subpopulations, some of which are irrelevant at test time (Ben-David et al., 2010; Hoffman et al., 2018) This can be expressed as $P_{\text{pool},X} = \sum_{k=1}^{K} \pi_k Q_{k,X}$ with an unknown mixture, while $P_{\text{test},X}$ concentrates on a subset of components. A useful extension is to make the ratio estimation *component–aware*. Concretely, one can first identify latent subpopulations (e.g., by clustering a representation $\phi(X)$ or fitting a mixture density model), then estimate target–to–pool ratios within each component and run rejection sampling restricted to the components that have substantial target mass. This avoids proposing many points from irrelevant components, improving the acceptance rate, and reduces instability when global overlap is weak due to rare target–relevant components.

Estimating $w(x)$ without unlabelled target data via exponential tilting. Our current ratio estimator assumes access to unlabelled covariates from both $P_{\text{pool},X}$ and $P_{\text{test},X}$. In some applications, unlabelled test covariates are unavailable, but we may still have *side information* about the test feature distribution (e.g. summary statistics, published demographics, or moments of a semantically rich embedding such as CLIP, (Radford et al., 2021)). A principled approach is to assume that $P_{\text{test},X}$ is an *exponentially tilted* version of $P_{\text{pool},X}$:

$$p_{\text{test}}(x) = \frac{\exp(\lambda^\top \phi(x))}{Z(\lambda)}\, p_{\text{pool}}(x), \qquad Z(\lambda) := \mathbb{E}_{P_{\text{pool}}}\big[\exp(\lambda^\top \phi(X))\big], \tag{9}$$

where $\phi : \mathcal{X} \to \mathbb{R}^d$ is a chosen feature map (e.g. $\phi(x)$ could be a low-dimensional projection of a frozen CLIP embedding), and $\lambda \in \mathbb{R}^d$. Then the density ratio has the closed form

$$w(x) = \frac{p_{\text{test}}(x)}{p_{\text{pool}}(x)} = \frac{\exp(\lambda^\top \phi(x))}{Z(\lambda)}. \tag{10}$$

To identify λ without target samples, suppose we know (or can estimate externally) a moment constraint of the form

$$\mathbb{E}_{P_{\text{test}}}[\phi(X)] = m \in \mathbb{R}^d. \tag{11}$$

Under (9), this constraint becomes

$$m = \nabla_\lambda \log Z(\lambda) = \frac{\mathbb{E}_{P_{\text{pool}}}[\phi(X)\exp(\lambda^\top \phi(X))]}{\mathbb{E}_{P_{\text{pool}}}[\exp(\lambda^\top \phi(X))]}.$$

Equivalently, λ is the unique minimiser of the convex objective

$$\lambda^\star \in \arg\min_{\lambda \in \mathbb{R}^d} \Big\{ \log Z(\lambda) - \lambda^\top m \Big\}, \tag{12}$$

which is the standard maximum–entropy solution (Csiszár, 1975). Given only a finite unlabelled pool $D_{\text{pool}} = \{x_j\}_{j=1}^M$, we can approximate $Z(\lambda)$ by $\widehat{Z}(\lambda) = \frac{1}{M}\sum_{j=1}^M \exp(\lambda^\top \phi(x_j))$ and solve (12) by gradient methods. This yields an explicit, normalised ratio estimate

$$\widehat{w}(x) = \frac{\exp((\lambda^\star)^\top \phi(x))}{\widehat{Z}(\lambda^\star)}, \tag{13}$$

which can be plugged directly into our rejection–sampling acceptance probability. This extension is attractive when target covariates are unavailable but informative target *summaries* are, and it naturally supports semantically meaningful $\phi(x)$ (e.g. CLIP-based) that may capture the aspects of shift most relevant for calibration.

Appendix B. Experimental Setup

Below, we give the full details for our experimental setup.

Data–generating process. Test covariates follow $X \sim P_{\text{test},X} = \mathcal{N}(\mu_{\text{test}}, \sigma^2_{\text{test}})$, and corrupt covariates follow $X \sim P_{\text{corr},X} = \mathcal{N}(\mu_{\text{corr}}, \sigma^2_{\text{corr}})$. The unlabelled pool exhibits covariate shift via contamination:

$$P_{\text{pool},X} \;=\; (1-\gamma)\, P_{\text{test},X} \;+\; \gamma\, P_{\text{corr},X}, \qquad \gamma \in [0,1]. \tag{14}$$

Conditionals are shared, with the labels given by

$$Y = g(X) + \varepsilon, \qquad \varepsilon \mid X \sim \mathcal{N}(0, \sigma(X)^2), \tag{15}$$

with $g(x) = \sin(1.5x) + 0.3x$ and heteroscedastic noise $\sigma(x) = 0.15 + 0.10|x|$.

Pool and test sets. We draw an unlabelled pool of size $M = 100{,}000$ from $P_{\text{pool},X}$. We also draw a test set of size $n_{\text{test}} = 1000$ from P_{test}.

Predictor training. We train a fixed random forest regressor f using $n_{\text{train}} = 100$ labelled samples drawn from P_{test} and the default hyperparameters from `scikit--learn` (Pedregosa et al., 2011). The resulting f is held fixed throughout all calibration trials.

Oracle and estimated importance weights. The oracle density ratio used throughout is

$$w(x) \;=\; \frac{dP_{\text{test},X}}{dP_{\text{pool},X}}(x) \;=\; \frac{p_{\text{test}}(x)}{(1-\gamma)p_{\text{test}}(x) + \gamma p_{\text{corr}}(x)}, \tag{16}$$

where $p_{\text{test}}, p_{\text{corr}}$ denote the corresponding Gaussian densities.

To estimate weights, we train a probabilistic discriminator using only unlabelled covariates. Specifically, we sample $n_w = 100$ points from $P_{\text{test},X}$ and $n_w = 100$ points from $P_{\text{pool},X}$ and fit a logistic regression classifier with the default hyperparameters from `scikit--learn` (Pedregosa et al., 2011) to distinguish *test* vs. *pool*. Let $\hat{\eta}(x)$ be the estimated probability that x came from the test distribution. With balanced class sampling, we use the classifier–odds estimator

$$\hat{w}(x) \;\propto\; \frac{\hat{\eta}(x)}{1-\hat{\eta}(x)}, \tag{17}$$

and (when needed) rescale $\hat{w}$ so that its empirical mean over the pool equals 1, matching $\mathbb{E}_{P_{\text{pool},X}}[w(X)] = 1$.

Estimating the envelope $W_{\max}$. For rejection sampling we require an envelope $W_{\max}$ such that $w(x) \le W_{\max}$ (or similarly for $\hat{w}$). In practice we estimate it from the finite pool:

$$\hat{W}_{\max} \;=\; \max_{x \in \mathcal{D}_{\text{pool}}} \hat{w}(x), \tag{18}$$

(and analogously $W_{\text{oracle}} = \max_{x \in \mathcal{D}_{\text{pool}}} w(x)$ when using oracle weights).

Conformity score and prediction intervals. We use the residual score

$$s(x,y) \;=\; |y - f(x)|. \tag{19}$$

Given a labelled calibration set of size B, standard CP computes the $(1-\alpha)(1+1/B)$ empirical quantile $\hat{q}$ of $\{s(X_i,Y_i)\}_{i=1}^{B}$ and returns the interval

$$C(x) \;=\; \big[f(x) - \hat{q},\; f(x) + \hat{q}\big]. \tag{20}$$

For **Random + Weighted CP**, we implement weighted conformal prediction as in Tibshirani et al. (2019) (with either oracle w or estimated $\hat{w}$).

Appendix C. Proofs

Here, we formally state Theorem 1 and prove it.

Theorem 2 (Finite–sample target–marginal coverage under density–based rejection–sampling). *Let P_{pool} and P_{test} be distributions on $\mathcal{X} \times \mathcal{Y}$, with $P_{\text{pool}}(Y \mid X) = P_{\text{test}}(Y \mid X)$ and $P_{\text{test},X} \ll P_{\text{pool},X}$. Moreover, define the Radon–Nikodym derivative*

$$w(x) \;:=\; \frac{dP_{\text{test},X}}{dP_{\text{pool},X}}(x),$$

and suppose that there exists $W_{\max} < \infty$ such that $0 \le w(x) \le W_{\max}$ for $P_{pool,X}$–a.e. x. Fix a label budget $B \in \mathbb{N}$, sample $(X_i, Y_i)_{i=1}^{B}$ via the density–based rejection sampling procedure in Section 3, and let $(X_{B+1}, Y_{B+1}) \sim P_{\text{test}}$ be an independent test point.

Let $f : \mathcal{X} \to \mathcal{Y}$ be a fixed predictor, $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any nonconformity score, and define $S_i := s(X_i, Y_i)$ for $i = 1, \ldots, B$. Let $k := \lceil (B+1)(1-\alpha) \rceil$ and let $S_{(k)}$ denote the kth order statistic of $\{S_1, \ldots, S_B\}$. Define the conformal prediction set

$$C(x) \;:=\; \{y \in \mathcal{Y} : s(x, y) \le S_{(k)}\}.$$

Then $C(\cdot)$ satisfies the finite–sample target–marginal coverage guarantee

$$\mathbb{P}\{Y_{B+1} \in C(X_{B+1})\} \;\ge\; 1 - \alpha.$$

Proof. Let $Q := P_{\text{pool},X}$ and $P := P_{\text{test},X}$.

Exchangeability. By standard rejection sampling results (Robert and Casella, 2004), if $X \sim Q$ is accepted with probability $w(X)/W_{\max}$, then the accepted draw $X^\star$ satisfies $X^\star \sim P$. Moreover, independent repetitions of this procedure yields i.i.d. accepted draws, so $X_1, \ldots, X_B \overset{\text{iid}}{\sim} P$. Since the accepted covariates satisfy $X_i \overset{\text{iid}}{\sim} P$ and labels are generated from the common conditional $P(Y \mid X)$ under covariate shift, we have $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{\text{test}}$. Together with an independent test point $(X_{B+1}, Y_{B+1}) \sim P_{\text{test}}$, the $B+1$ pairs are therefore exchangeable.

Conformal validity. Define scores $S_i := s(X_i, Y_i)$ for $i = 1, \ldots, B+1$. Since the underlying pairs are exchangeable, the scores $(S_1, \ldots, S_B, S_{B+1})$ are also exchangeable. Let $k = \lceil (B+1)(1-\alpha) \rceil$ and $S_{(k)}$, $S^+_{(k)}$ be the kth order statistics of $\{S_1, \ldots, S_B\}$, and $\{S_1, \ldots, S_B, S_{B+1}\}$ respectively. The conformal set is $C(x) = \{y : s(x, y) \le S_{(k)}\}$, and the coverage event is given by:

$$\{Y_{B+1} \in C(X_{B+1})\} = \{S_{B+1} \le S_{(k)}\}.$$

With randomized tie–breaking (or if the scores are a.s. distinct), the rank R of S_{B+1} among the $B+1$ scores is uniform on $\{1, \ldots, B+1\}$. Then

$$\mathbb{P}\{S_{B+1} \le S^+_{(k)}\} = \mathbb{P}\{R \le k\} = \frac{k}{B+1}.$$

Moreover, since $S_{(k)}$ is the kth order statistic of $\{S_1, \ldots, S_B\}$, we have $S_{(k)} \ge S^+_{(k)}$, and hence

$$\mathbb{P}\{S_{B+1} \le S_{(k)}\} \ge \mathbb{P}\{S_{B+1} \le S^+_{(k)}\} = \frac{k}{B+1} \ge 1 - \alpha.$$

This is exactly the desired finite–sample target–marginal coverage guarantee. □

Appendix D. Approximation Errors

This section analyses two independent departures from the idealised setting of Theorem 1: (i) drawing proposals by resampling from a finite pool rather than from $P_{\text{pool},X}$ and (ii) using an estimated density ratio $\widehat{w}$ in place of the true ratio w. Throughout, write

$$Q := P_{\text{pool},X}, \qquad P := P_{\text{test},X}, \qquad w(x) := \frac{\mathrm{d}P}{\mathrm{d}Q}(x).$$

D.1 Estimated Weights

In Theorem 1 we assume access to the true density ratio to perform rejection sampling from the pool covariate marginal Q and obtain accepted covariates i.i.d. from P. In practice, the weight function is often replaced by an estimator $\hat{w}$. This induces a mismatch between the distribution of accepted calibration covariates and the test covariate distribution, breaking exact exchangeability. In this section we quantify the resulting error and obtain an *approximate* marginal coverage guarantee.

Acceptance with estimated weights. Assume we run rejection sampling with acceptance probability

$$a_{\hat{w}}(x) = \frac{\hat{w}(x)}{\hat{W}_{\max}}, \qquad \text{where} \qquad 0 \le \hat{w}(x) \le \hat{W}_{\max} \;\; \forall x, \tag{21}$$

so that $a_{\hat{w}}(x) \in [0,1]$ without clipping.[3] Let $X^\star$ denote an accepted proposal when $X \sim Q$ is proposed and accepted with probability (21). A standard calculation shows that the accepted covariate distribution $\hat{P}_X$ satisfies, for any measurable set $A \subseteq \mathcal{X}$,

$$\mathbb{P}(X^\star \in A) = \frac{\int_A a_{\hat{w}}(x)\, Q(dx)}{\int a_{\hat{w}}(x)\, Q(dx)} = \frac{\int_A \hat{w}(x)\, Q(dx)}{\int \hat{w}(x)\, Q(dx)}. \tag{22}$$

Define the normalising constant

$$Z := \mathbb{E}_{X\sim Q}[\hat{w}(X)] = \int \hat{w}(x)\, Q(dx) \in (0,\infty). \tag{23}$$

Then (22) is equivalent to the Radon–Nikodym form

$$\frac{d\hat{P}_X}{dQ}(x) = \frac{\hat{w}(x)}{Z}. \tag{24}$$

In contrast, under the idealised setting of Theorem 1, $\frac{dP}{dQ}(x) = w(x)$ and $Z = 1$.

Induced joint distribution of accepted labelled pairs. Under the covariate shift assumption, if each accepted covariate $X_i^\star$ is labelled via the true conditional $Y_i^\star \mid X_i^\star = x \sim P_{\text{test}}(Y \mid X = x)$, then the accepted labelled pairs are i.i.d. from the joint distribution

$$\hat{P}(dx, dy) := \hat{P}_X(dx)\, P_{\text{test}}(dy \mid x), \tag{25}$$

whereas the test point (X_{B+1}, Y_{B+1}) is distributed as

$$P_{\text{test}}(dx, dy) = P(dx)\, P_{\text{test}}(dy \mid x). \tag{26}$$

Thus the only source of mismatch is the marginal covariate distribution $\hat{P}_X \neq P$.

3. If clipping is used, $a(x) = \min\{1, \hat{w}(x)/\hat{M}\}$, then the analysis below holds with $\hat{w}$ replaced by $\tilde{w}(x) = \min\{\hat{w}(x), \hat{M}\}$.

A distance–to–target quantity. We quantify the mismatch between $\hat{P}_X$ and P using total variation:

$$\delta_X := \mathrm{TV}(\hat{P}_X, P) := \sup_{A \subseteq \mathcal{X}} |\hat{P}_X(A) - P(A)| = \frac{1}{2} \int \Big| \frac{\hat{w}(x)}{Z} - w(x) \Big| Q(dx)$$

$$= \frac{1}{2} \mathbb{E}_{X \sim Q} \Big[\Big| \frac{\hat{w}(X)}{Z} - w(X) \Big| \Big]. \tag{27}$$

Since $\hat{P}$ and P_{test} share the same conditional $Y \mid X$, we also have

$$\mathrm{TV}(\hat{P}, P_{\text{test}}) = \mathrm{TV}(\hat{P}_X, P) = \delta_X. \tag{28}$$

The below result shows, not surprisingly, that δ_X can be bounded by the error resulting from the estimated weights.

Lemma 3 (Bounding δ_X by weight estimation error). *Let $e(x) := \hat{w}(x) - w(x)$. Then*

$$\delta_X \le \mathbb{E}_{X \sim Q}[|e(X)|] \qquad \textit{and} \qquad \delta_X \le \sqrt{\mathbb{E}_{X \sim Q}[e(X)^2]}. \tag{29}$$

Moreover, if there exists $\varepsilon \in (0, 1)$ such that

$$\sup_{x \in \mathcal{X}} \Big| \frac{\hat{w}(x)}{w(x)} - 1 \Big| \le \varepsilon, \tag{30}$$

then

$$\delta_X \le \frac{\varepsilon}{1 - \varepsilon}. \tag{31}$$

Proof. Recall $Z = \mathbb{E}_Q[\hat{w}]$ and $\mathbb{E}_Q[w] = 1$. Then $Z - 1 = \mathbb{E}_Q[e]$ and hence $|Z - 1| \le \mathbb{E}_Q[|e|]$. Using the triangle inequality,

$$\Big| \frac{\hat{w}}{Z} - w \Big| \le \Big| \frac{\hat{w}}{Z} - \hat{w} \Big| + |\hat{w} - w| = |\hat{w}| \Big| \frac{1}{Z} - 1 \Big| + |e|.$$

Taking expectations under Q and using $\mathbb{E}_Q[\hat{w}] = Z$ yields

$$\mathbb{E}_Q \Big[\Big| \frac{\hat{w}}{Z} - w \Big| \Big] \le Z \Big| \frac{1}{Z} - 1 \Big| + \mathbb{E}_Q[|e|] = |1 - Z| + \mathbb{E}_Q[|e|] \le 2\, \mathbb{E}_Q[|e|].$$

Substituting into (27) gives $\delta_X \le \mathbb{E}_Q[|e|]$. The L_2 bound follows from Cauchy–Schwarz: $\mathbb{E}_Q[|e|] \le \sqrt{\mathbb{E}_Q[e^2]}$.

For the uniform relative error bound, assume (30). Then $\hat{w}(x) \in [(1 - \varepsilon) w(x), (1 + \varepsilon) w(x)]$ for all x, so $Z = \mathbb{E}_Q[\hat{w}] \in [1 - \varepsilon,\, 1 + \varepsilon]$. Therefore

$$\frac{d\hat{P}_X}{dP}(x) = \frac{\frac{d\hat{P}_X}{dQ}(x)}{\frac{dP}{dQ}(x)} = \frac{\hat{w}(x)/(Z)}{w(x)} = \frac{\hat{w}(x)}{Z\, w(x)} \in \Big[\frac{1 - \varepsilon}{1 + \varepsilon}, \frac{1 + \varepsilon}{1 - \varepsilon} \Big].$$

Hence $\big| \frac{d\hat{P}_X}{dP}(x) - 1 \big| \le \frac{2\varepsilon}{1 - \varepsilon}$ for all x. Using $\mathrm{TV}(\hat{P}_X, P) = \frac{1}{2} \int \big| \frac{d\hat{P}_X}{dP} - 1 \big| P(dx)$ gives $\delta_X \le \varepsilon/(1 - \varepsilon)$. □

Approximate marginal coverage under weight estimation We now translate the mismatch δ_X into an approximate coverage bound.

Let $\widehat{C}(\cdot)$ denote the (unweighted) split conformal prediction set built from B calibration pairs. Let A denote the coverage event $A := \{Y_{B+1} \in \widehat{C}(X_{B+1})\}$. Under the assumptions of Theorem 1 (true weights w), $\mathbb{P}(A) \ge 1 - \alpha$. When $\hat{w}$ is used, the calibration set is i.i.d. from $\hat{P}$ defined in (25), while the test point is from P_{test} as in (26). Thus the joint distribution is $\hat{P}^B \otimes P_{\text{test}}$ instead of P_{test}^{B+1}.

Proposition 4 (Approximate target–marginal coverage with estimated weights). *Assume covariate shift and that the calibration pairs are constructed via rejection sampling using $\hat{w}$ as in (21), yielding i.i.d. calibration pairs $(X_1, Y_1), \ldots, (X_B, Y_B) \sim \hat{P}$, and let $(X_{B+1}, Y_{B+1}) \sim P_{\text{test}}$ be an independent test point. Then the standard split conformal set $\widehat{C}(\cdot)$ built from $\{(X_i, Y_i)\}_{i=1}^{B}$ satisfies*

$$\mathbb{P}\{Y_{B+1} \in \widehat{C}(X_{B+1})\} \;\geq\; 1 - \alpha - \Big(1 - (1 - \delta_X)^B\Big), \tag{32}$$

where $\delta_X = TV(\hat{P}_X, P)$ is defined in (27). In particular,

$$\mathbb{P}\{Y_{B+1} \in \widehat{C}(X_{B+1})\} \;\geq\; 1 - \alpha - B\,\delta_X. \tag{33}$$

Proof. Let $\mu := \hat{P}^B \otimes P_{\text{test}}$ and $\nu := P_{\text{test}}^B \otimes P_{\text{test}} = P_{\text{test}}^{B+1}$. For any event A, the variational characterization of total variation yields

$$\mu(A) \geq \nu(A) - \mathrm{TV}(\mu, \nu).$$

Applying this to the coverage event $A = \{Y_{B+1} \in \widehat{C}(X_{B+1})\}$ gives

$$\mathbb{P}_{\hat{P}^B \otimes P_{\text{test}}}(A) \geq \mathbb{P}_{P_{\text{test}}^{B+1}}(A) - \mathrm{TV}(\hat{P}^B \otimes P_{\text{test}},\, P_{\text{test}}^B \otimes P_{\text{test}}).$$

Because the test factor P_{test} is identical in both product measures we have:

$$\mathrm{TV}(\hat{P}^B \otimes P_{\text{test}},\, P_{\text{test}}^B \otimes P_{\text{test}}) = \mathrm{TV}(\hat{P}^B, P_{\text{test}}^B).$$

Moreover, by (28) we have $\mathrm{TV}(\hat{P}, P_{\text{test}}) = \delta_X$. A standard tensorization bound for total variation gives

$$\mathrm{TV}(\hat{P}^B, P_{\text{test}}^B) \leq 1 - (1 - \mathrm{TV}(\hat{P}, P_{\text{test}}))^B = 1 - (1 - \delta_X)^B.$$

Finally, under the ideal setting (true weights w), Theorem 1 implies $\mathbb{P}_{P_{\text{test}}^{B+1}}(A) \geq 1 - \alpha$. Combining these proves (32). The linear bound (33) follows from $1 - (1 - \delta_X)^B \leq B\,\delta_X$ for $\delta_X \in [0, 1]$. □

Explicit coverage bounds in terms of weight estimation error. Combining Proposition 4 with Lemma 3 yields directly:

$$\mathbb{P}\{Y_{B+1} \in \widehat{C}(X_{B+1})\} \;\geq\; 1 - \alpha - B\,\mathbb{E}_{X \sim Q}[|\hat{w}(X) - w(X)|], \tag{34}$$

and

$$\mathbb{P}\{Y_{B+1} \in \widehat{C}(X_{B+1})\} \;\geq\; 1 - \alpha - B\sqrt{\mathbb{E}_{X \sim Q}[(\hat{w}(X) - w(X))^2]}. \tag{35}$$

If the uniform relative error condition (30) holds, then (31) and (33) give

$$\mathbb{P}\{Y_{B+1} \in \widehat{C}(X_{B+1})\} \;\geq\; 1 - \alpha - B\,\frac{\varepsilon}{1 - \varepsilon}. \tag{36}$$

Connection to classifier-based ratio estimation. In our implementation we form $\hat{w}$ via a probabilistic classifier $\hat{\eta}(x) \approx \eta(x) := \mathbb{P}(D = 1 \mid X = x)$ that discriminates test from pool covariates. With class prior $\pi := \mathbb{P}(D = 1)$, the implied ratio estimator is

$$\hat{w}(x) = \frac{1 - \pi}{\pi}\,\frac{\hat{\eta}(x)}{1 - \hat{\eta}(x)}. \tag{37}$$

Taking logarithms yields $\log \hat{w}(x) = \log \frac{1-\pi}{\pi} + \mathrm{logit}(\hat{\eta}(x))$. Therefore, uniform control of the logit error, $|\mathrm{logit}(\hat{\eta}(x)) - \mathrm{logit}(\eta(x))| \leq t$, implies a multiplicative bound $e^{-t} \leq \hat{w}(x)/w(x) \leq e^t$, which can be plugged into (36) (for sufficiently small t) to obtain an explicit coverage degradation bound.

D.2 Finite–Pool Approximations

Algorithm 1 proposes covariates by resampling from a finite unlabelled pool $\mathcal{D}_{\text{pool}} = \{X^{(m)}\}_{m=1}^{M}$ with $X^{(m)} \overset{\text{iid}}{\sim} P_{\text{pool},X}$, rather than drawing $X \sim P_{\text{pool},X}$ directly. Here, we describe informally how this *finite–pool approximation* affects the target–marginal coverage guarantee of Theorem 1.

With oracle weights $w(x) = \frac{dP_{\text{test},X}}{dP_{\text{pool},X}}(x)$ and $0 \le w(x) \le W_{\max}$, rejection sampling based on resampling from $\mathcal{D}_{\text{pool}}$ produces accepted covariates from the *weighted empirical* measure:

$$\widehat{P}_M(\cdot) \;:=\; \frac{\sum_{m=1}^{M} w(X^{(m)})\,\mathbf{1}\{X^{(m)} \in \cdot\}}{\sum_{m=1}^{M} w(X^{(m)})}. \tag{38}$$

Consequently, the calibration pairs are exchangeable with a test point drawn from $\widehat{P}_M(dx)\,P_{\text{test}}(dy \mid x)$, whereas the true test point is drawn from $P_{\text{test},X}(dx)\,P_{\text{test}}(dy \mid x)$. Thus, the impact on coverage is determined by how close $\widehat{P}_M$ is to $P_{\text{test},X}$, measured using an appropriate metric (e.g. an integral probability metric over a restricted function class).[4]

More concretely, writing $C_M(\cdot)$ for the conformal set constructed from B accepted labelled pairs, the coverage gap can be expressed as

$$\begin{aligned}&\Big|\mathbb{P}_{(X,Y)\sim P_{\text{test}}}\{Y \in C_M(X)\} - \mathbb{P}_{(X,Y)\sim \widehat{P}_M \otimes P_{\text{test}}(\cdot|X)}\{Y \in C_M(X)\}\Big| \\ &\qquad = \Big|\mathbb{E}_{X\sim P_{\text{test},X}}\big[g_{C_M}(X)\big] - \mathbb{E}_{X\sim \widehat{P}_M}\big[g_{C_M}(X)\big]\Big|\end{aligned}$$

where $g_{C_M}(x) := \mathbb{P}\{Y \in C_M(x) \mid X = x\} \in [0,1]$. Importantly, g_{C_M} depends on the *data-dependent* prediction set C_M, and hence on the underlying conformity score and predictor. Therefore, obtaining *finite* quantitative bounds typically requires additional structure, e.g. regularity/complexity control on the induced class of functions $\{g_{C_M}\}$, which is model– and score–dependent.

Nonetheless, it is straightforward to show that the finite–pool sampling step becomes asymptotically equivalent to sampling from the true target covariate distribution as $M \to \infty$. Indeed, by standard self–normalised importance sampling results, $\widehat{P}_M$ consistently approximates $P_{\text{test},X}$ as the pool size grows:

Proposition 5 (Consistency of the weighted empirical target). *Let $Q := P_{\text{pool},X}$ and $P := P_{\text{test},X}$ with $P \ll Q$, and define $w = \frac{dP}{dQ}$. Let $X^{(1)}, \dots, X^{(M)} \overset{iid}{\sim} Q$ and define $\widehat{P}_M$ by (38). Then, for any bounded measurable $g : \mathcal{X} \to \mathbb{R}$,*

$$\int g(x)\,\widehat{P}_M(dx) \;\xrightarrow[M\to\infty]{a.s.}\; \int g(x)\,P(dx). \tag{39}$$

Proof. By definition,

$$\int g\,d\widehat{P}_M = \frac{\frac{1}{M}\sum_{m=1}^{M} w(X^{(m)})\,g(X^{(m)})}{\frac{1}{M}\sum_{m=1}^{M} w(X^{(m)})}.$$

Since g is bounded and w is essentially bounded, both $w(X)g(X)$ and $w(X)$ are integrable under Q. Hence, by the strong law of large numbers,

$$\frac{1}{M}\sum_{m=1}^{M} w(X^{(m)})\,g(X^{(m)}) \xrightarrow{a.s.} \mathbb{E}_Q[w(X)g(X)], \qquad \frac{1}{M}\sum_{m=1}^{M} w(X^{(m)}) \xrightarrow{a.s.} \mathbb{E}_Q[w(X)] = 1.$$

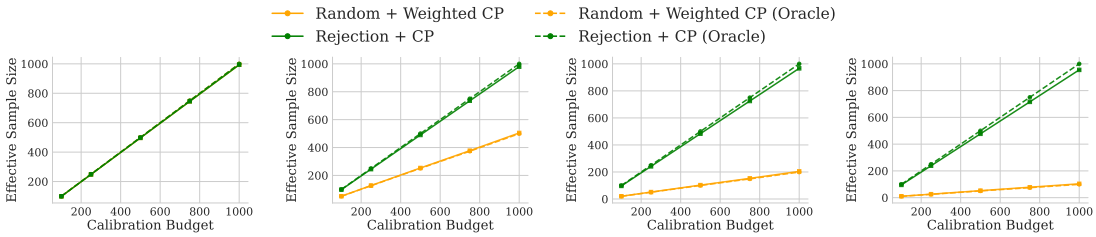
Taking the ratio yields $\int g\,d\widehat{P}_M \xrightarrow{a.s.} \mathbb{E}_Q[w(X)g(X)] = \int g\,dP$, which proves (39). □

4. Note that total variation is generally not informative here when $P_{\text{test},X}$ is continuous, since $\widehat{P}_M$ is discrete.

Proposition 5 shows P_M is a consistent approximation to $P_{\text{test},\,X}$. Under additional regularity assumptions (e.g., continuity/no–ties for the score distribution and stability of the induced prediction-set functional), one expects the resulting loss of coverage from finite–pool resampling to vanish as $M \to \infty$. Moreover, whether the pool is resampled with or without replacement becomes asymptotically irrelevant when $M \gg B$, since repeated selections from the pool occur with vanishing probability.

Appendix E. Additional Results

Figure 2 shows the effective sample sizes for **Rejection + CP** and **Random + Weighted CP**, using both oracle and estimated weights. As expected, **Random + Weighted CP** suffers from a significant loss in effective sample size as the level covariate shift increases, whereas our approach remains stable as a result of directly target the test distribution.



(a) $\gamma = 0$ (b) $\gamma = 0.5$ (c) $\gamma = 0.8$ (d) $\gamma = 0.9$

Figure 2. Average effective sample size for different labelling budgets and levels of covariate shift for our synthetic dataset. Results show the mean $\pm$ standard err. over 250 trials.

Appendix F. Additional Details

Here, we provide full details for the weighted CP procedure described in Section 2.2.

Weighted conformal prediction. Assume the setup in Section 2.1 and define the importance weights:

$$w(x) = \frac{dP_{\text{test},X}}{dP_{\text{pool},X}}(x). \tag{40}$$

Let $\{(X_i, Y_i)\}_{i=1}^{B}$ be calibration pairs and fix a test covariate x. For any candidate label y, define the candidate score $S_{B+1}(y) := s(x, y)$ and calibration scores $S_i := s(X_i, Y_i)$ as normal. Weighted CP (Tibshirani et al., 2019) replaces the usual *unweighted* rank of $S_{B+1}(y)$ among $\{S_1, \ldots, S_B, S_{B+1}(y)\}$ by a *weighted* rank in which each calibration score carries weight $w(X_i)$ and the candidate score carries weight $w(x)$. Concretely, writing $w_i := w(X_i)$ and $w_{B+1} := w(x)$, the (randomised) weighted conformal p–value is

$$p_w(y) = \frac{\sum_{i=1}^{B} w_i \, \mathbf{1}\{S_i \geq S_{B+1}(y)\} + w_{B+1} U}{\sum_{i=1}^{B+1} w_i}, \qquad U \sim \text{Unif}(0, 1), \tag{41}$$

and the weighted conformal prediction set is $C_w(x) := \{y : p_w(y) > \alpha\}$. With known w, this yields the target–marginal guarantee $\mathbb{P}_{(X,Y)\sim P_{\text{test}}}\{Y \in C_w(X)\} \geq 1 - \alpha$ under covariate shift (Tibshirani et al., 2019).

Appendix G. Algorithm Block

Algorithm 1 Density–Based Rejection Sampling for Calibration

Require: Unlabelled pool $D_u = \{x_j\}_{j=1}^{M}$, miscoverage level $\alpha \in (0,1)$, labelling budget B, fixed predictor f, nonconformity score $s(x,y)$, estimate of $\hat{w}(x)$

Ensure: Conformal prediction set function $C(\cdot)$ with target marginal coverage

Initialise labelled calibration set $\mathcal{D}_{\text{cal}} \leftarrow \emptyset$
Set envelope $\hat{W}_{\max} \leftarrow \max_{x_j \in D_u} \max\{0,\, \hat{w}(x_j)\}$
while $|\mathcal{D}_{\text{cal}}| < B$ **do**
    Sample $x \sim \text{Unif}(D_u)$ ▷ sampling $x \sim \text{Unif}(D_u)$ is with replacement
    Sample $u \sim \text{Unif}(0,1)$
    Compute $\tilde{w} \leftarrow \min\{1,\, \hat{w}(x)\}$ ▷ (clip to nonnegative)
    Set acceptance probability $p_{\text{acc}} \leftarrow \min\{1, \tilde{w}/\hat{W}_{\max}\}$
    if $u \le p_{\text{acc}}$ **then**
        Query label $y \sim P(Y \mid X = x)$ and add (x,y) to $\mathcal{D}_{\text{cal}}$
    end if
end while
Compute calibration scores $\{s_i\}_{i=1}^{B}$ where $s_i \leftarrow s(x_i, y_i)$ for $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$
Let $k \leftarrow \lceil (B+1)(1-\alpha) \rceil$ and $q_b \leftarrow s_{(k)}$, where $s_{(k)}$ is the k–th order statistic of $\{s_i\}_{i=1}^{B}$
Define $C(x) \leftarrow \{y \in \mathcal{Y} : s(x,y) \le q_b\}$
return $C(\cdot)$

Remark 6. *Note that in typical regimes with a large pool and a small label budget ($B \ll M$), sampling with or without replacement differs negligibly, so the practical effect is minimal. See also Appendix D.2.*