TOWARDS ROBUST AND EFFICIENT CONFORMAL PREDICTION UNDER COVARIATE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction provides model-agnostic prediction sets with nominal coverage, enabling rigorous uncertainty quantification for machine-learning predictions. A major challenge in conformal prediction is covariate shift, where source and target covariate distributions differ while the conditional outcome distribution remains unchanged. Existing methods often fail to balance robustness and efficiency: weighted conformal prediction (WCP) may yield excessively wide or infinite sets under severe covariate shift, whereas doubly robust prediction (DRP) may substantially undercover because it relies on a single pair of component-model estimates whose finite-sample errors can reinforce one another. We propose doubly propagated calibration (DPC) to attain nominal target coverage while keeping prediction sets informative. DPC repeatedly regenerates the DRP component-model estimates across data splits and propagates the resulting finite-sample estimation uncertainty through downstream calibration. Within each regeneration run, DPC balances source and target samples on estimated coverage probabilities and then takes the union of candidate sets across runs, so that one reliable candidate can protect final coverage. Across severe-shift simulations and California Housing experiments, DPC attains near-nominal 90% coverage in settings where DRP substantially undercovers, with intervals about 40% narrower than WCP in simulations and widths less than 9% above the oracle benchmark on California Housing.

1 INTRODUCTION

Conformal prediction converts a fitted point predictor into a prediction set with finite-sample marginal coverage when calibration and future observations are exchangeable (Vovk et al., 2005; Shafer and Vovk, 2008; Lei et al., 2018). Because its validity comes primarily from calibration rather than correct specification of the outcome distribution, conformal prediction can be combined with complex black-box predictors (Romano et al., 2019; Gibbs and Candes, 2021). It therefore provides a general framework for quantifying uncertainty in machine-learning predictions.

A major practical challenge for conformal prediction is covariate shift, where the source and target populations have different covariate distributions (Shimodaira, 2000; Tibshirani et al., 2019). The target population may place greater mass on regions where the prediction model performs poorly, causing a threshold calibrated on source data to misrepresent the target score distribution. This problem is especially concerning under severe covariate shift, when target-relevant covariate profiles have few labeled source observations (Gibbs and Candes, 2021; Fannjiang et al., 2022; Qiu et al., 2023; Prinster et al., 2022; Hore and Barber, 2025).

Existing methods largely follow two types of approaches. Weighted conformal prediction (WCP) reweights source observations using the target-to-source density ratio (Tibshirani et al., 2019). Under severe covariate shift or limited covariate overlap, estimated weights can concentrate on a few observations, sharply reducing the effective calibration sample size and producing overly wide or even infinite prediction sets (Qiu et al., 2023; Prinster et al., 2022; 2023; Hore and Barber, 2025). WCP may therefore be *robust* (i.e., attaining nominal coverage even under severe covariate shift) but *inefficient* (i.e., producing prediction sets that are too wide to be informative).

Doubly robust prediction (DRP) augments density-ratio weighting with a conditional score-CDF model that estimates conditional coverage at each candidate threshold (Yang et al., 2024). Its population calibration equation remains valid if either nuisance model is correctly specified. However,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

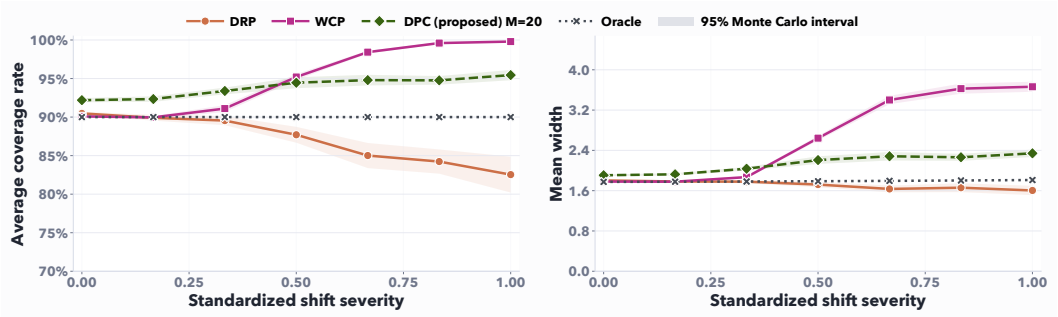


Figure 1: Pooled mean coverage and interval width across the three covariate-shift simulations described in Section 4.1, with DPC using $M = 20$, where M is the number of regeneration runs. Oracle uses the empirical 90% residual-score quantile from the independent new target sample.

this population-level double robustness does not guarantee that a single finite-sample fit is reliable. Conventional DRP relies on one pair of nuisance estimates and carries them directly into threshold selection without propagating their estimation uncertainty through downstream calibration. Its remaining systematic error depends on the product of the density-ratio and conditional score-CDF estimation errors (Chernozhukov et al., 2018; Yang et al., 2024). Under limited covariate overlap, these errors may reinforce one another in target-relevant regions, leading DRP to select a threshold that is too small and produce short but undercovered (i.e., efficient but unrobust) prediction sets.

We propose doubly propagated calibration (DPC), a new calibration principle designed to balance robustness and efficiency. Rather than committing the final prediction set to a single pair of nuisance estimates, DPC repeatedly regenerates the density-ratio and conditional score-CDF estimators across subsamples and fold assignments and propagates the resulting finite-sample uncertainty through the regenerated downstream calibrations (Heng et al., 2026). Within each regeneration run, DPC directly balances estimated conditional coverage between the weighted source and target samples, targeting the quantity most relevant to final coverage rather than requiring globally accurate estimation of the density ratio (Graham et al., 2012; Imai and Ratkovic, 2014; Chan et al., 2016; Zhao, 2019). DPC then takes the union of the resulting nested prediction sets, yielding a *one-success principle* (Guo, 2023; Zheng et al., 2025; Heng et al., 2026): it need not identify the best nuisance fit, because a single candidate with sufficiently small downstream error can protect final coverage. DPC thus replaces reliance on one globally accurate nuisance fit with a propagated collection of candidates in which one sufficiently reliable candidate is enough.

Theoretical results formalize this one-success principle and show how the tilting step improves reliability without excessive width. Across various covariate-shift simulations and California Housing experiments (Pedregosa et al., 2011), DPC maintains near-nominal coverage where DRP substantially undercovers, while avoiding WCP's excessive width inflation in simulations and remaining close to the infeasible oracle width; see Figure 1 for an illustration.

2 Notation and Background

Let X denote the covariates, namely the observed features used to predict a response Y. We distinguish a labeled source domain from a target domain in which predictions will be conducted. Let P_{src} and P_{tar} denote the source and target joint laws of (X, Y), with covariate marginals $P_{\mathrm{src},X}$ and $P_{\mathrm{tar},X}$. We observe labeled source data $\mathcal{D}_{\mathrm{src}} := \{(X_i^{\mathrm{src}}, Y_i^{\mathrm{src}})\}_{i=1}^{n_{\mathrm{src}}}$ and unlabeled target covariates $\mathcal{D}_{\mathrm{tar}} := \{X_j^{\mathrm{tar}}\}_{j=1}^{n_{\mathrm{tar}}}$. The central challenge is that the target covariate distribution may differ from the source covariate distribution ($P_{\mathrm{src},X} \neq P_{\mathrm{tar},X}$), so regions that are common in the target population may be sparsely represented in the labeled source data. A calibration rule based only on the source distribution may therefore fail to reflect uncertainty in the target population.

Assumption 2.1. The target covariate law is absolutely continuous with respect to the source covariate law, $P_{\mathrm{tar},X} \ll P_{\mathrm{src},X}$, and the conditional law of Y given X is the same across the source and target domains (Shimodaira, 2000; Tibshirani et al., 2019). All of the study units are independent.

For a target miscoverage level $\alpha \in (0, 1)$, conformal prediction seeks a prediction set $\mathcal{C}$ satisfying $\mathbb{P}_{\mathrm{tar}}\{Y \in \mathcal{C}(X)\} \geq 1 - \alpha$. A nonnegative score $R(x, y)$ parameterizes nested prediction sets as $\mathcal{C}_\theta(x) := \{y : R(x, y) \leq \theta\}$. Because $\mathcal{C}_\theta(x)$ expands as θ increases, calibration represents choosing

an appropriate threshold θ. For example, for the absolute-residual score $R(x, y) = |y - \widehat{\mu}(x)|$, where $\widehat{\mu}$ is a fitted regression function, the prediction set $\mathcal{C}_\theta(x) = [\widehat{\mu}(x) - \theta, \widehat{\mu}(x) + \theta]$ (Vovk et al., 2005).

Under covariate shift, define the oracle density ratio $w^\star(x) := dP_{\mathrm{tar},X}/dP_{\mathrm{src},X}(x)$, which reweights the source covariate distribution toward the target covariate distribution. For a threshold θ, define the oracle conditional score CDF $m^\star(\theta, x) := \mathbb{P}\{R(X, Y) \le \theta \mid X = x\}$. The function $m^\star(\theta, x)$ is shared across the source and target domains under covariate shift. We use w and m for candidate density-ratio and conditional score-CDF functions, respectively, and $\widehat{w}$ and $\widehat{m}$ for their estimates. Write $\mathbb{E}_{\mathrm{src}}$ and $\mathbb{E}_{\mathrm{tar}}$ for expectations under the source and target laws. The oracle density ratio satisfies the transport identity $\mathbb{E}_{\mathrm{src}}[w^\star(X)h(X)] = \mathbb{E}_{\mathrm{tar}}[h(X)]$ for any $h(X)$, which allows source covariate averages to represent the target population.

For nuisance fitting of $\widehat{w}$ and $\widehat{m}$, we partition the source and target data into disjoint fitting and calibration samples. Specifically, $\mathcal{D}_{\mathrm{src}}$ is divided into $\mathcal{D}_{\mathrm{src}}^{\mathrm{fit}}$ and $\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}$, while $\mathcal{D}_{\mathrm{tar}}$ is divided into $\mathcal{D}_{\mathrm{tar}}^{\mathrm{fit}}$, $\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}}$ and $D_{\mathrm{tar}}^{\mathrm{new}}$. Write $n_{\mathrm{src}}^{\mathrm{fit}} := |\mathcal{D}_{\mathrm{src}}^{\mathrm{fit}}|$, $n_{\mathrm{src}}^{\mathrm{cal}} := |\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}|$, $n_{\mathrm{tar}}^{\mathrm{fit}} := |\mathcal{D}_{\mathrm{tar}}^{\mathrm{fit}}|$, and $n_{\mathrm{tar}}^{\mathrm{cal}} := |\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}}|$. The calibration observations are written as $\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}} = \{(X_i^{\mathrm{src,cal}}, Y_i^{\mathrm{src,cal}})\}_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}$ and $\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}} = \{X_j^{\mathrm{tar,cal}}\}_{j=1}^{n_{\mathrm{tar}}^{\mathrm{cal}}}$. The independent new target dataset $\mathcal{D}_{\mathrm{tar}}^{\mathrm{new}}$ is used to construct prediction sets $\mathcal{C}_\theta(X_j^{\mathrm{tar,new}})$, and the outcomes $Y_j^{\mathrm{tar,new}}$ are never used for nuisance fitting or threshold calibration. To construct prediction sets for all target observations, we cyclically permute the roles of $\mathcal{D}_{\mathrm{tar}}^{\mathrm{fit}}$, $\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}}$, and $\mathcal{D}_{\mathrm{tar}}^{\mathrm{new}}$, yielding prediction sets for all target observations.

2.1 Weighted Conformal Prediction (WCP)

WCP addresses covariate shift by assigning larger calibration weights to source observations that are more common in the target population relative to the source population. It then selects a weighted $1 - \alpha$ quantile of the source calibration scores (Tibshirani et al., 2019). The density-ratio estimator $\widehat{w}$ is fitted using source covariates from $\mathcal{D}_{\mathrm{src}}^{\mathrm{fit}}$ and target covariates from $\mathcal{D}_{\mathrm{tar}}^{\mathrm{fit}}$.

For a candidate density ratio estimate $\widehat{w}$ and target covariate $X_j^{\mathrm{tar,new}}$, WCP defines

$$\widehat{\theta}_{\mathrm{WCP}}\big(X_j^{\mathrm{tar,new}}; \widehat{w}\big) = \inf\left\{\theta : \frac{\sum_{i \in \mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}} \widehat{w}(X_i^{\mathrm{src}}) \mathbf{1}\{R(X_i^{\mathrm{src}}, Y_i^{\mathrm{src}}) \le \theta\}}{\sum_{i \in \mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}} \widehat{w}(X_i^{\mathrm{src}}) + \widehat{w}(X_j^{\mathrm{tar,new}})} \ge 1 - \alpha\right\}. \tag{1}$$

Write $\theta_\alpha^\star$ as the smallest score threshold that attains the desired target-population coverage. As covariate shift becomes more severe, the importance weights $\widehat{w}(X_i^{\mathrm{src}})$ among source calibration observations can become highly concentrated, with a small number of observations receiving much larger weights than the rest. If these high-weight source calibration observations also have large $R(X_i^{\mathrm{src}}, Y_i^{\mathrm{src}})$, then for a small candidate threshold θ, the corresponding indicators $\mathbf{1}\{R(X_i^{\mathrm{src}}, Y_i^{\mathrm{src}}) \le \theta\}$ are zero. To make the weighted cumulative proportion in WCP reach $1 - \alpha = 0.9$, WCP therefore need to increase $\widehat{\theta}_{\mathrm{WCP}}\big(X_j^{\mathrm{tar,new}}; \widehat{w}\big)$, so $\widehat{\theta}_{\mathrm{WCP}}\big(X_j^{\mathrm{tar,new}}; \widehat{w}\big) > \theta_\alpha^\star$. A larger threshold directly enlarges the prediction set, reducing the *efficiency* of WCP (Qiu et al., 2023).

2.2 Doubly Robust Prediction (DRP)

Motivated by improving the efficiency of WCP and related methods, DRP constructs a calibration moment using semiparametric efficiency theory (Yang et al., 2024). The density-ratio estimator $\widehat{w}$ is fitted using $\mathcal{D}_{\mathrm{src}}^{\mathrm{fit}}$ and $\mathcal{D}_{\mathrm{tar}}^{\mathrm{fit}}$, while the conditional score-CDF estimator $\widehat{m}$ is fitted using the labeled source observations in $\mathcal{D}_{\mathrm{src}}^{\mathrm{fit}}$. For each trial threshold θ, DRP then uses $\widehat{m}$ evaluated on $\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}}$ to estimate target coverage and applies an importance-weighted correction using the observed score indicators from $\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}$. Specifically, DRP combines $\widehat{w}$ and $\widehat{m}$ through the empirical moment

$$\widehat{\Psi}_{\mathrm{DRP}}(\theta) := \frac{1}{n_{\mathrm{src}}^{\mathrm{cal}}} \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} \widehat{w}\Big(X_i^{\mathrm{src,cal}}\Big)\Big\{\mathbf{1}\Big[R\Big(X_i^{\mathrm{src,cal}}, Y_i^{\mathrm{src,cal}}\Big) \le \theta\Big] - \widehat{m}\Big(\theta, X_i^{\mathrm{src,cal}}\Big)\Big\} + \frac{1}{n_{\mathrm{tar}}^{\mathrm{cal}}} \sum_{j=1}^{n_{\mathrm{tar}}^{\mathrm{cal}}} \Big\{\widehat{m}\Big(\theta, X_j^{\mathrm{tar,cal}}\Big) - (1 - \alpha)\Big\}. \tag{2}$$

Holding R, $\widehat{w}$, and $\widehat{m}$ fixed, let $\Psi_{\mathrm{DRP}}(\theta)$ denote the corresponding population moment, obtained by replacing the sample averages with their respective expectations. Under covariate shift, it satisfies the product-bias identity

$$\Psi_{\mathrm{DRP}}(\theta) = \mathbb{E}_{\mathrm{tar}}[m^\star(\theta, X)] - (1 - \alpha) - \mathbb{E}_{\mathrm{src}}[\{\widehat{w}(X) - w^\star(X)\}\{\widehat{m}(\theta, X) - m^\star(\theta, X)\}]. \tag{3}$$

Here $\mathbb{E}_{\text{tar}}[m^\star(\theta, X)] - (1-\alpha)$ is the true target coverage gap at threshold θ. Let $\theta^\star_\alpha$ denote the oracle $(1-\alpha)$-quantile of the target score distribution. For $\theta < \theta^\star_\alpha$, this gap is negative because target coverage has not yet reached $1-\alpha$. If either $\widehat{w} = w^\star$ or $\widehat{m} = m^\star$, the product-bias term vanishes, and the population DRP moment recovers the true target coverage gap. This removes nuisance-induced bias but not sampling variability in the empirical moment. We select $\widehat{\theta}_{\text{DRP}}$ as the smallest threshold on the ordered solver grid for which $\widehat{\Psi}_{\text{DRP}}(\theta) \geq 0$.

3 The Proposed Approach: Doubly Propagated Calibration (DPC)

Both $\widehat{w}$ and $\widehat{m}$ in DRP are estimated from finite nuisance-fitting samples. Conventional plug-in DRP calibrates using a single fitted pair $(\widehat{w}, \widehat{m})$, so its threshold $\widehat{\theta}_{\text{DRP}}$ can be sensitive to the particular fitting sample and data split.

DPC instead varies the subsampling and fold assignments to generate nuisance pairs $(\widehat{w}_m, \widehat{m}_m)$, $m = 1, \ldots, M$, each yielding a candidate threshold $\widehat{\theta}^{\text{tilt}}_m$ after the tilting step. This *regeneration step* reflects nuisance-fitting uncertainty (i.e., variability) in downstream calibration and provides multiple opportunities to obtain a candidate with small coverage-relevant error (Section 3.1).

Within each regeneration run, DPC tilts $\widehat{w}_m$ to better align the weighted source calibration sample $\mathcal{D}^{\text{cal}}_{\text{src}}$ with the target calibration covariates in $\mathcal{D}^{\text{cal}}_{\text{tar}}$ along score-CDF directions determined by $\widehat{m}_m$. This *tilting step* addresses imbalance directly relevant to coverage and aims to reduce *transport discrepancy*, namely the nuisance-induced bias in estimating target coverage, thereby keeping candidate thresholds $\widehat{\theta}^{\text{tilt}}_m$ closer to the oracle threshold $\theta^\star_\alpha$ (Section 3.2).

Finally, DPC takes the union of the candidate prediction sets $\mathcal{C}_m$ (i.e., the *union* step), where $\mathcal{C}_m$ denotes the prediction set based on $\widehat{\theta}^{\text{tilt}}_m$. Its robustness follows from a *one-success principle*: a single candidate prediction set $\mathcal{C}_m$ with sufficiently accurate target coverage can protect the union against under-coverage, without requiring DPC to identify that candidate (Section 3.3).

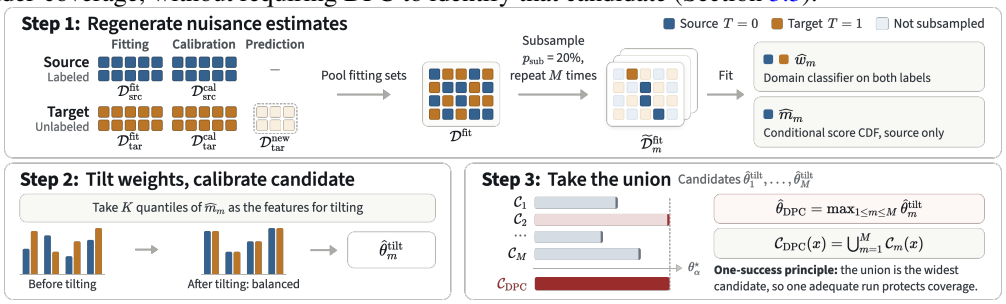


Figure 2: Overview of the proposed Doubly Propagated Calibration (DPC).

3.1 Step 1: Regenerate Nuisance Estimates

The DRP framework already accommodates nonparametric estimation of the density ratio and conditional score CDF (Yang et al., 2024). DPC further uses repeated subsampling to reflect, rather than ignore, finite-sample uncertainty in nuisance estimation. Different fitting samples and data splits can produce different estimates of $\widehat{w}$ and $\widehat{m}$; we refer to repeatedly generating such nuisance estimates as *regeneration* (Heng et al., 2026). In contrast, relying on a single fitted nuisance pair $(\widehat{w}, \widehat{m})$ does not propagate this source of nuisance-fitting uncertainty to downstream calibration.

After fixing $\mathcal{D}^{\text{fit}}_{\text{src}}$, $\mathcal{D}^{\text{cal}}_{\text{src}}$, $\mathcal{D}^{\text{fit}}_{\text{tar}}$, and $\mathcal{D}^{\text{cal}}_{\text{tar}}$, DPC performs M regeneration runs. For each run $m = 1, \ldots, M$, DPC first pools the source and target nuisance-fitting covariates and assigns a domain label $T \in \{0, 1\}$, forming $\mathcal{D}^{\text{fit}} = \{(X^{\text{src}}_i, 0) : (X^{\text{src}}_i, Y^{\text{src}}_i) \in \mathcal{D}^{\text{fit}}_{\text{src}}\} \cup \{(X^{\text{tar}}_j, 1) : X^{\text{tar}}_j \in \mathcal{D}^{\text{fit}}_{\text{tar}}\}$, with $T = 0$ for source observations and $T = 1$ for target observations. The pooled fitting sample has size $|\mathcal{D}^{\text{fit}}| = n^{\text{fit}}_{\text{src}} + n^{\text{fit}}_{\text{tar}}$. DPC then randomly subsamples 20% (or another prespecified subsampling rate) of the observations from $\mathcal{D}^{\text{fit}}$, yielding a run-specific fitting subsample $\widetilde{\mathcal{D}}^{\text{fit}}_m \subseteq \mathcal{D}^{\text{fit}}$, while retaining the corresponding labeled source observations for fitting $\widehat{m}_m$. Thus, different regeneration runs use different fitting subsamples and consequently produce different nuisance estimates, as described in detail in the next two paragraphs.

Density-ratio regeneration. To construct each regenerated density-ratio estimate $\widehat{w}_m$, DPC uses the source and target covariates in $\widetilde{\mathcal{D}}^{\text{fit}}_m$ together with their domain labels. It fits a domain classifier

$\widehat{p}_m(x)$ on $\widetilde{\mathcal{D}}_m^{\mathrm{fit}}$, where $\widehat{p}_m(x)$ estimates $\mathbb{P}(T=1 \mid X=x)$. The fitted classifier $\widehat{e}_m$ is then applied to the fixed source calibration covariates $X_i^{\mathrm{src,cal}}$, $i=1,\dots,n_{\mathrm{src}}^{\mathrm{cal}}$, in $\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}$. Converting $\widehat{p}_m(x)$ to target-to-source odds yields the preliminary density-ratio estimate $\widehat{w}_m(x)$ for regeneration run m.

Conditional score-CDF regeneration. To construct each regenerated conditional score-CDF estimate $\widehat{m}_m$, DPC uses only the source observations in $\widetilde{\mathcal{D}}_m^{\mathrm{fit}}$, because only source observations have outcomes and hence observed conformal scores. Let $\widetilde{\mathcal{D}}_{m,\mathrm{src}}^{\mathrm{fit}}$ denote the source subset of $\widetilde{\mathcal{D}}_m^{\mathrm{fit}}$. Using the source covariates X_i^{src} with $(X_i^{\mathrm{src}},0)\in\widetilde{\mathcal{D}}_{m,\mathrm{src}}^{\mathrm{fit}}$ and the associated conformal scores $R(X_i^{\mathrm{src}},Y_i^{\mathrm{src}})$ from $\mathcal{D}_{\mathrm{src}}^{\mathrm{fit}}$, DPC fits a conditional score-CDF model $\widehat{m}_m(\theta,x)$. The fitted model $\widehat{m}_m$ is then applied to $X_i^{\mathrm{src,cal}}$, $i=1,\dots,n_{\mathrm{src}}^{\mathrm{cal}}$, and $X_j^{\mathrm{tar,cal}}$, $j=1,\dots,n_{\mathrm{tar}}^{\mathrm{cal}}$.

Thus, regeneration run m produces one matched nuisance pair $(\widehat{w}_m,\widehat{m}_m)$. DPC passes this regenerated pair to the subsequent tilting step (as will be described in Step 2), yielding the corresponding candidate threshold $\widehat{\theta}_m^{\mathrm{tilt}}$ and candidate prediction set $\mathcal{C}_m$ based on $\widehat{\theta}_m^{\mathrm{tilt}}$. Repeating this procedure converts finite-sample uncertainty in nuisance fitting into a collection of downstream candidate thresholds that can be propagated through the final union of prediction sets, namely $\bigcup_{m=1}^{M}\mathcal{C}_m$. This mechanism is formalized in Subsection A.1 and illustrated in Figure 3, where propagation improves threshold accuracy as covariate shift becomes more severe.

3.2 Step 2: Tilt Weights and Calibrate Candidate Thresholds

For the nuisance pair $(\widehat{w}_m,\widehat{m}_m)$ obtained in regeneration run m, the corresponding DRP moment $\widehat{\Psi}_{\mathrm{DRP}}(\theta)$ may still exhibit deviation from the true target coverage gap. As shown by the exact decomposition in Appendix Subsection A.1.1, the DRP moment contains a transport discrepancy, namely, the gap between the average estimation error of $\widehat{m}_m$ in the reweighted source sample and that in the target sample:

$$\Delta_m(\theta)=\frac{1}{n_{\mathrm{src}}^{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{w}_m(X_i^{\mathrm{src,cal}})e_m(\theta,X_i^{\mathrm{src,cal}})-\frac{1}{n_{\mathrm{tar}}^{\mathrm{cal}}}\sum_{j=1}^{n_{\mathrm{tar}}^{\mathrm{cal}}}e_m(\theta,X_j^{\mathrm{tar,cal}}),$$

in which $e_m(\theta,x)=m^\star(\theta,x)-\widehat{m}_m(\theta,x)$. When $\Delta_m(\theta)$ shifts the DRP moment upward, the first nonnegative crossing of the DRP moment using $(\widehat{w}_m,\widehat{m}_m)$ may occur below $\theta_\alpha^\star$. DPC therefore applies the tilting step to $\widehat{w}_m$ so that the weighted source calibration sample $\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}$ and the target calibration sample $\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}}$ are balanced on score-CDF features that approximate $e_m(\theta,x)$, to reduce the transport error after the tilting step.

Because $e_m(\theta,x)$ depends on the unknown $m^\star(\theta,x)$, DPC cannot balance this error directly. Instead, DPC constructs observable balance features from $\widehat{m}_m$ at selected score thresholds, whose span is designed to capture the components of $e_m(\theta,x)$ that are most relevant to target coverage. Choose $\tau_{m1}<\cdots<\tau_{mK_m}$ and set

$$g_m(x)=\{1,\widehat{m}_m(\tau_{m1},x),\dots,\widehat{m}_m(\tau_{mK_m},x)\}^{\mathsf{T}}. \tag{4}$$

The intercept targets normalization and the remaining features target score-event features. In regeneration run m, we choose $\tau_{m1}<\cdots<\tau_{mK_m}$ as equally spaced quantiles used to fit $\widehat{m}_m$. For the oracle density ratio, $\mathbb{E}_{\mathrm{src}}[w^\star(X)g_m(X)]=\mathbb{E}_{\mathrm{tar}}[g_m(X)]$.

In the tilting step (Schennach, 2007), with $\lambda=(\lambda_0,\dots,\lambda_{K_m})^{\mathsf{T}}$, DPC computes

$$\widehat{\lambda}_m=\operatorname*{arg\,min}_{\lambda\in\mathbb{R}^{K_m+1}}\left[\frac{1}{n_{\mathrm{src}}^{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{w}_m(X_i^{\mathrm{src,cal}})\exp\left\{\lambda^{\mathsf{T}}g_m(X_i^{\mathrm{src,cal}})\right\}-\lambda^{\mathsf{T}}\frac{1}{n_{\mathrm{tar}}^{\mathrm{cal}}}\sum_{j=1}^{n_{\mathrm{tar}}^{\mathrm{cal}}}g_m(X_j^{\mathrm{tar,cal}})\right] \tag{5}$$

and set the tilted weights

$$\widehat{w}_m^{\mathrm{tilt}}(x)=\widehat{w}_m(x)\exp\{\widehat{\lambda}_m^{\mathsf{T}}g_m(x)\}. \tag{6}$$

Ignoring optimization error, the first-order condition is

$$\frac{1}{n_{\mathrm{src}}^{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{w}_m^{\mathrm{tilt}}(X_i^{\mathrm{src,cal}})g_m(X_i^{\mathrm{src,cal}})-\frac{1}{n_{\mathrm{tar}}^{\mathrm{cal}}}\sum_{j=1}^{n_{\mathrm{tar}}^{\mathrm{cal}}}g_m(X_j^{\mathrm{tar,cal}})=0. \tag{7}$$

We then plug the tilted nuisance pair $(\widehat{w}_m^{\mathrm{tilt}},\widehat{m}_m)$ into the DRP moment in (2), evaluated on $\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}$ and $\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}}$, to obtain the candidate calibration threshold $\widehat{\theta}_m^{\mathrm{tilt}}$ for regeneration run m.

Write $\Delta_m^{\text{tilt}}(\theta)$ for the transport error obtained by replacing $\widehat{w}_m$ with $\widehat{w}_m^{\text{tilt}}$ in $\Delta_m(\theta)$. The role of the tilting step is to reduce the run-specific transport error. If the score-CDF features g_m capture an important component of $e_m(\theta, x)$, exact balance removes the contribution of the component represented by g_m to $\Delta_m^{\text{tilt}}(\theta)$. When the remaining unrepresented error is sufficiently small, the tilting step therefore increases the chance that a propagated regeneration run yields an accurate candidate calibration threshold $\widehat{\theta}_m^{\text{tilt}}$. As a result, fewer propagated regeneration runs may be needed to obtain a reliable candidate. Formal conditions and guarantees are provided in Subsubsections A.2.1–A.2.3. Figure 3 shows the empirical counterpart: DPC with $M = 20$ attains coverage comparable to propagation-only (i.e., DPC without Step 2) $M = 50$, with an evident reduction in threshold error.

3.3 Step 3: Take the union of candidate prediction sets

After the first two steps, regeneration run m produces a candidate calibration threshold $\widehat{\theta}_m^{\text{tilt}}$ and, for each $X_j^{\text{tar,new}} \in \mathcal{D}_{\text{tar}}^{\text{new}}$, the corresponding prediction set $\mathcal{C}_m(X_j^{\text{tar,new}}) = \{y : R(X_j^{\text{tar,new}}, y) \leq \widehat{\theta}_m^{\text{tilt}}\}$. Because the prediction sets are nested in the threshold, $\widehat{\theta}_{m_1}^{\text{tilt}} \leq \widehat{\theta}_{m_2}^{\text{tilt}} \to \mathcal{C}_{m_1}(X_j^{\text{tar,new}}) \subseteq \mathcal{C}_{m_2}(X_j^{\text{tar,new}})$, the union of the M candidate prediction sets can be written as

$$\bigcup_{m=1}^{M} \mathcal{C}_m(X_j^{\text{tar,new}}) = \left\{ y : R(X_j^{\text{tar,new}}, y) \leq \max_{1 \leq m \leq M} \widehat{\theta}_m^{\text{tilt}} \right\}. \tag{8}$$

Consequently, if at least one regeneration run produces a candidate threshold with adequate target coverage, the union of prediction sets has coverage no smaller than that candidate. This yields the one-success principle: DPC requires only one reliable candidate among $\widehat{\theta}_1^{\text{tilt}}, \ldots, \widehat{\theta}_M^{\text{tilt}}$, rather than requiring every regeneration run to be accurate. The final DPC calibration threshold is

$$\widehat{\theta}_{\text{DPC}} = \max_{1 \leq m \leq M} \widehat{\theta}_m^{\text{tilt}}, \tag{9}$$

with the corresponding prediction set $\mathcal{C}_{\text{DPC}}(X_j^{\text{tar,new}}) = \{y : R(X_j^{\text{tar,new}}, y) \leq \widehat{\theta}_{\text{DPC}}\}$. The complete process of DPC is described in Algorithm 1. Repeated regeneration reduces the probability that all candidates miss the success region geometrically (Guo, 2023; Zheng et al., 2025; Heng et al., 2026). This argument is formalized in Subsection A.1.2 and Theorem A.1. Although the union may enlarge the prediction set, Figures 1 and 3 show that the resulting robustness need not entail a substantial efficiency loss: DPC maintains coverage near the nominal 90% level where DRP undercovers, is about 40% narrower than WCP, and remains within roughly 15% of Oracle width.

Algorithm 1 Doubly Propagated Calibration (DPC)

Input: Labeled source sample $\mathcal{D}_{\text{src}}$, unlabeled target sample $\mathcal{D}_{\text{tar}}$, fixed score R, miscoverage level α, number of regeneration runs M, tilting grid counts K_m, and a subsampling rate $p_{\text{sub}} \in (0, 1)$.

1: Split $\mathcal{D}_{\text{src}}$ into $\mathcal{D}_{\text{src}}^{\text{fit}}, \mathcal{D}_{\text{src}}^{\text{cal}}$ and $\mathcal{D}_{\text{tar}}$ into $\mathcal{D}_{\text{tar}}^{\text{fit}}, \mathcal{D}_{\text{tar}}^{\text{cal}}$ and $\mathcal{D}_{\text{tar}}^{\text{new}}$. Keep $\mathcal{D}_{\text{tar}}^{\text{new}}$ separate for prediction. (Step 1)
2: $\mathcal{D}^{\text{fit}} = \{(X_i^{\text{src}}, 0) : (X_i^{\text{src}}, Y_i^{\text{src}}) \in \mathcal{D}_{\text{src}}^{\text{fit}}\} \cup \{(X_j^{\text{tar}}, 1) : X_j^{\text{tar}} \in \mathcal{D}_{\text{tar}}^{\text{fit}}\}$. (Step 1)
3: **for** $m = 1, \ldots, M$ **do**
4: Randomly draw $\widetilde{\mathcal{D}}_m^{\text{fit}} \subseteq \mathcal{D}^{\text{fit}}$ at the prescribed subsampling rate p_{sub}. (Step 1)
5: Fit $\widehat{w}_m$ from domain labels and $\widehat{m}_m$ from labeled source scores in $\widetilde{\mathcal{D}}_m^{\text{fit}}$. (Step 1)
6: Choose K_m score quantiles $\tau_{m1} < \cdots < \tau_{mK_m}$ and construct g_m by (4). (Step 2)
7: Solve (5) for $\widehat{\lambda}_m$ and obtain $\widehat{w}_m^{\text{tilt}}$ by (6). (Step 2)
8: On $\mathcal{D}_{\text{src}}^{\text{cal}}$ and $\mathcal{D}_{\text{tar}}^{\text{cal}}$, obtain $\widehat{\theta}_m^{\text{tilt}}$ as the first nonnegative crossing of (2) using $(\widehat{w}_m^{\text{tilt}}, \widehat{m}_m)$. (Step 2)
9: **end for**
10: Fit baseline $(\widehat{w}, \widehat{m})$ on the full fitting samples and obtain $\widehat{\theta}_{\text{DRP}}$ on $\mathcal{D}_{\text{src}}^{\text{cal}}$ and $\mathcal{D}_{\text{tar}}^{\text{cal}}$ using (2). (Step 3)
11: Set $\widehat{\theta}_{\text{DPC}} = \max_{1 \leq m \leq M} \widehat{\theta}_m^{\text{tilt}}$. (Step 3)
12: **return** $\mathcal{C}_{\text{DPC}}(X_j^{\text{tar,new}}) = \{y : R(X_j^{\text{tar,new}}, y) \leq \widehat{\theta}_{\text{DPC}}\}$ for $X_j^{\text{tar,new}} \in \mathcal{D}_{\text{tar}}^{\text{new}}$. (Step 3)

Theorem 3.1. *Suppose Assumption 2.1 and Conditions A.1–A.2 hold for the tilted candidate thresholds $\widehat{\theta}_m^{tilt}$. Let $\varepsilon_n \to 0$ be a fixed sequence of bias tolerances from the oracle threshold θ^* such that $\epsilon_N \to 0$, where $n = \min\{n_{\text{src}}^{\text{cal}}, n_{\text{tar}}^{\text{cal}}\}$; see Condition A.3 for details. Conditional on the $\mathcal{D}^{\text{fit}}$, assume independent regeneration runs and a per-run success probability of at least $p_{\varepsilon_n} > 0$ on an event of probability $1 - o(1)$ over the fitting and calibration data, as detailed in Condition A.5. Then, for every fixed $\eta > 0$,*

$$\mathbb{P}\left\{ \mathbb{E}_{\text{tar}}\left[m^\star(\widehat{\theta}_{\text{DPC}}, X) \right] < 1 - \alpha - \varepsilon_n - \eta \right\} \leq \exp\{-M p_{\varepsilon_n}\} + o(1). \tag{10}$$

The quantity $\mathbb{E}_{\text{tar}}[m^\star(\widehat{\theta}_{\text{DPC}}, X)]$ is the target-population coverage with the fitting and calibration steps fixed. For fixed p_{ε_n}, the regeneration-failure term decreases exponentially in the number of regeneration runs M. The full statement and proof appear in Theorem A.1 and Appendix B.3.

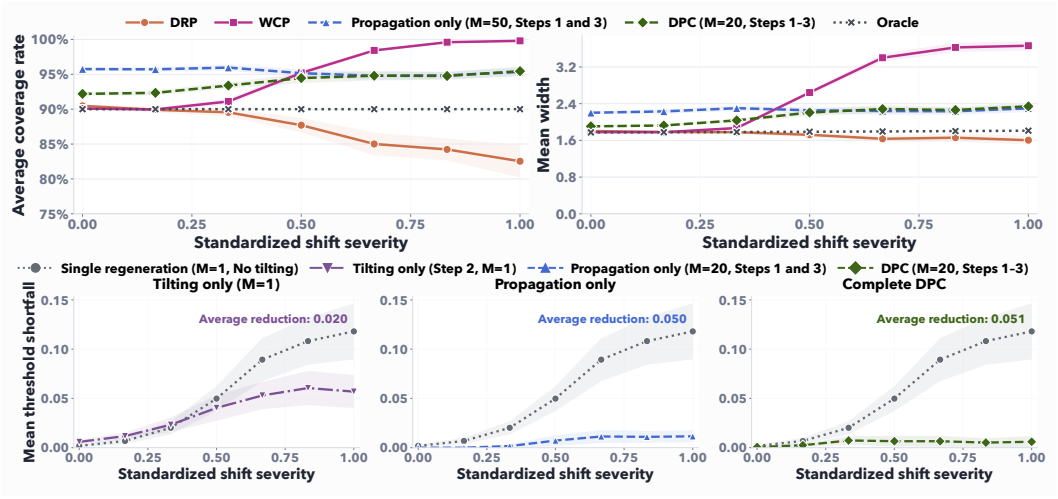


Figure 3: Top: Pooled propagation-only (DRP without Step 2) with $M = 50$ versus DPC (Steps 1–3) with $M = 20$, DRP, WCP, and Oracle. Bottom: reductions in mean absolute threshold error from the tilting step (Step 2 with $M = 1$), propagation (Steps 1 and 3), and both (Steps 1–3). Pooled results across the three covariate-shift simulations. The simulation setup is described in Section 4.1.

4 Numerical Studies

4.1 Simulation Studies

In each Monte Carlo replication, we generate a labeled source sample $\mathcal{D}_{\text{src}}$ of size 1000, an unlabeled target sample $\mathcal{D}_{\text{tar}}$ of size 1000, and an independent new target sample $\mathcal{D}_{\text{tar}}^{\text{new}}$ of size 1000 for prediction and coverage calculation. For DRP and DPC, $\mathcal{D}_{\text{src}}$ and $\mathcal{D}_{\text{tar}}$ are each randomly divided in half, giving $|\mathcal{D}_{\text{src}}^{\text{fit}}| = |\mathcal{D}_{\text{src}}^{\text{cal}}| = |\mathcal{D}_{\text{tar}}^{\text{fit}}| = |\mathcal{D}_{\text{tar}}^{\text{cal}}| = 500$. The fitting samples are used to estimate the nuisance functions, while $\mathcal{D}_{\text{src}}^{\text{cal}}$ and $\mathcal{D}_{\text{tar}}^{\text{cal}}$ are used to determine the calibration threshold θ. We use 30 Monte Carlo replications and set $1 - \alpha = 0.90$. DRP, WCP, and DPC use XGBoost nuisance learners (Chen and Guestrin, 2016). DPC uses $M \in \{20, 30\}$ regeneration runs and $K_m = 31$ score-CDF balance features in each regeneration run. In the following tables and figures, Oracle denotes the infeasible minimum interval width required to attain 90% coverage. All three synthetic designs first construct the source and target shift in the latent covariates $U \in \mathbb{R}^5$, and then apply the same coordinatewise nonlinear transformation $U \mapsto X$ to obtain the observed covariates X.

Additive mean shift. The source latent covariates satisfy $U_{\text{src}} \sim \mathcal{N}(0, I_5)$, where I_5 is the 5×5 identity matrix, so the five latent coordinates have unit variance and are mutually uncorrelated. The target latent covariates satisfy $U_{\text{tar}} \sim \mathcal{N}(\mu, I_5)$, where $\mu = (\|\mu\|_2/\sqrt{5})\mathbf{1}_5$ and $\|\mu\|_2 \in \{0, 1, \ldots, 6\}$. Thus, the difficulty of the additive design is indexed by the magnitude of the latent mean shift $\|\mu\|_2$.

Variance-tail shift. The variance-tail design extends the additive mean shift by allowing the latent covariates to come from a two-component Gaussian mixture. The source latent covariates satisfy $U_{\text{src}} \sim 0.95\mathcal{N}(0, I_5) + 0.05\mathcal{N}(0, 1.8^2 I_5)$. The target latent covariates satisfy $U_{\text{tar}} \sim \left(0.95 - \frac{0.03}{4.15}\|\mu\|_2\right)\mathcal{N}(\mu, I_5) + \left(0.05 + \frac{0.03}{4.15}\|\mu\|_2\right)\mathcal{N}(\mu, 1.8^2 I_5)$, where $\mu = (\|\mu\|_2/\sqrt{5})\mathbf{1}_5$.

Correlation-interaction shift. The source latent covariates satisfy $U_{\text{src}} \sim 0.92\mathcal{N}(0, I_5) + 0.08\mathcal{N}(0, \Sigma_{\text{corr}})$, where $\Sigma_{\text{corr}} = 0.60 I_5 + 0.40\mathbf{1}_5\mathbf{1}_5^{\top}$. Thus, both components have unit marginal variance, but the second component has pairwise correlation 0.4 between any two latent coordinates. The target latent covariates satisfy $U_{\text{tar}} \sim \left(0.92 - \frac{0.01}{4.15}\|\mu\|_2\right)\mathcal{N}(\mu, I_5) + \left(0.08 + \frac{0.01}{4.15}\|\mu\|_2\right)\mathcal{N}(\mu, \Sigma_{\text{corr}})$, where $\mu = (\|\mu\|_2/\sqrt{5})\mathbf{1}_5$.

Appendix D.1 provides the complete details of the above simulation scenarios. For concise reporting, results are presented in the format average coverage rate/average interval width.

Simulation results under the additive mean shift: At the covariate shift endpoint, DRP, WCP, and Oracle give 82.3%/1.59, 99.9%/3.64, and width 1.81; DPC gives 96.0%/2.35 for $M = 20$ and 96.4%/2.44 for $M = 30$. Thus, DPC gains 13.7–14.1 coverage points over DRP, with widths 29.5%–34.9% above Oracle and 32.8%–35.5% below WCP (Figure 4).

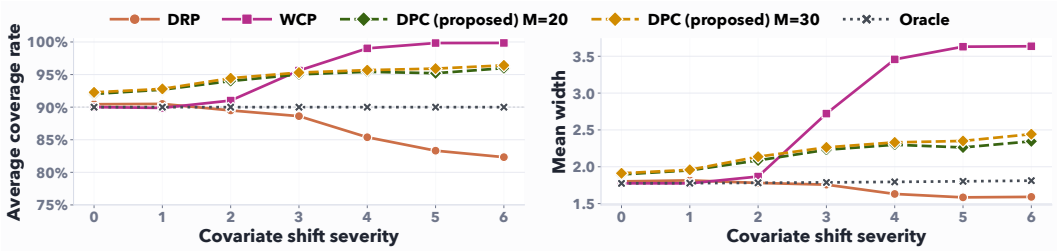


Figure 4: Average coverage and width under additive mean shift.

Simulation results under the variance-tail shift: At the covariate shift endpoint, DRP, WCP, and Oracle give $83.5\%/1.68$, $99.8\%/3.65$, and width 1.81; DPC gives $95.8\%/2.38$ and $96.6\%/2.50$. DPC gains 12.3–13.1 coverage points, with widths 31.5%–38.0% above Oracle and 31.6%–34.8% below WCP (Figure 5).

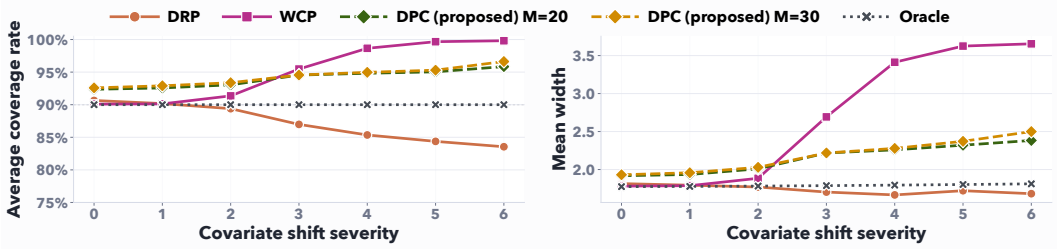


Figure 5: Average coverage and width under variance-tail shift.

Simulation results under the correlation-interaction shift: At the covariate shift endpoint, DRP, WCP, and Oracle give $81.7\%/1.54$, $99.7\%/3.70$, and width 1.81; DPC gives $94.5\%/2.30$ and $94.7\%/2.33$. DPC is within 4.7 points of nominal, gains 12.8–13.0 points over DRP, and stays 37.0%–37.9% narrower than WCP (Figure 6).

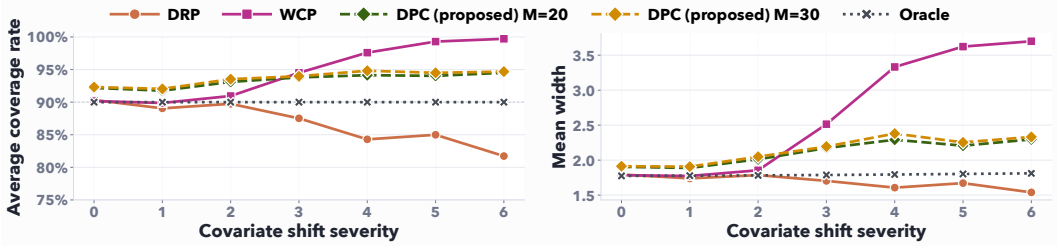


Figure 6: Average coverage and width under correlation-interaction shift.

Table 1 additionally reports WSCPS, CWCP, and WR-CP (Jonkers et al., 2024; Xu et al., 2025), all of which are extensions of WCP. These additional comparisons show that WSCPS and CWCP report infinite mean interval widths under moderate and severe shift, whereas the fixed-score WR-CP implementation substantially undercovers, with coverage of only 77.2%–89.7%. In these settings, DPC maintains finite mean widths of 2.2–2.5, compared with Oracle widths of 1.8, although its coverage is somewhat conservative.

Table 1: Additional simulation baselines over 30 Monte Carlo replications ($\alpha = 0.10$): target coverage (%)/mean interval width; DPC uses $M = 30$.

Setting	Severity	WSCPS	CWCP	WR-CP	**DPC**	Oracle Width
Additive	No shift	90.3 / 1.8	88.8 / 1.7	90.0 / 1.8	**92.3 / 1.9**	1.8
Additive	Moderate ($\ell_2 = 3$)	97.0 / ∞	99.7 / ∞	89.2 / 1.8	**95.3 / 2.3**	1.8
Additive	Hardest ($\ell_2 = 6$)	100.0 / ∞	100.0 / ∞	82.0 / 1.6	**96.4 / 2.4**	1.8
Variance-tail	No shift	90.3 / 1.8	89.4 / 1.7	89.9 / 1.8	**92.6 / 1.9**	1.8
Variance-tail	Moderate ($\ell_2 = 3$, tail +2.2 pp)	96.8 / ∞	99.8 / ∞	89.4 / 1.8	**94.6 / 2.2**	1.8
Variance-tail	Hardest ($\ell_2 = 6$, tail +4.3 pp)	100.0 / ∞	100.0 / ∞	82.7 / 1.7	**96.6 / 2.5**	1.8
Corr.-interaction	No shift	90.5 / 1.8	89.5 / 1.8	90.2 / 1.8	**92.3 / 1.9**	1.8
Corr.-interaction	Moderate ($\ell_2 = 3$, corr. +0.7 pp)	96.0 / ∞	99.8 / ∞	89.7 / 1.8	**94.0 / 2.2**	1.8
Corr.-interaction	Hardest ($\ell_2 = 6$, corr. +1.4 pp)	99.9 / ∞	100.0 / ∞	77.2 / 1.5	**94.7 / 2.3**	1.8

4.2 California Housing Illustration

The California Housing data contain 20,640 census block groups, with eight covariates and median house value as the response (Pace and Barry, 1997; Pedregosa et al., 2011). In each Monte Carlo replication, we first randomly partition the full dataset into four disjoint pools for point-prediction model fitting, labeled source calibration, unlabeled target calibration, and prediction on new target observations. We sample 1000 observations from each pool. The point-prediction model is fitted using the first source pool, and absolute-residual conformal scores are then computed for the labeled source sample. Following the notation introduced above, DRP and DPC further split the labeled source sample $\mathcal{D}_{\mathrm{src}}$ and the unlabeled target sample $\mathcal{D}_{\mathrm{tar}}$ into $\mathcal{D}_{\mathrm{src}}^{\mathrm{fit}}$, $\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}$, $\mathcal{D}_{\mathrm{tar}}^{\mathrm{fit}}$, $\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}}$, with each subset containing 500 observations. The independent new target sample $\mathcal{D}_{\mathrm{tar}}^{\mathrm{new}}$ is used for prediction.

Let $\omega_{\mathrm{tar}}^{(s,\gamma)}(x) = \omega_{\mathrm{src}}(x)\exp\{0.70\gamma d_s(x)\}$, denotes the sampling weight assigned to covariate profile x in the target population under shift setting s and difficulty level γ. Relative to the source sampling weight $\omega_{\mathrm{src}}(x)$, the factor $\exp\{0.70\gamma d_s(x)\}$ changes the sampling probability of different covariate profiles in the target sample. The parameter $\gamma \geq 0$ controls the overall severity of the source and target shift. When $\gamma = 0$, $\omega_{\mathrm{tar}}^{(s,0)}(x) = \omega_{\mathrm{src}}(x)$, so the source and target populations follow the same sampling rule. As γ increases, the target sampling weights become more sensitive to differences in $d_s(x)$, inducing a more severe covariate shift. The function $d_s(x)$ determines the direction of the shift, where $s \in \{\mathrm{add}, \mathrm{var}, \mathrm{corr}\}$ corresponds to the Additive, Variance-tail, and Correlation-interaction settings, respectively. For a given setting, observations with larger $d_s(x)$ receive larger target sampling weights. Appendix D.2 provides the detailed construction of the covariate laws.

The simulations show that WCP can achieve nominal or conservative coverage at the cost of excessively wide intervals. The California Housing results further show that WCP with estimated density ratios and DRP can both fall below nominal coverage at moderate severity. At moderate severity, DPC covers 92.8%–92.9%, compared with 85.9%–88.1% for DRP and 87.5%–88.8% for WCP. At the hardest severity level, DRP covers 78.7%–84.8% and WCP covers 90.6%–95.6%, while DPC remains at 92.7%–93.7%. However, WCP attains this nominal or conservative coverage with infinite mean interval widths in all three hardest-shift settings; it also yields an infinite mean width under moderate correlation-interaction shift. DPC gains 8.8, 10.0, and 14.0 percentage points over DRP in the additive, variance-tail, and correlation-interaction settings, respectively; its corresponding widths of 4.36, 4.43, and 4.86 are only 32.7%, 31.7%, and 32.7% above Oracle.

Table 2: California Housing over 30 paired replications ($\alpha = 0.10$); DPC uses $M = 60$.

Setting	Severity	DRP	WCP	**DPC**	Oracle Width
Additive	No shift ($\gamma = 0$)	90.0 / 1.70	89.9 / 1.67	**93.4 / 2.08**	1.67
Additive	Mild shift ($\gamma = 1$)	89.9 / 2.14	89.1 / 2.02	**93.3 / 2.57**	2.12
Additive	Moderate shift ($\gamma = 2$)	87.8 / 2.63	87.5 / 2.49	**92.8 / 3.32**	2.79
Additive	Hardest shift ($\gamma = 3$)	84.8 / 2.94	90.6 / ∞	**93.6 / 4.36**	3.29
Variance-tail	No shift ($\gamma = 0$)	90.0 / 1.70	89.9 / 1.67	**93.4 / 2.08**	1.67
Variance-tail	Mild shift ($\gamma = 1$)	89.8 / 2.15	89.2 / 2.06	**93.2 / 2.62**	2.14
Variance-tail	Moderate shift ($\gamma = 2$)	88.1 / 2.78	87.7 / 2.59	**92.8 / 3.38**	2.87
Variance-tail	Hardest shift ($\gamma = 3$)	83.6 / 2.93	91.8 / ∞	**93.7 / 4.43**	3.36
Corr.-interaction	No shift ($\gamma = 0$)	90.0 / 1.70	89.9 / 1.67	**93.4 / 2.08**	1.67
Corr.-interaction	Mild shift ($\gamma = 1$)	89.1 / 2.19	88.1 / 2.03	**92.8 / 2.64**	2.22
Corr.-interaction	Moderate shift ($\gamma = 2$)	85.9 / 2.83	88.8 / ∞	**92.9 / 4.06**	3.17
Corr.-interaction	Hardest shift ($\gamma = 3$)	78.7 / 3.00	95.6 / ∞	**92.7 / 4.86**	3.66

5 Discussion

We introduced doubly propagated calibration (DPC) to improve the robustness-efficiency trade-off of conformal prediction under covariate shift. DPC propagates finite-sample nuisance-estimation uncertainty by repeatedly regenerating nuisance estimates and carrying them through downstream calibration, with exponential tilting improving each candidate calibration (Steps 1 and 2). It then aggregates the resulting candidate calibration thresholds by taking the union of their nested prediction sets (Step 3). Our theory shows how one reliable candidate calibration threshold can protect target coverage, while the numerical results demonstrate that DPC avoids both the substantial under-coverage of DRP and the excessive widths of weighted conformal methods, including WCP and its extensions. More broadly, DPC suggests propagating uncertainty in learned nuisance components through downstream calibration rather than relying on a single nuisance fit, as conventional DRP does. This principle may inform conformal prediction procedures involving other estimated nuisances, such as missingness or censoring mechanisms, treatment propensities, and outcome regressions. Extending this principle to these settings are promising future directions.

AI USE STATEMENT

In this work, a generative AI assistant was used to check compliance with the ICLR 2027 template, edit and condense the manuscript language, and refine the LaTeX formatting, including the presentation of cited baseline methods and reported results.

REPRODUCIBILITY STATEMENT

The DPC implementation is available in our anonymous code repository.

ETHICS STATEMENT

The work studies uncertainty calibration using synthetic data and a public census-block-group benchmark; it does not recruit participants or use individual-level human-subject data.

REFERENCES

Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic Learning in a Random World*, New York: Springer.

Shafer, G. and Vovk, V. (2008), "A tutorial on conformal prediction," *Journal of Machine Learning Research*, 9, 371–421.

Lei, J. and Wasserman, L. (2014), "Distribution-free prediction bands for non-parametric regression," *Journal of the Royal Statistical Society: Series B*, 76, 71–96.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018), "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, 113, 1094–1111.

Tibshirani, R. J., Barber, R. F., Candes, E. J., and Ramdas, A. (2019), "Conformal prediction under covariate shift," in *Advances in Neural Information Processing Systems 32*, 2526–2536.

Romano, Y., Patterson, E., and Candes, E. J. (2019), "Conformalized quantile regression," in *Advances in Neural Information Processing Systems 32*.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021a), "Predictive inference with the jackknife+," *The Annals of Statistics*, 49, 486–507.

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

Politis, D. N. and Romano, J. P. (1994), "Large sample confidence regions based on subsamples under minimal assumptions," *The Annals of Statistics*, 22(4), 2031–2050.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2024), "Robust validation: Confident predictions even when distributions shift," *Journal of the American Statistical Association*, 119, 3033–3044.

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

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846–866.

Bang, H. and Robins, J. M. (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–973.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. M. (2018), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2024), "Doubly robust calibration of prediction sets under covariate shift," *Journal of the Royal Statistical Society: Series B*, 86, 943–965.

Hainmueller, J. (2012), "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, 20, 25–46.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Imai, K. and Ratkovic, M. (2014), "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B*, 76, 243–263.

Graham, B. S., Pinto, C. C. X., and Egel, D. (2012), "Inverse probability tilting for moment condition models with missing data," *Review of Economic Studies*, 79, 1053–1079.

van der Laan, M. J. and Rubin, D. (2006), "Targeted maximum likelihood learning," *The International Journal of Biostatistics*, 2, Article 11.

Heng, S., Shen, Y., and Guo, Z. (2026), "Propensity score propagation: A general framework for design-based inference with unknown propensity scores," arXiv:2601.13150.

Shimodaira, H. (2000), "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, 90, 227–244.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007), "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, 8, 985–1005.

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

Pace, R. K. and Barry, R. (1997), "Sparse spatial autoregressions," *Statistics & Probability Letters*, 33, 291–297.

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

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016), "Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting," *Journal of the Royal Statistical Society: Series B*, 78, 673–700.

Chen, R., Chen, G., and Yu, M. (2023), "Entropy balancing for causal generalization with target sample summary information," *Biometrics*, 79, 3179–3190.

Zhao, Q. (2019), "Covariate balancing propensity score by tailored loss functions," *The Annals of Statistics*, 47, 965–993.

Wang, Y. and Zubizarreta, J. R. (2020), "Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations," *Biometrika*, 107, 93–105.

Zubizarreta, J. R. (2015), "Stable weights that balance covariates for estimation with incomplete outcome data," *Journal of the American Statistical Association*, 110, 910–922.

Jonkers, J., Van Wallendael, G., Duchateau, L., and Van Hoecke, S. (2024), "Conformal predictive systems under covariate shift," in *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, *Proceedings of Machine Learning Research*, 230, 406–423; https://proceedings.mlr.press/v230/jonkers24a.html; arXiv:2404.15018.

Wang, J. and Goel, S. (2026), "Weight clipping for robust conformal inference under unbounded covariate shifts," arXiv:2605.02072v1; https://arxiv.org/abs/2605.02072.

Xu, R., Chen, C., Sun, Y., Venkitasubramaniam, P., and Xie, S. (2025), "Wasserstein-regularized conformal prediction under general distribution shift," in *The Thirteenth International Conference on Learning Representations*; https://openreview.net/forum?id=aJ3tiX1Tu4; arXiv:2501.13430v2.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35(2), 634–672.

Chen, T. and Guestrin, C. (2016), "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Gibbs, I. and Candes, E. (2021), "Adaptive conformal inference under distribution shift," in *Advances in Neural Information Processing Systems 34*, 1660–1672.

Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J., and Jordan, M. I. (2022), "Conformal prediction under feedback covariate shift for biomolecular design," *Proceedings of the National Academy of Sciences*, 119(43), e2204569119.

Prinster, D., Liu, A., and Saria, S. (2022), "JAWS: Auditing predictive uncertainty under covariate shift," in *Advances in Neural Information Processing Systems 35*, 35907–35920.

Prinster, D., Saria, S., and Liu, A. (2023), "JAWS-X: Addressing efficiency bottlenecks of conformal prediction under standard and feedback covariate shift," in *Proceedings of the 40th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, 202, 28167–28190.

Zheng, M., Bonvini, M., and Guo, Z. (2025), "Perturbed double machine learning: Nonstandard inference beyond the parametric length," arXiv:2511.01222.

Kiyani, S., Pappas, G. J., and Hassani, H. (2024), "Length optimization in conformal prediction," in *Advances in Neural Information Processing Systems 37*, 99519–99563.

Hore, R. and Barber, R. F. (2025), "Conformal prediction with local weights: Randomization enables robust guarantees," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87, 549–578.

Guo, Z. (2023), "Causal inference with invalid instruments: Post-selection problems and a solution using searching and sampling," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3), 959–985.

A THEORY FOR DOUBLY PROPAGATED CALIBRATION (DPC)

The propagation argument is presented first because it does not depend on the particular nuisance learners or on the comparison of the tilting step. Throughout Subsection A.1, fix $q \in \{\text{prop}, \text{tilt}\}$, where $q = \text{prop}$ denotes propagation-only and $q = \text{tilt}$ includes the tilting step. Regeneration run $m \in \{1, \dots, M\}$ produces the nuisance pair $(\widehat{w}_m, \widehat{m}_m)$, consisting of the regenerated density-ratio estimate $\widehat{w}_m$ and the regenerated conditional score-CDF estimate $\widehat{m}_m$. For $q = \text{prop}$, the DRP moment uses the nuisance pair $(\widehat{w}_m, \widehat{m}_m)$. For $q = \text{tilt}$, the DRP moment uses the tilted nuisance pair $(\widehat{w}_m^{\text{tilt}}, \widehat{m}_m)$, where $\widehat{w}_m^{\text{tilt}}$ denotes the tilted weights. The notation $\widehat{w}_m^q$ denotes $\widehat{w}_m$ for $q = \text{prop}$ and $\widehat{w}_m^{\text{tilt}}$ for $q = \text{tilt}$. The candidate calibration threshold $\widehat{\theta}_m^q$ is selected by the first nonnegative crossing of the DRP moment on the ordered solver grid $\mathcal{S}_{m,n}$. Similarly, the transport discrepancy $\Delta_m^q(\theta)$ denotes $\Delta_m(\theta)$ for $q = \text{prop}$ and $\Delta_m^{\text{tilt}}(\theta)$ for $q = \text{tilt}$. The propagation mechanism itself is the repeated regeneration of nuisance pairs $(\widehat{w}_m, \widehat{m}_m)$. No comparison specific to the tilting step is used in this subsection. For compactness, the appendix writes empirical source and target averages through the operators

$$\mathbb{P}_{\text{src},n}^{w} h := \frac{1}{n_{\text{src}}^{\text{cal}}} \sum_{i=1}^{n_{\text{src}}^{\text{cal}}} w(X_i^{\text{src,cal}}) h(X_i^{\text{src,cal}}, Y_i^{\text{src,cal}}), \qquad \mathbb{P}_{\text{tar},n} h := \frac{1}{n_{\text{tar}}^{\text{cal}}} \sum_{j=1}^{n_{\text{tar}}^{\text{cal}}} h(X_j^{\text{tar,cal}}). \quad (11)$$

Here w is any scalar multiplier, and h may be real- or vector-valued. When a source function depends only on X, its unused Y argument is suppressed; $\mathbb{P}_{\text{src},n}^{1}$ is the ordinary unweighted source average. The main text retains explicit summation notation for readability, whereas the appendix uses these compact operators throughout. Unless another sample is stated explicitly, they refer to the source calibration sample $\mathcal{D}_{\text{src}}^{\text{cal}}$ and target calibration sample $\mathcal{D}_{\text{tar}}^{\text{cal}}$.

Recall the oracle conditional score CDF

$$m^\star(\theta, x) = \mathbb{P}\{R(X, Y) \le \theta \mid X = x\}.$$

The target-population coverage of the prediction set $\mathcal{C}_\theta(x)$ is $\mathbb{E}_{\text{tar}}[m^\star(\theta, X)]$. The conditional score-CDF estimation error is $e_m(\theta, x) = m^\star(\theta, x) - \widehat{m}_m(\theta, x)$.

A.1 PROPAGATION ROBUSTNESS

A.1.1 PROPAGATION IN THE DRP MOMENT AND SUCCESSFUL CANDIDATES

For regeneration run m, the DRP moment in (2), evaluated using $(\widehat{w}_m^q, \widehat{m}_m)$, is

$$\widehat{\Psi}_{\text{DRP}}(\theta; \widehat{w}_m^q, \widehat{m}_m) = \mathbb{P}_{\text{src},n}^{\widehat{w}_m^q} \{\mathbf{1}\{R(X, Y) \le \theta\} - \widehat{m}_m(\theta, X)\} + \mathbb{P}_{\text{tar},n} \{\widehat{m}_m(\theta, \cdot) - (1 - \alpha)\}. \quad (12)$$

The candidate calibration threshold is selected by the first nonnegative crossing on the ordered solver grid,

$$\widehat{\theta}_m^q = \inf\{\theta \in \mathcal{S}_{m,n} \cup \{+\infty\} : \widehat{\Psi}_{\text{DRP}}(\theta; \widehat{w}_m^q, \widehat{m}_m) \ge 0\}.$$

Adding and subtracting the oracle conditional score CDF $m^\star(\theta, x)$ in the source and target terms gives the exact decomposition

$$\widehat{\Psi}_{\text{DRP}}(\theta; \widehat{w}_m^q, \widehat{m}_m) = Z_m^q(\theta) + G_n(\theta) + \Delta_m^q(\theta), \quad (13)$$

where

$$Z_m^q(\theta) = \mathbb{P}_{\text{src},n}^{\widehat{w}_m^q} [\mathbf{1}\{R(X, Y) \le \theta\} - m^\star(\theta, X)], \quad (14)$$

$$G_n(\theta) = \mathbb{P}_{\text{tar},n} m^\star(\theta, \cdot) - (1 - \alpha), \quad (15)$$

$$\Delta_m^q(\theta) = \mathbb{P}_{\text{src},n}^{\widehat{w}_m^q} \{m^\star(\theta, \cdot) - \widehat{m}_m(\theta, \cdot)\} - \mathbb{P}_{\text{tar},n} \{m^\star(\theta, \cdot) - \widehat{m}_m(\theta, \cdot)\}. \quad (16)$$

The decomposition separates the roles of the three terms. The empirical target coverage gap $G_n(\theta)$ is common to all candidate calibration thresholds because the target covariates, the oracle conditional score CDF $m^\star$, and the nominal coverage level $1 - \alpha$ are fixed across runs. The zero of the empirical target coverage gap $G_n(\theta)$ determines the target score quantile. The fluctuation $Z_m^q(\theta)$ can change numerically when propagation changes the weights $\widehat{w}_m^q$, but it remains conditionally centered given the calibration covariates and nuisance-fitting data; it is handled by uniform concentration rather than by selecting a particular regeneration run. In contrast, $\Delta_m^q(\theta)$ depends directly on the nuisance pairs $(\widehat{w}_m, \widehat{m}_m)$, need not be centered, and can systematically move the first nonnegative crossing downward or upward. Propagation therefore acts on the distribution of Δ_m^q across candidates. The DRP moment $\widehat{\Psi}_{\text{DRP}}(\theta; \widehat{w}_m^q, \widehat{m}_m)$ is the sum of these components, and the first nonnegative crossing rule maps their combined effect into $\widehat{\theta}_m^q$.

The following conditions isolate the centered empirical-process terms from the candidate-specific transport discrepancy.

Condition A.1. For every regeneration run under consideration, the weights $\widehat{w}_m^q$ are nonnegative and uniformly bounded by a fixed constant $W < \infty$. The regenerated conditional score-CDF estimate $\widehat{m}_m(\theta, x)$ takes values in $[0, 1]$ and is nondecreasing and right-continuous in θ. Conditional on the source covariates and the fitting data, the source calibration scores are independent. The weights $\widehat{w}_m^q$ may use fitting data and calibration covariates but do not use source calibration outcomes; the nuisance pairs $(\widehat{w}_m, \widehat{m}_m)$ are fitted using data independent of the calibration outcomes, and target outcomes are never used.

Condition A.2. Uniformly over the M regeneration runs and the thresholds on the ordered solver grids $\mathcal{S}_{m,n}$,

$$\max_{1\le m\le M} \sup_{\theta\in\mathcal{S}_{m,n}} |Z_m^q(\theta)| = o_{\mathbb{P}}(1), \tag{17}$$

and

$$\sup_{\theta\in\mathbb{R}} |G_n(\theta) - \{\mathbb{E}_{\mathrm{tar}}[m^\star(\theta, X)] - (1-\alpha)\}| = o_{\mathbb{P}}(1). \tag{18}$$

A standard sufficient regime under honest calibration and bounded weights $\widehat{w}_m^q$ is $\log M / n_{\mathrm{src}}^{\mathrm{cal}} \to 0$ and $n_{\mathrm{tar}}^{\mathrm{cal}} \to \infty$, together with uniform empirical-process control for the data-adaptive weights and ordered solver grids.

Condition A.3. Let $\varepsilon_n \ge 0$. A finite candidate calibration threshold $\widehat{\theta}_m^q$ is successful when

$$\sup_{\theta\in\mathcal{S}_{m,n}} |\Delta_m^q(\theta)| + \left[-\widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}_m^q; \widehat{w}_m^q, \widehat{m}_m)\right]_+ \le \varepsilon_n, \tag{19}$$

where $[z]_+ = \max\{z, 0\}$. Under the exact first nonnegative crossing used for (12), the numerical-deficit term is zero, so success reduces to

$$\sup_{\theta\in\mathcal{S}_{m,n}} |\Delta_m^q(\theta)| \le \varepsilon_n.$$

A candidate calibration threshold $\widehat{\theta}_m^q = +\infty$ is successful by convention because the corresponding target-population coverage is one.

Proposition A.1. *Suppose Assumption 2.1 and Conditions A.1– A.2 hold. If the candidate calibration threshold $\widehat{\theta}_m^q$ is successful at tolerance ε_n, then*

$$\mathbb{E}_{\mathrm{tar}}\left[m^\star(\widehat{\theta}_m^q, X)\right] \ge 1 - \alpha - \varepsilon_n - o_{\mathbb{P}}(1). \tag{20}$$

The proposition follows directly by evaluating (13) at $\widehat{\theta}_m^q$:

$$G_n(\widehat{\theta}_m^q) = \widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}_m^q; \widehat{w}_m^q, \widehat{m}_m) - Z_m^q(\widehat{\theta}_m^q) - \Delta_m^q(\widehat{\theta}_m^q).$$

The first nonnegative crossing condition controls the first term on the right, Condition A.2 makes the second term negligible and replaces G_n by the true target coverage gap, and successful candidate status controls the third term. The complete derivation is given in Subsection B.1.

A.1.2 Repeated propagation, vanishing tolerance, and union

The candidate condition above is downstream: it does not require either nuisance estimate to be globally accurate. The following rate specialization shows how the doubly robust product structure can generate a tolerance $\varepsilon_n \to 0$. Let

$$n = \min(n_{\mathrm{src}}^{\mathrm{cal}}, n_{\mathrm{tar}}^{\mathrm{cal}}), \qquad N = n_{\mathrm{src}}^{\mathrm{cal}} + n_{\mathrm{tar}}^{\mathrm{cal}},$$

and allow the number of regeneration runs $M = M_n$ to depend on the sample size.

Condition A.4. The oracle density ratio $w^\star$ is bounded by a fixed constant $W_0 < \infty$. Fix constants $C_w, C_m < \infty$ and exponents $\beta_w, \beta_m \ge 0$ satisfying $\beta_w + \beta_m > 0$. A finite candidate calibration threshold $\widehat{\theta}_m^q$ is *rate-controlled* when

$$\left[\mathbb{P}_{\mathrm{src},n}^1 \{\widehat{w}_m^q - w^\star\}^2\right]^{1/2} \le C_w n^{-\beta_w}, \tag{21}$$

$$\sup_{\theta\in\mathcal{S}_{m,n}} \left[\mathbb{P}_{\mathrm{src},n}^1 \{\widehat{m}_m(\theta, \cdot) - m^\star(\theta, \cdot)\}^2\right]^{1/2} \le C_m n^{-\beta_m}. \tag{22}$$

One exponent may be zero, so one nuisance may remain only bounded while the other converges.

For every regeneration run m and threshold θ, the transport discrepancy admits the exact product-plus-transport decomposition

$$\Delta_m^q(\theta) = \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q - w^\star}\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\} + \left[\mathbb{P}_{\mathrm{src},n}^{w^\star}\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\} - \mathbb{P}_{\mathrm{tar},n}\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\}\right]. \tag{23}$$

The first term is bounded by the product of the two empirical L_2 errors. The bracketed term is centered by the transport identity and is controlled uniformly over the candidate collection.

To make the resulting tolerance explicit, first apply Cauchy–Schwarz to the nuisance-product term in (23). For every rate-controlled candidate calibration threshold $\widehat{\theta}_m^q$,

$$\sup_{\theta\in\mathcal{S}_{m,n}}\left|\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q - w^\star}\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\}\right| \le \left[\mathbb{P}_{\mathrm{src},n}^{1}\{\widehat{w}_m^q - w^\star\}^2\right]^{1/2} \times \sup_{\theta\in\mathcal{S}_{m,n}}\left[\mathbb{P}_{\mathrm{src},n}^{1}\{\widehat{m}_m(\theta,\cdot) - m^\star(\theta,\cdot)\}^2\right]^{1/2}$$
$$\le C_w C_m n^{-(\beta_w+\beta_m)}. \tag{24}$$

For the remaining oracle-transport term, the transport identity

$$\mathbb{E}_{\mathrm{src}}[w^\star(X)h(X)] = \mathbb{E}_{\mathrm{tar}}[h(X)]$$

centers the source–target difference for $h(\cdot) = m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)$. Because both $m^\star(\theta,x)$ and $\widehat{m}_m(\theta,x)$ are bounded and nondecreasing in θ, uniform concentration for pointwise ordered classes, followed by a union bound over the M DRP moments, yields, for a universal constant $C<\infty$,

$$\max_{1\le m\le M}\sup_{\theta\in\mathbb{R}}\left|\mathbb{P}_{\mathrm{src},n}^{w^\star}\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\} - \mathbb{P}_{\mathrm{tar},n}\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\}\right|$$
$$\le C(W_0+1)\left[\sqrt{\frac{\log(32M/\delta)}{n_{\mathrm{src}}^{\mathrm{cal}}}} + \sqrt{\frac{\log(32M/\delta)}{n_{\mathrm{tar}}^{\mathrm{cal}}}}\right] \tag{25}$$

with probability at least $1-\delta/3$.

The preceding bound controls the candidate-specific transport discrepancy Δ_m^q and is therefore sufficient for the successful-candidate criterion. To obtain an explicit target-population coverage bound at the returned candidate calibration threshold, however, two additional empirical-process terms must also be controlled. Recall the exact decomposition

$$\widehat{\Psi}_{\mathrm{DRP}}(\theta;\widehat{w}_m^q,\widehat{m}_m) = Z_m^q(\theta) + G_n(\theta) + \Delta_m^q(\theta), \tag{26}$$

where

$$Z_m^q(\theta) = \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left[\mathbf{1}\{R(X,Y)\le\theta\} - m^\star(\theta,X)\right], \tag{27}$$
$$G_n(\theta) = \mathbb{P}_{\mathrm{tar},n}m^\star(\theta,\cdot) - (1-\alpha). \tag{28}$$

Here Z_m^q is the centered weighted source-score fluctuation, whereas G_n is the empirical target coverage gap, common to all candidates. At an exact finite first nonnegative crossing, $\widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}_m^q;\widehat{w}_m^q,\widehat{m}_m)\ge 0$, so

$$\mathbb{E}_{\mathrm{tar}}[m^\star(\widehat{\theta}_m^q,X)] - (1-\alpha) \ge -\left|\Delta_m^q(\widehat{\theta}_m^q)\right| - \left|Z_m^q(\widehat{\theta}_m^q)\right| - \left|G_n(\widehat{\theta}_m^q) - \left\{\mathbb{E}_{\mathrm{tar}}[m^\star(\widehat{\theta}_m^q,X)] - (1-\alpha)\right\}\right|. \tag{29}$$

Thus the oracle-transport bound above determines the explicit tolerance needed to certify $\sup_{\theta\in\mathcal{S}_{m,n}}|\Delta_m^q(\theta)|$, while Z_m^q and the difference between the empirical target coverage gap $G_n(\theta)$ and the true target coverage gap $\mathbb{E}_{\mathrm{tar}}[m^\star(\theta,X)] - (1-\alpha)$ enter only when that candidate-level control is translated into a finite-sample target-population coverage bound. We therefore control these two terms simultaneously over the candidate collection and absorb all three contributions into a common, slightly enlarged tolerance in the proposition below.

The same ordered-class argument controls the two empirical-process terms that enter when the DRP moment is applied to the returned candidate calibration threshold: with probability at least $1-\delta/3$,

$$\max_{1\le m\le M}\sup_{\theta\in\mathbb{R}}|Z_m^q(\theta)| \le CW\sqrt{\frac{\log(32M/\delta)}{n_{\mathrm{src}}^{\mathrm{cal}}}}, \tag{30}$$

and, with probability at least $1-\delta/3$,

$$\sup_{\theta\in\mathbb{R}}|G_n(\theta) - \{\mathbb{E}_{\mathrm{tar}}[m^\star(\theta,X)] - (1-\alpha)\}| \le C\sqrt{\frac{\log(32M/\delta)}{n_{\mathrm{tar}}^{\mathrm{cal}}}}. \tag{31}$$

Taking $\delta = N^{-2}$, where $N = n_{\mathrm{src}}^{\mathrm{cal}} + n_{\mathrm{tar}}^{\mathrm{cal}}$, and intersecting these events gives probability at least $1-N^{-2}$. Combining (24)– (31), and enlarging the universal constant C to absorb the separate

envelopes W, W_0, and 1, motivates the common tolerance

$$C_w C_m n^{-(\beta_w+\beta_m)} + C(W+W_0+1)\sqrt{\frac{\log(32MN^2)}{n_{\mathrm{src}}^{\mathrm{cal}}}} + C(W+W_0+1)\sqrt{\frac{\log(32MN^2)}{n_{\mathrm{tar}}^{\mathrm{cal}}}}.$$

The first term is the doubly robust nuisance-product contribution, whereas the two square-root terms collect the uniform source- and target-sample fluctuations needed both to control the candidate-specific transport discrepancy and to transfer the empirical first nonnegative crossing condition to target-population coverage.

Proposition A.2. *Suppose Assumption 2.1 and Conditions A.1 and A.4 hold. There is a universal constant $C < \infty$ such that, with probability at least $1 - N^{-2}$, simultaneously for every rate-controlled candidate calibration threshold among the M candidate calibration thresholds, the candidate calibration threshold is successful at tolerance*

$$\varepsilon_n = C_w C_m n^{-(\beta_w+\beta_m)} + C(W+W_0+1)\sqrt{\frac{\log(32MN^2)}{n_{\mathrm{src}}^{\mathrm{cal}}}} + C(W+W_0+1)\sqrt{\frac{\log(32MN^2)}{n_{\mathrm{tar}}^{\mathrm{cal}}}}. \tag{32}$$

Consequently,

$$\mathbb{E}_{\mathrm{tar}}\left[m^\star(\widehat{\theta}_m^q, X)\right] \geq 1 - \alpha - \varepsilon_n \tag{33}$$

on the same event. If a finite numerical fallback is used, its deterministic positive-part moment-deficit bound is added to (32). In particular, if $n_{\mathrm{src}}^{\mathrm{cal}}, n_{\mathrm{tar}}^{\mathrm{cal}} \to \infty$ and

$$\frac{\log(MN^2)}{n} \to 0, \tag{34}$$

then

$$\varepsilon_n \to 0.$$

The first term in ε_n is the doubly robust product of the density-ratio and conditional score-CDF errors. The remaining terms are uniform calibration-sample fluctuations over the M DRP moments. Thus increasing sample size shrinks the success tolerance, provided the number of regeneration runs grows slowly enough for (34) to hold.

Condition A.5. Conditional on the complete shared fitting and calibration data, the run-specific randomizations are independent across $m = 1, \ldots, M$. On an event whose probability tends to one, every run has conditional probability at least $p_{\varepsilon_n} > 0$ of producing a candidate calibration threshold that either equals $+\infty$ or is successful at tolerance ε_n. The quantity p_{ε_n} may depend on n and may tend to zero.

Let $S_m^q(\varepsilon_n)$ denote the successful-candidate event in run m. Here, shared data comprises all observed data and analysis inputs held fixed across regeneration runs, including the source calibration sample $\mathcal{D}_{\mathrm{src}}^{\mathrm{cal}}$, target calibration sample $\mathcal{D}_{\mathrm{tar}}^{\mathrm{cal}}$, fitting samples $\mathcal{D}_{\mathrm{src}}^{\mathrm{fit}}$ and $\mathcal{D}_{\mathrm{tar}}^{\mathrm{fit}}$, score R, ordered solver grids $\mathcal{S}_{m,n}$, and nominal coverage level $1-\alpha$; after conditioning on these quantities, the remaining randomness comes only from the run-specific subsampling, fold assignments, and nuisance regeneration. Conditional independence therefore gives

$$\mathbb{P}\left\{\bigcap_{m=1}^{M} S_m^q(\varepsilon_n)^c \,\middle|\, \text{shared data}\right\} \leq (1-p_{\varepsilon_n})^M \leq \exp\{-Mp_{\varepsilon_n}\}. \tag{35}$$

Hence propagation does not need to improve every nuisance pair $(\widehat{w}_m, \widehat{m}_m)$. It exponentially reduces the probability that all nuisance pairs $(\widehat{w}_m, \widehat{m}_m)$ miss the downstream success region.

For the nested candidate prediction sets $\mathcal{C}_{\widehat{\theta}_m^q}(x) = \{y : R(x,y) \leq \widehat{\theta}_m^q\}$, with $\mathcal{C}_{\widehat{\theta}_m^{\mathrm{tilt}}}(x) = \mathcal{C}_m(x)$,

$$\bigcup_{m=1}^{M} \mathcal{C}_{\widehat{\theta}_m^q}(x) = \left\{y : R(x,y) \leq \max_{1\leq m\leq M} \widehat{\theta}_m^q\right\},$$

$$\mathbb{E}_{\mathrm{tar}}\left[m^\star\left(\max_{1\leq m\leq M}\widehat{\theta}_m^q, X\right)\right] = \max_{1\leq m\leq M}\mathbb{E}_{\mathrm{tar}}[m^\star(\widehat{\theta}_m^q, X)]. \tag{36}$$

Therefore, once at least one successful candidate appears, taking the union, equivalently the maximum threshold, transfers that candidate's target coverage to the reported set without identifying which run succeeded.

Theorem A.1. *Suppose Assumption 2.1, Conditions A.1– A.2, and Condition A.5 hold. Then, for every fixed $\eta > 0$,*

$$\mathbb{P}\left\{\mathbb{E}_{\mathrm{tar}}\left[m^{\star}\left(\max_{1\le m\le M}\widehat{\theta}^{q}_{m}, X\right)\right] < 1-\alpha-\varepsilon_n-\eta\right\} \le (1-p_{\varepsilon_n})^M + o(1) \le \exp\{-Mp_{\varepsilon_n}\} + o(1). \tag{37}$$

The baseline-safe threshold

$$\widehat{\theta}^{q}_{\mathrm{safe}}(M) = \max\left\{\widehat{\theta}_{\mathrm{DRP}}, \max_{1\le m\le M}\widehat{\theta}^{q}_{m}\right\}$$

satisfies the same bound. The final DPC calibration threshold in (9) is $\widehat{\theta}_{\mathrm{DPC}} = \max_{1\le m\le M}\widehat{\theta}^{\mathrm{tilt}}_{m}$, without the additional baseline DRP threshold $\widehat{\theta}_{\mathrm{DRP}}$. For an independent new target observation $(X_j^{\mathrm{tar,new}}, Y_j^{\mathrm{tar,new}})$ indexed by any fixed $j \in \{1,\dots,n_{\mathrm{tar}}^{\mathrm{new}}\}$,

$$\mathbb{P}\left\{R(X_j^{\mathrm{tar,new}}, Y_j^{\mathrm{tar,new}}) \le \max_{1\le m\le M}\widehat{\theta}^{q}_{m}\right\} \ge 1-\alpha-\varepsilon_n-\eta-(1-p_{\varepsilon_n})^M - o(1). \tag{38}$$

Corollary A.1. *Under the assumptions of Proposition A.2, suppose that, conditional on the complete shared fitting and calibration data, the run randomizations are independent and every run has conditional probability at least p_{ε_n} of returning $+\infty$ or producing a rate-controlled candidate. Then*

$$\mathbb{P}\left\{\mathbb{E}_{\mathrm{tar}}\left[m^{\star}\left(\max_{1\le m\le M}\widehat{\theta}^{q}_{m}, X\right)\right] < 1-\alpha-\varepsilon_n\right\} \le (1-p_{\varepsilon_n})^M + N^{-2} + o(1), \tag{39}$$

and, for any fixed $j \in \{1,\dots,n_{\mathrm{tar}}^{\mathrm{new}}\}$,

$$\mathbb{P}\left\{R(X_j^{\mathrm{tar,new}}, Y_j^{\mathrm{tar,new}}) \le \max_{1\le m\le M}\widehat{\theta}^{q}_{m}\right\} \ge 1-\alpha-\varepsilon_n-(1-p_{\varepsilon_n})^M - N^{-2} - o(1). \tag{40}$$

If (34) holds and

$$Mp_{\varepsilon_n} \to \infty, \tag{41}$$

then $\varepsilon_n \to 0$, $(1-p_{\varepsilon_n})^M \le e^{-Mp_{\varepsilon_n}} \to 0$, and the reported propagated set has target coverage at least $1-\alpha-o(1)$.

For a desired all-run failure probability $\delta \in (0,1)$, it is sufficient to take

$$M \ge \frac{\log(1/\delta)}{-\log(1-p_{\varepsilon_n})}.$$

This expression makes the benefit of additional propagation explicit: holding the per-run success probability fixed, the probability of missing every successful candidate decreases geometrically, and at most exponentially, in the number of regeneration runs. The complete candidatewise, concentration, amplification, and coverage-transfer proofs for the union appear in Subsections B.1– B.3.

A.2 Tilting Step Efficiency: Faster Candidate Tolerances and Effective Propagation

The propagation results above apply to either $q = \mathrm{prop}$ or $q = \mathrm{tilt}$, but do not yet explain why the second implementation can be statistically more efficient. We now compare the two candidates within the same regeneration run. A paired regeneration run m uses the same regenerated conditional score-CDF estimate $\widehat{m}_m$, source and target calibration covariates, ordered solver grid $\mathcal{S}_{m,n}$, and first nonnegative crossing rule. The propagation-only candidate calibration threshold $\widehat{\theta}^{\mathrm{prop}}_m$ uses the nuisance pair $(\widehat{w}_m, \widehat{m}_m)$. The candidate calibration threshold $\widehat{\theta}^{\mathrm{tilt}}_m$ uses the tilted nuisance pair $(\widehat{w}^{\mathrm{tilt}}_m, \widehat{m}_m)$. Consequently, the comparison is entirely through the candidate-specific term $\Delta^q_m(\theta)$ in (13); the DRP moment, nominal coverage level $1-\alpha$, and union rule are unchanged.

Recall that the score-CDF balance features $g_m(x)$ and the non-intercept component $\widetilde{g}_m(x)$ are

$$g_m(x) = \{1, \widehat{m}_m(\tau_{m1}, x), \dots, \widehat{m}_m(\tau_{mK_m}, x)\}^{\mathsf{T}}, \qquad \widetilde{g}_m(x) = \{\widehat{m}_m(\tau_{m1}, x), \dots, \widehat{m}_m(\tau_{mK_m}, x)\}^{\mathsf{T}}. \tag{42}$$

Here K_m is the number of non-intercept score-CDF balance features used in run m. Accordingly, write the fitted multiplier for the tilting step and its non-intercept component as

$$\widehat{\lambda}_m = \left(\widehat{\lambda}_{0m}, \widehat{\gamma}_m^{\mathsf{T}}\right)^{\mathsf{T}}, \qquad \widehat{\gamma}_m = \left(\widehat{\lambda}_{m1}, \dots, \widehat{\lambda}_{mK_m}\right)^{\mathsf{T}}. \tag{43}$$

Here $\widehat{\lambda}_{0m}$ is the intercept component, while $\widehat{\gamma}_m$ contains the multipliers associated with the non-intercept score-CDF balance features $\widetilde{g}_m$.

Recall the empirical-average operators $\mathbb{P}^{w}_{\mathrm{src},n}$ and $\mathbb{P}_{\mathrm{tar},n}$ defined at the beginning of Appendix A. The former is a possibly unnormalized weighted source calibration average, whereas the latter is the

ordinary target calibration average. These operators should be distinguished from the hatted symbols $\widehat{\mathbb{P}}_m^{\mathrm{prop}}$, $\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$, and $\widehat{\mathbb{P}}_{m,u}$ introduced below, which denote normalized run-specific empirical probability laws and therefore have total mass one.

Using this notation, the imbalance in the score-CDF balance features before and after the tilting step can be written as

$$\mathcal{B}_{m,n} := \left\| \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} g_m - \mathbb{P}_{\mathrm{tar},n} g_m \right\|_\infty, \tag{44}$$

$$\mathcal{I}_{m,n} := \left\| \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} g_m - \mathbb{P}_{\mathrm{tar},n} g_m \right\|_\infty. \tag{45}$$

Here $\mathcal{B}_{m,n}$ measures the feature imbalance remaining after propagation alone, while $\mathcal{I}_{m,n}$ measures the imbalance after the tilting step. By the first-order balance equation (7), an exactly solved tilting step satisfies

$$\mathcal{I}_{m,n} = 0. \tag{46}$$

A.2.1 The tilting step is a local information projection

To make explicit which empirical distributions are compared, consider the probability laws induced by the importance weights $\widehat{w}_m(X_i^{\mathrm{src,cal}})$ and the tilted weights $\widehat{w}_m^{\mathrm{tilt}}(X_i^{\mathrm{src,cal}})$ on the source calibration covariates. Since the importance weights $\widehat{w}_m(X_i^{\mathrm{src,cal}})$ need not have unit empirical mass, define the normalized propagation-only weighted empirical distribution by assigning the ith source calibration covariate the probability mass

$$\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\} := \frac{\widehat{w}_m(X_i^{\mathrm{src,cal}})}{\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} \widehat{w}_m(X_\ell^{\mathrm{src,cal}})}, \qquad i = 1, \dots, n_{\mathrm{src}}^{\mathrm{cal}}. \tag{47}$$

Thus $\widehat{\mathbb{P}}_m^{\mathrm{prop}}$ is the probability distribution obtained by normalizing the importance weights $\widehat{w}_m(X_i^{\mathrm{src,cal}})$ over the source calibration sample.

For the tilted weights $\widehat{w}_m^{\mathrm{tilt}}$, exact balance of the intercept coordinate in g_m gives

$$\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} 1 = 1,$$

or equivalently

$$\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} \widehat{w}_m^{\mathrm{tilt}}(X_\ell^{\mathrm{src,cal}}) = n_{\mathrm{src}}^{\mathrm{cal}}.$$

Hence the tilted weighted empirical distribution assigns probability mass

$$\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\} := \frac{\widehat{w}_m^{\mathrm{tilt}}(X_i^{\mathrm{src,cal}})}{\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} \widehat{w}_m^{\mathrm{tilt}}(X_\ell^{\mathrm{src,cal}})} = \frac{\widehat{w}_m^{\mathrm{tilt}}(X_i^{\mathrm{src,cal}})}{n_{\mathrm{src}}^{\mathrm{cal}}}, \qquad i = 1, \dots, n_{\mathrm{src}}^{\mathrm{cal}}. \tag{48}$$

Therefore, $\widehat{\mathbb{P}}_m^{\mathrm{prop}}$ and $\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$ are two probability distributions supported on the same source calibration sample: the former is induced by the importance weights $\widehat{w}_m(X_i^{\mathrm{src,cal}})$, whereas the latter is induced by the tilted weights $\widehat{w}_m^{\mathrm{tilt}}(X_i^{\mathrm{src,cal}})$.

Equivalently, for any real-valued function h on the source calibration covariates,

$$\widehat{\mathbb{P}}_m^{\mathrm{prop}} h = \frac{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} h}{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} 1}, \qquad \widehat{\mathbb{P}}_m^{\mathrm{tilt}} h = \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} h.$$

Having placed the two weight vectors on the same probability scale, measure the perturbation induced by the tilting step through

$$D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}} \,\middle\|\, \widehat{\mathbb{P}}_m^{\mathrm{prop}} \right) := \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} \widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\} \log \frac{\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\}}{\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}}. \tag{49}$$

This divergence is zero when the tilting step leaves the normalized propagation distribution unchanged and increases as the tilted probability masses move away from their propagation-only values.

To identify the admissible balance corrections, let P denote any probability distribution supported on $\{X_1^{\mathrm{src,cal}}, \dots, X_{n_{\mathrm{src}}^{\mathrm{cal}}}^{\mathrm{src,cal}}\}$. It achieves the same empirical score-CDF balance targeted by the tilting

step when

$$\sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} P\{X_i^{\mathrm{src,cal}}\} g_m(X_i^{\mathrm{src,cal}}) = \mathbb{P}_{\mathrm{tar},n} g_m. \tag{50}$$

Because the intercept is the first coordinate of g_m, the first coordinate of (50) records that the probability masses sum to one, while the remaining features match the means of the score-CDF balance features g_m. By (46) and (48), $\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$ satisfies this constraint.

Combining (6), (47), and (48) gives

$$\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\} = \frac{\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\} \exp\left\{\widehat{\gamma}_m^{\mathsf{T}} \widetilde{g}_m(X_i^{\mathrm{src,cal}})\right\}}{\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} \widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_\ell^{\mathrm{src,cal}}\} \exp\left\{\widehat{\gamma}_m^{\mathsf{T}} \widetilde{g}_m(X_\ell^{\mathrm{src,cal}})\right\}}, \qquad i = 1, \ldots, n_{\mathrm{src}}^{\mathrm{cal}}. \tag{51}$$

The intercept multiplier $\widehat{\lambda}_{0m}$ supplies the common normalizing constant, while $\widehat{\gamma}_m$ changes the relative probability masses along the non-intercept score-CDF balance features $\widetilde{g}_m$.

Proposition A.3. *Consider a run m for which the exact optimizer for the tilting step exists and satisfies the balance equation (7). Then $\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$ is the unique solution, on the support of $\widehat{\mathbb{P}}_m^{\mathrm{prop}}$, to*

$$\widehat{\mathbb{P}}_m^{\mathrm{tilt}} = \arg\min_P \left\{ D_{\mathrm{KL}}\left(P \,\middle\|\, \widehat{\mathbb{P}}_m^{\mathrm{prop}}\right) : \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} P\{X_i^{\mathrm{src,cal}}\} g_m(X_i^{\mathrm{src,cal}}) = \mathbb{P}_{\mathrm{tar},n} g_m \right\}, \tag{52}$$

where P ranges over probability distributions supported on the source calibration covariates and satisfying $P\{X_i^{\mathrm{src,cal}}\} = 0$ whenever $\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\} = 0$.

The proof is given in Subsection C.1 of Section C. Proposition A.3 is a minimality statement: the exact tilting step selects the balanced distribution with the smallest KL displacement, but the proposition alone does not show that this minimum is small.

To quantify the size of the minimum displacement, trace the exponential weighting in the tilting step continuously from the propagation-only distribution to the exactly balanced distribution. For $u \in [0, 1]$, define

$$\widehat{\mathbb{P}}_{m,u}\{X_i^{\mathrm{src,cal}}\} := \frac{\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\} \exp\left\{u\widehat{\gamma}_m^{\mathsf{T}} \widetilde{g}_m(X_i^{\mathrm{src,cal}})\right\}}{\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} \widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_\ell^{\mathrm{src,cal}}\} \exp\left\{u\widehat{\gamma}_m^{\mathsf{T}} \widetilde{g}_m(X_\ell^{\mathrm{src,cal}})\right\}}, \qquad i = 1, \ldots, n_{\mathrm{src}}^{\mathrm{cal}}. \tag{53}$$

By construction,

$$\widehat{\mathbb{P}}_{m,0} = \widehat{\mathbb{P}}_m^{\mathrm{prop}}, \qquad \widehat{\mathbb{P}}_{m,1} = \widehat{\mathbb{P}}_m^{\mathrm{tilt}}.$$

The scalar u indexes only this analytical path and is not an additional tuning parameter.

Define the realized curvature along this path by

$$\mu_{m,n} := \lambda_{\min}\left[\int_0^1 \mathrm{Cov}_{\widehat{\mathbb{P}}_{m,u}}\{\widetilde{g}_m(X)\}\, du\right]. \tag{54}$$

If some features of $\widetilde{g}_m$ are exactly linearly redundant on the support of $\widehat{\mathbb{P}}_m^{\mathrm{prop}}$, they may be removed before forming (54); this does not change the balance constraint or the resulting tilted distribution. The quantity $\mu_{m,n}$ is a realized property of the score-CDF balance features g_m and empirical support, not an algorithmic tuning parameter.

Theorem A.2. *Suppose Condition A.1 holds, $\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} 1 > 0$, and the exact optimizer for the tilting step exists. If $\mu_{m,n} > 0$, then*

$$\|\widehat{\gamma}_m\|_2 \le \frac{2\sqrt{K_m}}{\mu_{m,n} \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} 1} \mathcal{B}_{m,n}, \tag{55}$$

and

$$D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}} \,\middle\|\, \widehat{\mathbb{P}}_m^{\mathrm{prop}}\right) \le \frac{4K_m}{\mu_{m,n}\left(\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} 1\right)^2} \mathcal{B}_{m,n}^2. \tag{56}$$

Consequently, along any sequence of runs for which the preliminary weighted mass stays bounded away from zero and

$$\frac{K_m \mathcal{B}_{m,n}^2}{\mu_{m,n}} \xrightarrow{\mathbb{P}} 0, \tag{57}$$

the perturbation from the tilting step satisfies

$$D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}} \,\middle\|\, \widehat{\mathbb{P}}_m^{\mathrm{prop}}\right) = o_{\mathbb{P}}(1).$$

The path identity, normalized-imbalance bound, and proof of Theorem A.2 are given in Subsection C.2. The theorem shows that small preliminary imbalance leads to a small KL correction provided the retained score-CDF balance features g_m do not degenerate too quickly.

When $\mathcal{B}_{m,n} = 0$, the normalized propagation-only distribution is already feasible for exact balance, so $\widehat{\mathbb{P}}_m^{\mathrm{tilt}} = \widehat{\mathbb{P}}_m^{\mathrm{prop}}$. When $\mu_{m,n}$ is very small, the bound correctly allows a larger perturbation. If exact balance is infeasible, the information-projection and local-KL conclusions do not apply to that run, although the propagation-only candidate calibration threshold $\widehat{\theta}_m^{\mathrm{prop}}$ remains valid under the preceding propagation theory.

A.2.2 The tilting step contracts the candidate tolerance

The preceding subsubsection shows that exact score-CDF balance can be obtained through a controlled perturbation of the propagation-only empirical distribution. We now return to the successful-candidate criterion. Under the exact first nonnegative crossing used by the method, a finite candidate calibration threshold $\widehat{\theta}_m^q$ is successful at tolerance ε_n whenever

$$\sup_{\theta \in \mathcal{S}_{m,n}} |\Delta_m^q(\theta)| \le \varepsilon_n, \qquad q \in \{\mathrm{prop}, \mathrm{tilt}\}. \tag{58}$$

Using the empirical-average notation defined at the beginning of the appendix, $\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}$ is the weighted source calibration average and $\mathbb{P}_{\mathrm{tar},n}$ is the ordinary target calibration average. With this shorthand,

$$\Delta_m^q(\theta) = \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\} - \mathbb{P}_{\mathrm{tar},n}\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\}. \tag{59}$$

The paired regeneration runs therefore transport the same conditional score-CDF estimation error $e_m(\theta, x)$ and differ only through the importance weights $\widehat{w}_m(X_i^{\mathrm{src,cal}})$ and the tilted weights $\widehat{w}_m^{\mathrm{tilt}}(X_i^{\mathrm{src,cal}})$.

The relevant balance relations are

$$\left\|\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} g_m - \mathbb{P}_{\mathrm{tar},n} g_m\right\|_\infty = \mathcal{B}_{m,n}, \qquad \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} g_m - \mathbb{P}_{\mathrm{tar},n} g_m = 0. \tag{60}$$

For each $\theta \in \mathcal{S}_{m,n}$, choose a coefficient vector $a_m(\theta)$ and define

$$r_m(\theta, x) := e_m(\theta, x) - a_m(\theta)^{\mathsf{T}} g_m(x). \tag{61}$$

Equivalently,

$$e_m(\theta, x) = a_m(\theta)^{\mathsf{T}} g_m(x) + r_m(\theta, x).$$

This decomposition does not require the conditional score-CDF estimation error $e_m(\theta, x)$ to lie exactly in the span of the score-CDF balance features g_m; r_m records the unrepresented part.

Substituting (61) into (59) initially gives the same represented-plus-remainder decomposition for both implementations. The difference appears after applying the balance relations in (60). For the tilted weights $\widehat{w}_m^{\mathrm{tilt}}$, exact balance makes the represented component identically zero. We then add and subtract the propagation-only source average of the remainder only to separate the remainder transport already present before the tilting step from the additional change induced by the tilting step:

$$\Delta_m(\theta) = a_m(\theta)^{\mathsf{T}} \underbrace{\left\{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} g_m - \mathbb{P}_{\mathrm{tar},n} g_m\right\}}_{\ell_\infty\text{-norm equals } \mathcal{B}_{m,n}} + \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} r_m(\theta,\cdot) - \mathbb{P}_{\mathrm{tar},n} r_m(\theta,\cdot),$$

$$\begin{aligned}\Delta_m^{\mathrm{tilt}}(\theta) &= \underbrace{a_m(\theta)^{\mathsf{T}} \left\{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} g_m - \mathbb{P}_{\mathrm{tar},n} g_m\right\}}_{=0 \text{ by exact balance}} + \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} r_m(\theta,\cdot) - \mathbb{P}_{\mathrm{tar},n} r_m(\theta,\cdot) \\ &= \underbrace{\left\{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} r_m(\theta,\cdot) - \mathbb{P}_{\mathrm{tar},n} r_m(\theta,\cdot)\right\}}_{\text{propagation-only remainder transport}} + \underbrace{\left\{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} r_m(\theta,\cdot) - \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} r_m(\theta,\cdot)\right\}}_{\text{change in the remainder average induced by the tilting step}}.\end{aligned} \tag{62}$$

where exact balance removes the represented term from the transport discrepancy $\Delta_m^{\mathrm{tilt}}(\theta)$. The verification of (62) and the control of the remaining source-average perturbation are given in Subsection C.3.

To state the uniform comparison, define

$$A_{m,n} := \sup_{\theta \in \mathcal{S}_{m,n}} \|a_m(\theta)\|_1 , \tag{63}$$

$$\rho_{m,n} := \sup_{\theta \in \mathcal{S}_{m,n}} \max_{1 \le i \le n_{\mathrm{src}}^{\mathrm{cal}}} \left| r_m(\theta, X_i^{\mathrm{src,cal}}) \right| , \tag{64}$$

$$\mathcal{V}_{m,n} := \sup_{\theta \in \mathcal{S}_{m,n}} \left| \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} r_m(\theta, \cdot) - \mathbb{P}_{\mathrm{tar},n} r_m(\theta, \cdot) \right| . \tag{65}$$

Here $A_{m,n}$ measures the largest represented coefficient size, $\rho_{m,n}$ measures the largest source-support remainder, and $\mathcal{V}_{m,n}$ is the propagation-only transport of that remainder.

Theorem A.3. *Consider a paired run m satisfying the assumptions of Theorem A.2. Suppose that the tilting step is solved exactly and exact first nonnegative crossings are used, and that (61) holds for every $\theta \in \mathcal{S}_{m,n}$. Define*

$$\sup_{\theta \in \mathcal{S}_{m,n}} |\Delta_m(\theta)| \le \bar{\varepsilon}_{m,n}^{\mathrm{prop}} := A_{m,n}\mathcal{B}_{m,n} + \mathcal{V}_{m,n}, \tag{66}$$

$$\sup_{\theta \in \mathcal{S}_{m,n}} \left|\Delta_m^{\mathrm{tilt}}(\theta)\right| \le \bar{\varepsilon}_{m,n}^{\mathrm{tilt}} := \mathcal{V}_{m,n} + \rho_{m,n}\mathcal{B}_{m,n} \left[1 + \frac{2\sqrt{2K_m}}{\sqrt{\mu_{m,n}}\, \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} 1} \right] . \tag{67}$$

Then

$$\sup_{\theta \in \mathcal{S}_{m,n}} |\Delta_m(\theta)| \le \bar{\varepsilon}_{m,n}^{\mathrm{prop}}, \qquad \sup_{\theta \in \mathcal{S}_{m,n}} \left|\Delta_m^{\mathrm{tilt}}(\theta)\right| \le \bar{\varepsilon}_{m,n}^{\mathrm{tilt}}. \tag{68}$$

Consequently, for either $q \in \{\mathrm{prop}, \mathrm{tilt}\}$,

$$\bar{\varepsilon}_{m,n}^{q} \le \varepsilon_n \quad \rightarrow \quad \sup_{\theta \in \mathcal{S}_{m,n}} |\Delta_m^{q}(\theta)| \le \varepsilon_n, \tag{69}$$

so the candidate calibration threshold $\widehat{\theta}_m^q$ is successful at tolerance ε_n under Condition A.3.

The two certified tolerances satisfy

$$\rho_{m,n} \left[1 + \frac{2\sqrt{2K_m}}{\sqrt{\mu_{m,n}}\, \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} 1} \right] \le A_{m,n} \quad \rightarrow \quad \bar{\varepsilon}_{m,n}^{\mathrm{tilt}} \le \bar{\varepsilon}_{m,n}^{\mathrm{prop}}. \tag{70}$$

If the inequality on the left is strict and $\mathcal{B}_{m,n} > 0$, then the inequality between the certified tolerances is strict.

More strongly, consider a sequence of paired runs for which $A_{m,n}\mathcal{B}_{m,n} > 0$ with probability tending to one. If

$$\frac{\mathcal{V}_{m,n}}{A_{m,n}\mathcal{B}_{m,n}} \xrightarrow{\mathbb{P}} 0 \tag{71}$$

and

$$\frac{\rho_{m,n}}{A_{m,n}} \left[1 + \frac{2\sqrt{2K_m}}{\sqrt{\mu_{m,n}}\, \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} 1} \right] \xrightarrow{\mathbb{P}} 0, \tag{72}$$

then

$$\bar{\varepsilon}_{m,n}^{\mathrm{prop}} = A_{m,n}\mathcal{B}_{m,n}\{1 + o_{\mathbb{P}}(1)\}, \qquad \bar{\varepsilon}_{m,n}^{\mathrm{tilt}} = o_{\mathbb{P}}(A_{m,n}\mathcal{B}_{m,n}) , \tag{73}$$

and therefore

$$\frac{\bar{\varepsilon}_{m,n}^{\mathrm{tilt}}}{\bar{\varepsilon}_{m,n}^{\mathrm{prop}}} \xrightarrow{\mathbb{P}} 0. \tag{74}$$

The proof is given in Subsection C.3. It uses Theorem A.2 only to control the change in the unrepresented remainder; the successful-candidate implication itself is inherited from the propagation theory and is not reproved.

The finite-sample comparison separates the benefit of exact balance from the cost of changing the empirical source distribution. Propagation alone retains $A_{m,n}\mathcal{B}_{m,n}$, while the exact tilting step removes this represented term and introduces only a perturbation proportional to $\rho_{m,n}$. The ordering condition (70) is therefore a comparison of approximation quality, not an assumption of better coverage.

If $\mathcal{B}_{m,n} = 0$, the propagation-only distribution is already balanced. If $A_{m,n} = 0$, the score-CDF balance features g_m represent no part of the conditional score-CDF estimation error $e_m(\theta, x)$. If $r_m(\theta, x) = 0$ on both source and target calibration covariates, then $\rho_{m,n} = \mathcal{V}_{m,n} = 0$ and the exact tilting step gives $\Delta_m^{\mathrm{tilt}}(\theta) = 0$ throughout the ordered solver grid. Conversely, if $\mathcal{V}_{m,n}$ is the leading term, both methods share the same remainder-transport floor and the tilting step need not improve the rate. For an approximate numerical solution of the tilting step, the tilted certificate additionally contains $A_{m,n}\mathcal{I}_{m,n}$; any positive-part first nonnegative crossing deficit must likewise be added as in Condition A.3.

A.2.3 From candidate-tolerance contraction to effective propagation

The preceding theorem gives a candidatewise comparison within one regeneration run. Recall from the propagation analysis that a lower bound on the per-run success probability yields a geometric all-run failure bound, and the union transfers the coverage of any successful candidate to the reported set. It remains to compare how often the paired propagation-only and tilted certificates cross the same success threshold ε_n.

Define the actual success event

$$S_m^q(\varepsilon_n) := \left\{\widehat{\theta}_m^q = +\infty\right\} \cup \left\{\widehat{\theta}_m^q < +\infty,\ \sup_{\theta \in \mathcal{S}_{m,n}} |\Delta_m^q(\theta)| \le \varepsilon_n\right\}, \qquad q \in \{\mathrm{prop}, \mathrm{tilt}\}, \tag{75}$$

and, for paired regeneration runs with finite candidate calibration thresholds covered by Theorem A.3, define the certified success event

$$C_m^q(\varepsilon_n) := \left\{\bar{\varepsilon}_{m,n}^q \le \varepsilon_n\right\}. \tag{76}$$

Runs outside the scope of that theorem are counted as uncertified. By (69),

$$C_m^q(\varepsilon_n) \subseteq S_m^q(\varepsilon_n).$$

Define the conditional certified success mass by

$$\underline{p}_{\varepsilon_n}^q := \mathbb{P}\{C_m^q(\varepsilon_n) \,|\, \text{shared data}\}, \qquad q \in \{\mathrm{prop}, \mathrm{tilt}\}. \tag{77}$$

It is a lower bound on the corresponding actual per-run success probability.

Without assuming that the tilting step is superior, define

$$h_{\varepsilon_n}^+ := \mathbb{P}\left\{C_m^{\mathrm{tilt}}(\varepsilon_n) \setminus C_m^{\mathrm{prop}}(\varepsilon_n) \,\middle|\, \text{shared data}\right\}, \tag{78}$$

$$h_{\varepsilon_n}^- := \mathbb{P}\left\{C_m^{\mathrm{prop}}(\varepsilon_n) \setminus C_m^{\mathrm{tilt}}(\varepsilon_n) \,\middle|\, \text{shared data}\right\}. \tag{79}$$

Then

$$\underline{p}_{\varepsilon_n}^{\mathrm{tilt}} - \underline{p}_{\varepsilon_n}^{\mathrm{prop}} = h_{\varepsilon_n}^+ - h_{\varepsilon_n}^-. \tag{80}$$

This identity imposes no ordering: $h_{\varepsilon_n}^+$ records runs newly certified by the tilting step, while $h_{\varepsilon_n}^-$ records certified losses.

Suppose that, conditional on the shared data, every paired regeneration run with finite candidate calibration thresholds entering the comparison satisfies

$$\rho_{m,n}\left[1 + \frac{2\sqrt{2K_m}}{\sqrt{\mu_{m,n}}\,\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} 1}\right] \le A_{m,n}. \tag{81}$$

Then (70) gives $C_m^{\mathrm{prop}}(\varepsilon_n) \subseteq C_m^{\mathrm{tilt}}(\varepsilon_n)$, so

$$h_{\varepsilon_n}^- = 0. \tag{82}$$

Define the newly certified mass by

$$\begin{aligned}\pi_{\varepsilon_n} &:= \mathbb{P}\left\{C_m^{\mathrm{tilt}}(\varepsilon_n) \setminus C_m^{\mathrm{prop}}(\varepsilon_n) \,\middle|\, \text{shared data}\right\} \\ &= \mathbb{P}\left\{\bar{\varepsilon}_{m,n}^{\mathrm{tilt}} \le \varepsilon_n < \bar{\varepsilon}_{m,n}^{\mathrm{prop}} \,\middle|\, \text{shared data}\right\}.\end{aligned} \tag{83}$$

Consequently,

$$\underline{p}_{\varepsilon_n}^{\mathrm{tilt}} = \underline{p}_{\varepsilon_n}^{\mathrm{prop}} + \pi_{\varepsilon_n} \ge \underline{p}_{\varepsilon_n}^{\mathrm{prop}}. \tag{84}$$

The inequality is strict exactly when $\pi_{\varepsilon_n} > 0$. These set identities and the inclusion argument are verified in Subsection C.4. A run counted by π_{ε_n} is certified to succeed after the tilting step, but its propagation-only certificate exceeding ε_n does not prove that the propagation-only candidate calibration threshold $\widehat{\theta}_m^{\mathrm{prop}}$ fails.

The propagation-robustness theorem already converts a conditional per-run success lower bound into coverage and run-count guarantees. Applying that result with the certified lower bounds in (77) gives the following comparison.

Theorem A.4. *Suppose Assumption 2.1 and Conditions A.1–A.2 hold. Suppose also that the paired regeneration runs with finite candidate calibration thresholds used to define $C_m^{\mathrm{prop}}(\varepsilon_n)$ and $C_m^{\mathrm{tilt}}(\varepsilon_n)$*

satisfy the assumptions of Theorem A.3. Conditional on the shared fitting and calibration data, assume that the run-specific randomizations are independent and that $\underline{p}^{q}_{\varepsilon_n}$ is a common lower bound on the certified success probability of each run for $q \in \{\mathrm{prop}, \mathrm{tilt}\}$.

Then, for every fixed $\eta > 0$ and $q \in \{\mathrm{prop}, \mathrm{tilt}\}$, with probability tending to one over the shared data,

$$\mathbb{P}\left\{ \mathbb{E}_{\mathrm{tar}}\left[m^{\star}\left(\max_{1 \le m \le M} \widehat{\theta}^{q}_{m}, X \right) \right] < 1 - \alpha - \varepsilon_n - \eta \,\middle|\, \textit{shared data} \right\} \le \left(1 - \underline{p}^{q}_{\varepsilon_n}\right)^{M} \le \exp\left\{ -M \underline{p}^{q}_{\varepsilon_n} \right\}. \tag{85}$$

The same bound holds after retaining the baseline DRP threshold $\widehat{\theta}_{\mathrm{DRP}}$.

If (81) also holds, then

$$\underline{p}^{\mathrm{tilt}}_{\varepsilon_n} = \underline{p}^{\mathrm{prop}}_{\varepsilon_n} + \pi_{\varepsilon_n} \ge \underline{p}^{\mathrm{prop}}_{\varepsilon_n}, \tag{86}$$

and hence, for every $M \ge 1$,

$$\left(1 - \underline{p}^{\mathrm{tilt}}_{\varepsilon_n}\right)^{M} \le \left(1 - \underline{p}^{\mathrm{prop}}_{\varepsilon_n}\right)^{M}. \tag{87}$$

The inequality is strict whenever $\pi_{\varepsilon_n} > 0$.

For a desired certified all-run failure probability $\delta \in (0,1)$ and $0 < \underline{p}^{q}_{\varepsilon_n} < 1$, the smallest integer number of runs making the geometric failure bound no larger than δ is

$$\left\lceil \frac{\log(1/\delta)}{-\log\left(1 - \underline{p}^{q}_{\varepsilon_n}\right)} \right\rceil. \tag{88}$$

Consequently,

$$\left\lceil \frac{\log(1/\delta)}{-\log\left(1 - \underline{p}^{\mathrm{tilt}}_{\varepsilon_n}\right)} \right\rceil \le \left\lceil \frac{\log(1/\delta)}{-\log\left(1 - \underline{p}^{\mathrm{prop}}_{\varepsilon_n}\right)} \right\rceil. \tag{89}$$

If $\pi_{\varepsilon_n} > 0$, the corresponding unrounded requirement is strictly smaller for the tilting step, although integer rounding may make the displayed counts equal.

Finally, if

$$0 < \underline{p}^{\mathrm{prop}}_{\varepsilon_n} \le \underline{p}^{\mathrm{tilt}}_{\varepsilon_n} < 1,$$

then M tilted runs have a certified all-run failure bound no larger than that of

$$\left\lfloor M \frac{-\log\left(1 - \underline{p}^{\mathrm{tilt}}_{\varepsilon_n}\right)}{-\log\left(1 - \underline{p}^{\mathrm{prop}}_{\varepsilon_n}\right)} \right\rfloor \tag{90}$$

propagation-only runs. The multiplier is at least one and is strictly greater than one whenever $\pi_{\varepsilon_n} > 0$.

The proof of the certified-mass ordering and effective-run formulas is given in Subsection C.4. The coverage-transfer step follows directly from Proposition A.1, Theorem A.1, and the union argument already proved for propagation robustness; those arguments are not repeated.

The theorem concerns the reliability of obtaining at least one successful candidate and the corresponding number of regeneration runs.

B Technical Derivations for Propagation Robustness

This subsection contains the complete algebraic and probability derivations supporting the statements above. The first chain proves the exact DRP decomposition and the one-successful-candidate result. The second derives the sample-size tolerance from the doubly robust product structure. The third proves geometric amplification and coverage transfer through union. The remaining chains establish the growing-shift and width results stated in the preceding subsections.

B.1 Candidatewise decomposition and proof

Fix $q \in \{\mathrm{prop}, \mathrm{tilt}\}$. A regeneration run changes only the nuisance-fitting randomization within the fixed candidate construction. The score R, calibration samples $\mathcal{D}^{\mathrm{cal}}_{\mathrm{src}}$ and $\mathcal{D}^{\mathrm{cal}}_{\mathrm{tar}}$, nominal coverage level

$1-\alpha$, and first nonnegative crossing rule remain fixed across runs. We now expand the DRP moment for regeneration run m and prove Proposition A.1.

Fix regeneration run m and a threshold $\theta \in \mathcal{S}_{m,n}$ on the ordered solver grid. In the weighted source part of the DRP moment, add and subtract the oracle conditional score CDF $m^\star(\theta, X)$:

$$\begin{aligned}&\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left\{\mathbf{1}\{R(X,Y)\le\theta\}-\widehat{m}_m(\theta,X)\right\}\\&\quad=\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left[\mathbf{1}\{R(X,Y)\le\theta\}-m^\star(\theta,X)\right]+\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left[m^\star(\theta,X)-\widehat{m}_m(\theta,X)\right].\end{aligned}$$

In the target augmentation part, add and subtract the oracle conditional score CDF $m^\star(\theta, X)$:

$$\begin{aligned}&\mathbb{P}_{\mathrm{tar},n}\left\{\widehat{m}_m(\theta,\cdot)-(1-\alpha)\right\}\\&\quad=\mathbb{P}_{\mathrm{tar},n}m^\star(\theta,\cdot)-(1-\alpha)-\mathbb{P}_{\mathrm{tar},n}\left[m^\star(\theta,\cdot)-\widehat{m}_m(\theta,\cdot)\right].\end{aligned}$$

Adding $\mathbb{P}_{\mathrm{tar},n}\left\{\widehat{m}_m(\theta,\cdot)-(1-\alpha)\right\}$ and $\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left\{\mathbf{1}\{R(X,Y)\le\theta\}-\widehat{m}_m(\theta,X)\right\}$ gives the exact decomposition

$$\begin{aligned}&\underbrace{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left\{\mathbf{1}\{R(X,Y)\le\theta\}-\widehat{m}_m(\theta,X)\right\}+\mathbb{P}_{\mathrm{tar},n}\left\{\widehat{m}_m(\theta,\cdot)-(1-\alpha)\right\}}_{\widehat{\Psi}_{\mathrm{DRP}}(\theta;\widehat{w}_m^q,\widehat{m}_m)\text{: DRP moment}}\\&=\underbrace{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left[\mathbf{1}\{R(X,Y)\le\theta\}-m^\star(\theta,X)\right]}_{Z_m^q(\theta)\text{: weighted source-score fluctuation}}+\underbrace{\left[\mathbb{P}_{\mathrm{tar},n}m^\star(\theta,\cdot)-(1-\alpha)\right]}_{G_n(\theta)\text{: empirical target coverage gap}}\\&\quad+\underbrace{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left[m^\star(\theta,\cdot)-\widehat{m}_m(\theta,\cdot)\right]-\mathbb{P}_{\mathrm{tar},n}\left[m^\star(\theta,\cdot)-\widehat{m}_m(\theta,\cdot)\right]}_{\Delta_m^q(\theta)}.\end{aligned}\tag{91}$$

Combining these terms gives exactly the compact identity (13) stated in the main text.

For the source-score fluctuation, conditional on the source covariates and the fitting data,

$$\mathbb{E}\left[\mathbf{1}\{R(X_i^{\mathrm{src,cal}},Y_i^{\mathrm{src,cal}})\le\theta\}-m^\star(\theta,X_i^{\mathrm{src,cal}})\;\middle|\;X_i^{\mathrm{src,cal}}\right]=0.\tag{92}$$

At a fixed threshold, $G_n(\theta)$ contains no candidate-specific nuisance estimate and is therefore common to all regeneration runs. The second part of Condition A.2 controls $G_n(\theta)$ uniformly by the true target coverage gap $\mathbb{E}_{\mathrm{tar}}[m^\star(\theta, X)] - (1-\alpha)$.

For completeness, the compact candidate-success requirement in (19) expands to

$$\begin{aligned}&\sup_{\theta\in\mathcal{S}_{m,n}}\left|\underbrace{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left[m^\star(\theta,\cdot)-\widehat{m}_m(\theta,\cdot)\right]-\mathbb{P}_{\mathrm{tar},n}\left[m^\star(\theta,\cdot)-\widehat{m}_m(\theta,\cdot)\right]}_{\Delta_m^q(\theta)}\right|\\&+\underbrace{\left[-\underbrace{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q}\left\{\mathbf{1}\left[R(X,Y)\le\widehat{\theta}_m^q\right]-\widehat{m}_m(\widehat{\theta}_m^q,X)\right\}+\mathbb{P}_{\mathrm{tar},n}\left\{\widehat{m}_m(\widehat{\theta}_m^q,\cdot)-(1-\alpha)\right\}}_{\widehat{\Psi}_{\mathrm{DRP}}(\theta;\widehat{w}_m^q,\widehat{m}_m)\text{: DRP moment}}\right]_+}_{\text{positive-part numerical moment deficit}}\le\varepsilon_n.\end{aligned}\tag{93}$$

The preceding facts give the componentwise summary

$$\begin{cases}\widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}_m^q;\widehat{w}_m^q,\widehat{m}_m)\ge-\left[-\widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}_m^q;\widehat{w}_m^q,\widehat{m}_m)\right]_+,\\[2ex]\max\limits_{1\le m\le M}\sup\limits_{\theta\in\mathcal{S}_{m,n}}|Z_m^q(\theta)|=o_{\mathbb{P}}(1),\\[2ex]\sup\limits_{\theta\in\mathbb{R}}|G_n(\theta)-[\mathbb{E}_{\mathrm{tar}}[m^\star(\theta,X)]-(1-\alpha)]|=o_{\mathbb{P}}(1),\\[2ex]\sup\limits_{\theta\in\mathcal{S}_{m,n}}|\Delta_m^q(\theta)|+\left[-\widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}_m^q;\widehat{w}_m^q,\widehat{m}_m)\right]_+\le\varepsilon_n,\\[2ex]\quad\text{for a successful finite candidate calibration threshold.}\end{cases}\tag{94}$$

The DRP moment and its numerical deficit determine how the three components are converted into the selected candidate calibration threshold; Z_m^q is candidate-dependent but centered sampling noise; $G_n(\theta)$ is the empirical target coverage gap; and Δ_m^q is the transport discrepancy.

We now carry these bounds through the returned candidate calibration threshold and prove Proposition A.1.

Proof of Proposition A.1. Fix a candidate calibration threshold $\widehat{\theta}^q_m$ satisfying (19). Suppose first that $\widehat{\theta}^q_m$ is finite. Evaluate (13) at $\theta = \widehat{\theta}^q_m$ and solve for the empirical target coverage gap. This gives the exact identity at the candidate calibration threshold

$$G_n(\widehat{\theta}^q_m) = \widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}^q_m; \widehat{w}^q_m, \widehat{m}_m) - Z^q_m(\widehat{\theta}^q_m) - \Delta^q_m(\widehat{\theta}^q_m). \tag{95}$$

Apply the first inequality to $\widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}^q_m; \widehat{w}^q_m, \widehat{m}_m)$ and the second separately to $Z^q_m(\widehat{\theta}^q_m)$ and $\Delta^q_m(\widehat{\theta}^q_m)$. Expanding all three quantities yields

$$\underbrace{\left[\mathbb{P}_{\mathrm{tar},n} m^\star(\widehat{\theta}^q_m, \cdot) - (1-\alpha)\right]}_{G_n(\theta)\text{: empirical target coverage gap}}$$

$$\geq -\underbrace{\left|\mathbb{P}^{\widehat{w}^q_m}_{\mathrm{src},n}\left[\mathbf{1}\{R(X,Y) \leq \widehat{\theta}^q_m\} - m^\star(\widehat{\theta}^q_m, X)\right]\right|}_{Z^q_m(\theta)\text{: weighted source-score fluctuation}}$$

$$- \underbrace{\left|\mathbb{P}^{\widehat{w}^q_m}_{\mathrm{src},n}\left[m^\star(\widehat{\theta}^q_m, \cdot) - \widehat{m}_m(\widehat{\theta}^q_m, \cdot)\right] - \mathbb{P}_{\mathrm{tar},n}\left[m^\star(\widehat{\theta}^q_m, \cdot) - \widehat{m}_m(\widehat{\theta}^q_m, \cdot)\right]\right|}_{\Delta^q_m(\theta)}$$

$$- \underbrace{\left[-\mathbb{P}^{\widehat{w}^q_m}_{\mathrm{src},n}\left\{\mathbf{1}\{R(X,Y) \leq \widehat{\theta}^q_m\} - \widehat{m}_m(\widehat{\theta}^q_m, X)\right\} - \mathbb{P}_{\mathrm{tar},n}\left\{\widehat{m}_m(\widehat{\theta}^q_m, \cdot) - (1-\alpha)\right\}\right]_+}_{\text{positive-part numerical moment deficit}}.$$

Because $\widehat{\theta}^q_m \in \mathcal{S}_{m,n}$, the absolute transport discrepancy at the returned threshold is no larger than its supremum over the ordered solver grid. Condition A.3 therefore bounds the sum of the second and third negative terms by ε_n. Condition A.2 makes the absolute source-score fluctuation $o_{\mathbb{P}}(1)$ uniformly. Hence

$$\underbrace{\left[\mathbb{P}_{\mathrm{tar},n} m^\star(\widehat{\theta}^q_m, \cdot) - (1-\alpha)\right]}_{G_n(\widehat{\theta}^q_m)} \geq -\varepsilon_n - o_{\mathbb{P}}(1). \tag{96}$$

Adding $1-\alpha$ to both sides gives

$$\mathbb{P}_{\mathrm{tar},n} m^\star(\widehat{\theta}^q_m, \cdot) \geq 1 - \alpha - \varepsilon_n - o_{\mathbb{P}}(1). \tag{97}$$

It remains to replace the empirical target average by target-population coverage at the data-dependent candidate calibration threshold. Because (18) is uniform over all real thresholds, it remains valid at $\widehat{\theta}^q_m$. Explicitly,

$$\left|\mathbb{P}_{\mathrm{tar},n} m^\star(\widehat{\theta}^q_m, \cdot) - \mathbb{E}_{\mathrm{tar}}\left[m^\star(\widehat{\theta}^q_m, X)\right]\right| \leq \sup_{\theta \in \mathbb{R}} \left| \left[\mathbb{P}_{\mathrm{tar},n} m^\star(\theta, \cdot) - (1-\alpha)\right] - \left[\mathbb{E}_{\mathrm{tar}}[m^\star(\theta, X)] - (1-\alpha)\right] \right| = o_{\mathbb{P}}(1).$$

Combining this bound with (97) gives

$$\mathbb{E}_{\mathrm{tar}}\left[m^\star(\widehat{\theta}^q_m, X)\right] \geq 1 - \alpha - \varepsilon_n - o_{\mathbb{P}}(1). \tag{98}$$

This is (20). If $\widehat{\theta}^q_m = +\infty$, then $m^\star(+\infty, x) = 1$, and the target-population coverage is one, so the same conclusion is immediate. □

B.2 Derivation of the Sample-Size Candidate Tolerance

This subsection proves Proposition A.2.

Lemma B.1. *Let $U_1, \ldots, U_N$ be independent conditional on a sigma-field $\mathcal{H}$. For each i and $\theta \in \mathbb{R}$, let $f_{\theta,i}$ be $\mathcal{H}$-measurable as a function and suppose $0 \leq f_{\theta,i}(u) \leq B$ and $\theta \mapsto f_{\theta,i}(u)$ is nondecreasing for every u. Then, for all $\delta \in (0,1)$, conditional on $\mathcal{H}$, with probability at least $1-\delta$,*

$$\sup_{\theta \in \mathbb{R}} \left| \frac{1}{N} \sum_{i=1}^{N} [f_{\theta,i}(U_i) - \mathbb{E}\{f_{\theta,i}(U_i) \mid \mathcal{H}\}] \right| \leq \frac{4B}{\sqrt{N}} + B\sqrt{\frac{\log(1/\delta)}{2N}}. \tag{99}$$

Consequently, for any collection of $K \geq 1$ such ordered classes with a common envelope B, there is a universal constant C such that, conditional on $\mathcal{H}$, with probability at least $1-\delta$,

$$\max_{1 \leq k \leq K} \sup_{\theta \in \mathbb{R}} \left| \frac{1}{N} \sum_{i=1}^{N} [f_{k,\theta,i}(U_i) - \mathbb{E}\{f_{k,\theta,i}(U_i) \mid \mathcal{H}\}] \right| \leq CB\sqrt{\frac{\log(2K/\delta)}{N}}. \tag{100}$$

Proof. Conditional on $\mathcal{H}$, define the left side of (99) by A_N. Standard symmetrization for independent, not necessarily identically distributed, observations gives

$$\mathbb{E}(A_N \mid \mathcal{H}) \le 2\,\mathbb{E}\left[\sup_{\theta\in\mathbb{R}}\left|\frac{1}{N}\sum_{i=1}^{N}\xi_i f_{\theta,i}(U_i)\right| \,\middle|\, \mathcal{H}\right], \tag{101}$$

where $\xi_1,\dots,\xi_N$ are independent Rademacher variables independent of the observations. For every fixed realization of the observations, the layer-cake identity yields $\sup_\theta \left|\sum_{i=1}^N \xi_i f_{\theta,i}(U_i)\right| \le \int_0^B \sup_\theta \left|\sum_{i=1}^N \xi_i \mathbf{1}\{f_{\theta,i}(U_i)\ge u\}\right| \mathrm{d}u$. For fixed u, the index sets $\{i : f_{\theta,i}(U_i)\ge u\}$ are nested as θ increases. Order the indices by the first threshold at which they enter this nested family, breaking ties arbitrarily. The inner supremum is then no larger than the maximum absolute partial sum of a permuted Rademacher sequence. Doob's L_2 maximal inequality gives

$$\mathbb{E}_\xi\left[\max_{0\le k\le N}\left|\sum_{i=1}^{k}\xi_i\right|\right] \le \left\{\mathbb{E}_\xi\left[\max_{0\le k\le N}\left|\sum_{i=1}^{k}\xi_i\right|^2\right]\right\}^{1/2} \le 2\sqrt{N}. \tag{102}$$

Integrating over u and substituting into (101) therefore gives

$$\mathbb{E}(A_N \mid \mathcal{H}) \le \frac{4B}{\sqrt{N}}. \tag{103}$$

Changing one U_i changes A_N by at most B/N. McDiarmid's inequality, combined with (103), gives (99). Applying this result to each of the K classes with failure probability δ/K and using the union bound proves (100), after increasing the universal constant. □

We now derive the exact nuisance-product decomposition. Add and subtract the source calibration average formed with the oracle density ratio $w^\star$ in (16). For every regeneration run m and threshold θ,

$$\Delta_m^q(\theta) = \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q - w^\star}\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\} + \mathbb{P}_{\mathrm{src},n}^{w^\star}\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\} - \mathbb{P}_{\mathrm{tar},n}\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\}. \tag{104}$$

Cauchy–Schwarz gives

$$\sup_{\theta\in\mathcal{S}_{m,n}}\left|\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^q - w^\star}\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\}\right| \le \left[\mathbb{P}_{\mathrm{src},n}^1\left\{\widehat{w}_m^q - w^\star\right\}^2\right]^{1/2} \times \sup_{\theta\in\mathcal{S}_{m,n}}\left[\mathbb{P}_{\mathrm{src},n}^1\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\}^2\right]^{1/2}. \tag{105}$$

For a rate-controlled candidate calibration threshold $\widehat{\theta}_m^q$, the right side of (105) is bounded by

$$C_w C_m \{\min(n_{\mathrm{src}}^{\mathrm{cal}}, n_{\mathrm{tar}}^{\mathrm{cal}})\}^{-(\beta_w+\beta_m)}. \tag{106}$$

Conditional on the nuisance-fitting data and the run randomizations, the transport identity gives, for every θ,

$$\mathbb{E}_{\mathrm{src}}[w^\star(X)\{m^\star(\theta,X) - \widehat{m}_m(\theta,X)\}] = \mathbb{E}_{\mathrm{tar}}[m^\star(\theta,X) - \widehat{m}_m(\theta,X)]. \tag{107}$$

Both $\theta \mapsto m^\star(\theta,x)$ and $\theta\mapsto \widehat{m}_m(\theta,x)$ are nondecreasing and bounded by one. After multiplication by the nonnegative oracle density ratio $w^\star$, the corresponding source functions are pointwise ordered and bounded by W_0. Applying Lemma B.1 separately to the oracle conditional score CDF $m^\star$ and the regenerated conditional score-CDF estimates $\widehat{m}_m$, then taking a union bound over the M candidates, gives, with probability at least $1-\delta/3$,

$$\max_{1\le m\le M}\sup_{\theta\in\mathbb{R}}\left|\mathbb{P}_{\mathrm{src},n}^{w^\star}\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\} - \mathbb{P}_{\mathrm{tar},n}\left\{m^\star(\theta,\cdot) - \widehat{m}_m(\theta,\cdot)\right\}\right|$$

$$\le C(W_0+1)\left[\sqrt{\frac{\log(32M/\delta)}{n_{\mathrm{src}}^{\mathrm{cal}}}} + \sqrt{\frac{\log(32M/\delta)}{n_{\mathrm{tar}}^{\mathrm{cal}}}}\right]. \tag{108}$$

The same ordered-class argument controls the other two pieces of the identity at the candidate calibration threshold. Conditional on the source covariates, fitting data, and returned weights, the functions $\theta\mapsto \widehat{w}_m^q(X_i^{\mathrm{src,cal}})\mathbf{1}\{R(X_i^{\mathrm{src,cal}}, Y_i^{\mathrm{src,cal}})\le\theta\}$ are pointwise ordered and bounded by W, and their conditional expectations are the corresponding weighted oracle conditional score CDFs $\widehat{w}_m^q(X_i^{\mathrm{src,cal}})m^\star(\theta, X_i^{\mathrm{src,cal}})$.

Hence, using Lemma B.1 with probability at least $1-\delta/3$,

$$\max_{1\le m\le M}\sup_{\theta\in\mathbb{R}}|Z_m^q(\theta)| \le CW\sqrt{\frac{\log(32M/\delta)}{n_{\mathrm{src}}^{\mathrm{cal}}}}. \tag{109}$$

Also, with probability at least $1-\delta/3$,

$$\sup_{\theta\in\mathbb{R}} |G_n(\theta) - [\mathbb{E}_{\mathrm{tar}}[m^\star(\theta, X)] - (1-\alpha)]| \le C\sqrt{\frac{\log(32M/\delta)}{n_{\mathrm{tar}}^{\mathrm{cal}}}}. \tag{110}$$

The use of M in the logarithm of (110) is only a harmless common upper bound that simplifies the final expression.

Proof of Proposition A.2. Take $\delta = (n_{\mathrm{src}}^{\mathrm{cal}} + n_{\mathrm{tar}}^{\mathrm{cal}})^{-2}$ and intersect the three events in (108)– (110). Their intersection has probability at least $1-\delta$. For every rate-controlled finite candidate calibration threshold, combine (104), (106), and (108) to obtain

$$\begin{aligned}\sup_{\theta\in\mathcal{S}_{m,n}} |\Delta_m^q(\theta)| &\le C_w C_m \{\min(n_{\mathrm{src}}^{\mathrm{cal}}, n_{\mathrm{tar}}^{\mathrm{cal}})\}^{-(\beta_w+\beta_m)} \\ &\quad + C(W_0+1)\left[\sqrt{\frac{\log\{32M(n_{\mathrm{src}}^{\mathrm{cal}}+n_{\mathrm{tar}}^{\mathrm{cal}})^2\}}{n_{\mathrm{src}}^{\mathrm{cal}}}} + \sqrt{\frac{\log\{32M(n_{\mathrm{src}}^{\mathrm{cal}}+n_{\mathrm{tar}}^{\mathrm{cal}})^2\}}{n_{\mathrm{tar}}^{\mathrm{cal}}}}\right].\end{aligned} \tag{111}$$

At the exact first nonnegative crossing, $\widehat{\Psi}_{\mathrm{DRP}}(\widehat{\theta}_m^q; \widehat{w}_m^q, \widehat{m}_m) \ge 0$, so (111) also verifies Condition A.3 at tolerance ε_n. Rearranging (13) at $\widehat{\theta}_m^q$ and then using (109), (110), and (111) yields

$$\begin{aligned}&\mathbb{E}_{\mathrm{tar}}\left[m^\star(\widehat{\theta}_m^q, X)\right] - (1-\alpha) \\ &\quad \ge -|Z_m^q(\widehat{\theta}_m^q)| - |\Delta_m^q(\widehat{\theta}_m^q)| - \sup_{\theta\in\mathbb{R}} |G_n(\theta) - [\mathbb{E}_{\mathrm{tar}}[m^\star(\theta, X)] - (1-\alpha)]| \\ &\quad \ge -\varepsilon_n.\end{aligned}$$

This proves (33) simultaneously for all rate-controlled candidate calibration thresholds. A candidate calibration threshold $\widehat{\theta}_m^q = +\infty$ has target-population coverage one.

Finally, $\beta_w + \beta_m > 0$ makes the first term in (32) vanish. Under (34), each square-root term also vanishes, because its numerator is of smaller order than $\min(n_{\mathrm{src}}^{\mathrm{cal}}, n_{\mathrm{tar}}^{\mathrm{cal}})$. Hence the displayed sample-size tolerance converges to zero as claimed. □

B.3 Finite-run amplification and coverage transfer through union

Let S_m^q denote the event that regeneration run m, under candidate construction $q \in \{\mathrm{prop}, \mathrm{tilt}\}$, produces a successful candidate calibration threshold. On the high-probability event in Condition A.5,

$$\begin{aligned}&\mathbb{P}(S_m^q \mid \text{complete shared fitting and calibration data}) \ge p_{\varepsilon_n}, \\ &\mathbb{P}((S_m^q)^c \mid \text{complete shared fitting and calibration data}) \le 1 - p_{\varepsilon_n}.\end{aligned} \tag{112}$$

Because the run-specific randomizations are conditionally independent across m, the failure events $(S_1^q)^c, \ldots, (S_M^q)^c$ are conditionally independent, and therefore

$$\begin{aligned}&\mathbb{P}\left(\bigcap_{m=1}^{M} (S_m^q)^c \,\middle|\, \text{complete shared fitting and calibration data}\right) \\ &\quad = \prod_{m=1}^{M} \mathbb{P}((S_m^q)^c \mid \text{complete shared fitting and calibration data}) \\ &\quad \le (1-p_{\varepsilon_n})^M \le \exp(-Mp_{\varepsilon_n}).\end{aligned} \tag{113}$$

The last inequality follows from $\log(1-x) \le -x$ for $x \in [0,1)$: multiplying by M and exponentiating gives $(1-x)^M = \exp\{M\log(1-x)\} \le \exp(-Mx)$, with $x = p_{\varepsilon_n}$. Thus repetition does not improve the global convergence rate of either nuisance learner. It reduces the probability that every run misses the downstream successful-candidate region. Unconditioning adds only the $o(1)$ probability of the exceptional shared-data event on which the conditional lower bound p_{ε_n} is unavailable.

The algorithm does not identify which candidate is successful, and such identification is unnecessary. For every realized candidate collection, the oracle conditional score CDF $m^\star(\theta, x)$ is nondecreasing in the threshold, so

$$\mathbb{E}_{\mathrm{tar}}\left[m^\star\left(\max_{1\le m\le M} \widehat{\theta}_m^q, X\right)\right] = \max_{1\le m\le M} \mathbb{E}_{\mathrm{tar}}\left[m^\star(\widehat{\theta}_m^q, X)\right]. \tag{114}$$

Proof of Theorem A.1. On the event $\bigcup_{m=1}^{M} S_m^q$, at least one run is successful. Proposition A.1 supplies a candidate with target-population coverage at least $1-\alpha-\varepsilon_n-o_{\mathbb{P}}(1)$. By (114), the propagated maximum has coverage no smaller than that candidate. Therefore, for every fixed $\eta>0$, up to an event of probability $o(1)$ generated by the uniform empirical-process remainder,

$$\left\{\mathbb{E}_{\mathrm{tar}}\left[m^\star\left(\max_{1\le m\le M}\widehat{\theta}_m^q, X\right)\right] < 1-\alpha-\varepsilon_n-\eta\right\} \subseteq \bigcap_{m=1}^{M}(S_m^q)^c.$$

Using (113) and then unconditioning gives

$$\mathbb{P}\left\{\mathbb{E}_{\mathrm{tar}}\left[m^\star\left(\max_{1\le m\le M}\widehat{\theta}_m^q, X\right)\right] < 1-\alpha-\varepsilon_n-\eta\right\} \le (1-p_{\varepsilon_n})^M + o(1),$$

which is (37).

The baseline-safe threshold $\widehat{\theta}^q_{\mathrm{safe}}(M)$ satisfies $\widehat{\theta}^q_{\mathrm{safe}}(M) \ge \max_{1\le m\le M}\widehat{\theta}^q_m$. Monotonicity of the target score distribution therefore gives the same target-population coverage bound for $\widehat{\theta}^q_{\mathrm{safe}}(M)$. For $q=\mathrm{tilt}$, the candidate maximum is the final DPC calibration threshold $\widehat{\theta}_{\mathrm{DPC}}$ in (9).

For the joint marginal statement, conditional on all observed data and algorithmic randomization, the coverage probability of an independent new target observation $(X_j^{\mathrm{tar,new}}, Y_j^{\mathrm{tar,new}})$ is $\mathbb{E}_{\mathrm{tar}}\left[m^\star\left(\max_{1\le m\le M}\widehat{\theta}_m^q, X\right)\right]$. This quantity lies in $[0,1]$. Hence

$$\begin{aligned}
&\mathbb{P}\left\{R(X_j^{\mathrm{tar,new}}, Y_j^{\mathrm{tar,new}}) \le \max_{1\le m\le M}\widehat{\theta}_m^q\right\}\\
&= \mathbb{E}\left[\mathbb{E}_{\mathrm{tar}}\left[m^\star\left(\max_{1\le m\le M}\widehat{\theta}_m^q, X\right)\right]\right]\\
&\ge (1-\alpha-\varepsilon_n-\eta)\mathbb{P}\left\{\mathbb{E}_{\mathrm{tar}}\left[m^\star\left(\max_{1\le m\le M}\widehat{\theta}_m^q, X\right)\right] \ge 1-\alpha-\varepsilon_n-\eta\right\}\\
&\ge 1-\alpha-\varepsilon_n-\eta-\mathbb{P}\left\{\mathbb{E}_{\mathrm{tar}}\left[m^\star\left(\max_{1\le m\le M}\widehat{\theta}_m^q, X\right)\right] < 1-\alpha-\varepsilon_n-\eta\right\}\\
&\ge 1-\alpha-\varepsilon_n-\eta-(1-p_{\varepsilon_n})^M - o(1).
\end{aligned}$$

This is (38). □

C Technical Derivations for Tilting Step Efficiency

C.1 Proof of the information-projection characterization

Proof of Proposition A.3. Fix any probability distribution P satisfying the support restriction and balance constraint in (52). By (51), on the support of $\widehat{\mathbb{P}}_m^{\mathrm{prop}}$,

$$\log\frac{\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\}}{\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}} = \widehat{\gamma}_m^{\mathsf{T}}\widetilde{g}_m(X_i^{\mathrm{src,cal}}) - \log\left[\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_\ell^{\mathrm{src,cal}}\}\exp\left\{\widehat{\gamma}_m^{\mathsf{T}}\widetilde{g}_m(X_\ell^{\mathrm{src,cal}})\right\}\right].$$

Expand the KL divergence from P to $\widehat{\mathbb{P}}_m^{\mathrm{prop}}$ by inserting $\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$:

$$D_{\mathrm{KL}}\left(P\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{prop}}\right) = D_{\mathrm{KL}}\left(P\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\right) + \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} P\{X_i^{\mathrm{src,cal}}\}\log\frac{\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\}}{\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}}.$$

Both P and $\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$ satisfy (50). Their means of $\widetilde{g}_m$ are therefore identical, and both have total mass one. Substituting the preceding log-ratio formula shows that

$$\sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}} P\{X_i^{\mathrm{src,cal}}\}\log\frac{\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\}}{\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}} = \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\}\log\frac{\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\}}{\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}} = D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{prop}}\right).$$

Hence the exact Pythagorean identity

$$D_{\mathrm{KL}}\left(P\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{prop}}\right) = D_{\mathrm{KL}}\left(P\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\right) + D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{prop}}\right)$$

holds for every feasible P. Nonnegativity of KL divergence proves that $\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$ minimizes the objective. Equality can hold only when $P=\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$ on the common support, which proves uniqueness. □

C.2 Proof of the local KL perturbation theorem

Proof of Theorem A.2. Differentiate the probability mass in (53). Direct differentiation of the numerator and normalizing denominator gives, for every source calibration covariate,

$$\frac{d}{du}\widehat{\mathbb{P}}_{m,u}\{X_i^{\mathrm{src,cal}}\} = \widehat{\mathbb{P}}_{m,u}\{X_i^{\mathrm{src,cal}}\}\widehat{\gamma}_m^{\mathsf{T}}\left[\widetilde{g}_m(X_i^{\mathrm{src,cal}}) - \sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_{m,u}\{X_\ell^{\mathrm{src,cal}}\}\widetilde{g}_m(X_\ell^{\mathrm{src,cal}})\right].$$

Multiplying by $\widetilde{g}_m(X_i^{\mathrm{src,cal}})$ and summing over i yields

$$\frac{d}{du}\sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_{m,u}\{X_i^{\mathrm{src,cal}}\}\widetilde{g}_m(X_i^{\mathrm{src,cal}}) = \mathrm{Cov}_{\widehat{\mathbb{P}}_{m,u}}\{\widetilde{g}_m(X)\}\,\widehat{\gamma}_m.$$

Integrating from $u=0$ to $u=1$, using $\widehat{\mathbb{P}}_{m,0} = \widehat{\mathbb{P}}_m^{\mathrm{prop}}$, $\widehat{\mathbb{P}}_{m,1} = \widehat{\mathbb{P}}_m^{\mathrm{tilt}}$, and exact balance at $u=1$, gives

$$\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m - \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}\widetilde{g}_m(X_i^{\mathrm{src,cal}}) = \left[\int_0^1 \mathrm{Cov}_{\widehat{\mathbb{P}}_{m,u}}\{\widetilde{g}_m(X)\}\,du\right]\widehat{\gamma}_m. \tag{115}$$

We next bound the feature-mean discrepancy on the left. By (47),

$$\begin{aligned}&\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m - \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}\widetilde{g}_m(X_i^{\mathrm{src,cal}})\\ &\quad= \frac{1}{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}1}\left[\left(\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}1 - 1\right)\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m + \mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m - \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}\widetilde{g}_m\right].\end{aligned}$$

The intercept coordinate in (44) gives

$$\left|\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}1 - 1\right| \le \mathcal{B}_{m,n},$$

and every coordinate of $\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m$ lies in $[0,1]$. The remaining coordinatewise difference is also bounded by $\mathcal{B}_{m,n}$. Therefore

$$\left\|\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m - \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}\widetilde{g}_m(X_i^{\mathrm{src,cal}})\right\|_2 \le \frac{2\sqrt{K_m}}{\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}1}\mathcal{B}_{m,n}. \tag{116}$$

By the definition of $\mu_{m,n}$ and (115),

$$\mu_{m,n}\left\|\widehat{\gamma}_m\right\|_2 \le \left\|\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m - \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}\widetilde{g}_m(X_i^{\mathrm{src,cal}})\right\|_2.$$

Combining this inequality with (116) proves (55).

It remains to control the KL divergence. Taking logarithms in (51) and averaging under $\widehat{\mathbb{P}}_m^{\mathrm{tilt}}$ gives

$$D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{prop}}\right) = \widehat{\gamma}_m^{\mathsf{T}}\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m - \log\left[\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_\ell^{\mathrm{src,cal}}\}\exp\left\{\widehat{\gamma}_m^{\mathsf{T}}\widetilde{g}_m(X_\ell^{\mathrm{src,cal}})\right\}\right].$$

Jensen's inequality implies

$$\log\left[\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_\ell^{\mathrm{src,cal}}\}\exp\left\{\widehat{\gamma}_m^{\mathsf{T}}\widetilde{g}_m(X_\ell^{\mathrm{src,cal}})\right\}\right] \ge \widehat{\gamma}_m^{\mathsf{T}}\sum_{\ell=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_\ell^{\mathrm{src,cal}}\}\widetilde{g}_m(X_\ell^{\mathrm{src,cal}}).$$

Consequently,

$$\begin{aligned}D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{prop}}\right) &\le \left\|\widehat{\gamma}_m\right\|_2\left\|\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m - \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}\widetilde{g}_m(X_i^{\mathrm{src,cal}})\right\|_2\\ &\le \frac{1}{\mu_{m,n}}\left\|\mathbb{P}_{\mathrm{tar},n}\widetilde{g}_m - \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}\widetilde{g}_m(X_i^{\mathrm{src,cal}})\right\|_2^2.\end{aligned}$$

Applying (116) proves (56). If the preliminary weighted mass stays bounded away from zero and (57) holds, the right side of (56) converges to zero in probability, proving the final claim. □

C.3 Proof of candidate-tolerance contraction

Proof of Theorem A.3. Substituting (61) into (59) and applying exact balance gives (62). We now control the difference between the two remainder averages.

For any bounded real-valued function h on the source calibration support, Pinsker's inequality gives

$$\left|\sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\left[\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\}-\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}\right]h(X_i^{\mathrm{src,cal}})\right| \le \left[\max_{1\le i\le n_{\mathrm{src}}^{\mathrm{cal}}} h(X_i^{\mathrm{src,cal}})-\min_{1\le i\le n_{\mathrm{src}}^{\mathrm{cal}}} h(X_i^{\mathrm{src,cal}})\right]\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{prop}}\right)}. \tag{117}$$

Apply this inequality with $h = r_m(\theta,\cdot)$. From (48) and (47),

$$\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} r_m(\theta,\cdot) - \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} r_m(\theta,\cdot)$$
$$= \sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\left[\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\{X_i^{\mathrm{src,cal}}\}-\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\}\right] r_m(\theta,X_i^{\mathrm{src,cal}}) + \left[1-\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}1\right]\sum_{i=1}^{n_{\mathrm{src}}^{\mathrm{cal}}}\widehat{\mathbb{P}}_m^{\mathrm{prop}}\{X_i^{\mathrm{src,cal}}\} r_m(\theta,X_i^{\mathrm{src,cal}}).$$

The intercept coordinate in (44) implies

$$\left|1-\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}1\right| \le \mathcal{B}_{m,n}. \tag{118}$$

The oscillation of $r_m(\theta,\cdot)$ over the source support is no larger than twice its maximum absolute value. Combining (117)–(118) therefore gives

$$\begin{aligned}
&\left|\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m^{\mathrm{tilt}}} r_m(\theta,\cdot) - \mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} r_m(\theta,\cdot)\right| \\
&\quad\le \max_{1\le i\le n_{\mathrm{src}}^{\mathrm{cal}}}\left|r_m(\theta,X_i^{\mathrm{src,cal}})\right|\left[\mathcal{B}_{m,n}+\sqrt{2D_{\mathrm{KL}}\left(\widehat{\mathbb{P}}_m^{\mathrm{tilt}}\,\middle\|\,\widehat{\mathbb{P}}_m^{\mathrm{prop}}\right)}\right] \\
&\quad\le \max_{1\le i\le n_{\mathrm{src}}^{\mathrm{cal}}}\left|r_m(\theta,X_i^{\mathrm{src,cal}})\right|\mathcal{B}_{m,n}\left[1+\frac{2\sqrt{2K_m}}{\sqrt{\mu_{m,n}}\,\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}1}\right],
\end{aligned} \tag{119}$$

where the final inequality uses (56).

For propagation alone, (62) and Hölder's inequality yield

$$|\Delta_m(\theta)| \le \|a_m(\theta)\|_1\,\mathcal{B}_{m,n} + \left|\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} r_m(\theta,\cdot) - \mathbb{P}_{\mathrm{tar},n} r_m(\theta,\cdot)\right|, \tag{120}$$

$$\left|\Delta_m^{\mathrm{tilt}}(\theta)\right| \le \left|\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m} r_m(\theta,\cdot) - \mathbb{P}_{\mathrm{tar},n} r_m(\theta,\cdot)\right| + \max_{1\le i\le n_{\mathrm{src}}^{\mathrm{cal}}}\left|r_m(\theta,X_i^{\mathrm{src,cal}})\right|\mathcal{B}_{m,n}\left[1+\frac{2\sqrt{2K_m}}{\sqrt{\mu_{m,n}}\,\mathbb{P}_{\mathrm{src},n}^{\widehat{w}_m}1}\right].$$

Taking the supremum over $\theta\in\mathcal{S}_{m,n}$ and using (63) proves (68). The implication (69) then follows directly from Condition A.3, because the theorem assumes exact first nonnegative crossings.

Subtracting the two expressions in (66) shows that (70) is sufficient for the tilted certificate to be no larger. If the coefficient inequality is strict and $\mathcal{B}_{m,n} > 0$, the certificate inequality is strict.

Finally, under (71),

$$\mathcal{V}_{m,n} = o_{\mathbb{P}}(A_{m,n}\mathcal{B}_{m,n}),$$

so

$$\bar{\varepsilon}_{m,n}^{\mathrm{prop}} = A_{m,n}\mathcal{B}_{m,n}\{1+o_{\mathbb{P}}(1)\}.$$

Under (72), the additional tilted term is also $o_{\mathbb{P}}(A_{m,n}\mathcal{B}_{m,n})$. Together with the same bound for $\mathcal{V}_{m,n}$, this proves (73); division by the propagation-only certificate proves (74). □

C.4 Proof of effective propagation under the tilting step

Proof of Theorem A.4. First verify the certified-mass comparison. Decompose each certified success event into its intersection with the other event and its one-sided difference. Subtracting the resulting probabilities gives exactly

$$\underline{p}_{\varepsilon_n}^{\mathrm{tilt}} - \underline{p}_{\varepsilon_n}^{\mathrm{prop}} = h_{\varepsilon_n}^{+} - h_{\varepsilon_n}^{-},$$

which is (80). Under (81), Theorem A.3 gives

$$\bar{\varepsilon}_{m,n}^{\mathrm{tilt}} \le \bar{\varepsilon}_{m,n}^{\mathrm{prop}}$$

for every compared run. Hence $C_m^{\mathrm{prop}}(\varepsilon_n) \subseteq C_m^{\mathrm{tilt}}(\varepsilon_n)$, so $h_{\varepsilon_n}^- = 0$ and $h_{\varepsilon_n}^+ = \pi_{\varepsilon_n}$. This proves (84) and (86).

Because $C_m^q(\varepsilon_n) \subseteq S_m^q(\varepsilon_n)$, failure of every actual candidate implies that no run is certified. Conditional independence of the run-specific randomizations therefore gives

$$\mathbb{P}\left\{\bigcap_{m=1}^{M} S_m^q(\varepsilon_n)^c \,\middle|\, \text{shared data}\right\} \le \mathbb{P}\left\{\bigcap_{m=1}^{M} C_m^q(\varepsilon_n)^c \,\middle|\, \text{shared data}\right\} \le \left(1-\underline{p}_{\varepsilon_n}^q\right)^M \le \exp\left\{-M\underline{p}_{\varepsilon_n}^q\right\}.$$

This is the same amplification argument already proved in (35), now applied to the certified lower bound. Proposition A.1 and the union identity (36) then give (85); baseline retention can only increase the threshold. No additional coverage argument is required.

The ordering (87) follows immediately because $p \mapsto (1-p)^M$ is decreasing. More explicitly,

$$\left(1-\underline{p}_{\varepsilon_n}^{\mathrm{tilt}}\right)^M = \left(1-\underline{p}_{\varepsilon_n}^{\mathrm{prop}}-\pi_{\varepsilon_n}\right)^M \le \left(1-\underline{p}_{\varepsilon_n}^{\mathrm{prop}}\right)^M. \tag{121}$$

Since $p \mapsto -\log(1-p)$ is increasing on $(0,1)$, (86) yields (89). Strict certified-mass improvement gives a strict comparison before integer rounding.

For the effective number of regeneration runs, the floor in (90) is no larger than the unrounded quantity. Since $\log(1-\underline{p}_{\varepsilon_n}^{\mathrm{prop}}) < 0$,

$$\left\lfloor M\frac{-\log(1-\underline{p}_{\varepsilon_n}^{\mathrm{tilt}})}{-\log(1-\underline{p}_{\varepsilon_n}^{\mathrm{prop}})}\right\rfloor \log(1-\underline{p}_{\varepsilon_n}^{\mathrm{prop}}) \ge M\log(1-\underline{p}_{\varepsilon_n}^{\mathrm{tilt}}).$$

Exponentiating both sides gives

$$\left(1-\underline{p}_{\varepsilon_n}^{\mathrm{tilt}}\right)^M \le \left(1-\underline{p}_{\varepsilon_n}^{\mathrm{prop}}\right)^{\left\lfloor M\frac{-\log(1-\underline{p}_{\varepsilon_n}^{\mathrm{tilt}})}{-\log(1-\underline{p}_{\varepsilon_n}^{\mathrm{prop}})}\right\rfloor},$$

which proves the final claim. The ratio is at least one because $\underline{p}_{\varepsilon_n}^{\mathrm{tilt}} \ge \underline{p}_{\varepsilon_n}^{\mathrm{prop}}$, and it is strictly greater than one when $\pi_{\varepsilon_n} > 0$. □

D Additional Numerical Details

D.1 Simulation Design

This appendix records how the difficulty grid changes each synthetic data-generating process. These details are separated from Section 4 because the main numerical section uses only the source-target notation of Sections 2–3.

All synthetic designs use latent covariates $U \in \mathbb{R}^5$. The latent covariates U are first generated and then transformed coordinatewise to produce the observed covariates X. For the first five coordinates, the transformation is

$$\begin{aligned}
X_1 &= U_1 + 0.10\sin(0.65U_1) + 0.035\tanh(0.60U_1) + 0.006U_1^3,\\
X_2 &= U_2 + 0.08\{1-\cos(0.60U_2)\} + 0.035\,\mathrm{asinh}(U_2) + 0.006U_2^3,\\
X_3 &= U_3 + 0.090\tanh(0.70U_3) + 0.007U_3^3,\\
X_4 &= U_4 + 0.075\,\mathrm{asinh}(U_4) + 0.006U_4^3,\\
X_5 &= U_5 + 0.070\{\log(1+\exp(0.70U_5)) - \log 2\} + 0.006U_5^3.
\end{aligned}$$

For the additive mean shift, variance-tail shift, and correlation-interaction shift, the conditional score distribution is $R(X,Y) = |\sigma(X)Z|, \qquad Z \sim \mathcal{N}(0,1)$, where

$$\sigma(X) = \exp\left[-0.62 + 0.008\log\{1+\exp((s(X)-1.30)/1.20)\}\right],$$

$$s(X) = 0.70\,\frac{\sum_{j=1}^5 X_j}{\sqrt{5}} + 0.08\,\frac{1}{5}\sum_{j=1}^5 \sin(X_j) + 0.04\,\frac{1}{5}\sum_{j=1}^5 \tanh(0.70X_j).$$

The score-scale formula is fixed across the grid. Thus the grid changes $P_{\mathrm{tar},X}$, not the conditional law of $R(X,Y) \mid X$. The Oracle interval width changes because the marginal target distribution of $\sigma(X)$ changes.

D.1.1 Additive mean shift.

For each reported difficulty

$$\|\mu\|_2 \in \{0, 2, \ldots, 20\},$$

the latent source, target, and new target covariates are generated as

$$U_{\mathrm{src}} \sim \mathcal{N}(0, I_5), \qquad U_{\mathrm{tar}}, U_{\mathrm{tar}}^{\mathrm{new}} \sim \mathcal{N}(\mu, I_5), \qquad \mu = \frac{\|\mu\|_2}{\sqrt{5}} \mathbf{1}_5.$$

Indeed,

$$\left\| \frac{\|\mu\|_2}{\sqrt{5}} \mathbf{1}_5 \right\|_2 = \|\mu\|_2,$$

so each value in the reported grid enters the data generating process directly as the ℓ_2 magnitude of the target latent mean shift. In particular, $\|\mu\|_2 = 0$ gives identical source and target latent laws, whereas larger grid values shift all five target coordinates equally. The common coordinatewise transformation above is then applied to obtain X_i^{src}, X_j^{tar}, and $X_j^{\mathrm{tar,new}}$; the latent covariance and the conditional score law $R(X, Y) \mid X$ remain unchanged throughout the grid.

D.1.2 VARIANCE-TAIL SHIFT.

This design retains the additive mean shift

$$\mu = \frac{\|\mu\|_2}{\sqrt{5}} \mathbf{1}_5, \qquad \|\mu\|_2 \in \{0, 2, \ldots, 20\},$$

and introduces a high-variance latent component. The source law is

$$U_{\mathrm{src}} \sim 0.95\,\mathcal{N}(0, I_5) + 0.05\,\mathcal{N}(0, 1.80^2 I_5),$$

while, over the reported grid, the target and new target laws are

$$U_{\mathrm{tar}}, U_{\mathrm{tar}}^{\mathrm{new}} \sim \left(0.95 - \frac{0.03}{4.15}\|\mu\|_2\right) \mathcal{N}(\mu, I_5) + \left(0.05 + \frac{0.03}{4.15}\|\mu\|_2\right) \mathcal{N}(\mu, 1.80^2 I_5).$$

Thus both target mixture components retain the same mean shift μ, while the high-variance component becomes increasingly prevalent as $\|\mu\|_2$ grows. The conditional score law $R(X, Y) \mid X$ remains unchanged.

D.1.3 CORRELATION-INTERACTION SHIFT.

This design retains the additive mean shift

$$\mu = \frac{\|\mu\|_2}{\sqrt{5}} \mathbf{1}_5, \qquad \|\mu\|_2 \in \{0, 2, \ldots, 20\},$$

and introduces a correlated latent component. The source law is

$$U_{\mathrm{src}} \sim 0.92\,\mathcal{N}(0, I_5) + 0.08\,\mathcal{N}(0,\, \Sigma_{\mathrm{corr}})\,,$$

while, over the reported grid, the target and new target laws are

$$U_{\mathrm{tar}}, U_{\mathrm{tar}}^{\mathrm{new}} \sim \left(0.92 - \frac{0.01}{4.15}\|\mu\|_2\right) \mathcal{N}(\mu, I_5) + \left(0.08 + \frac{0.01}{4.15}\|\mu\|_2\right) \mathcal{N}(\mu,\, \Sigma_{\mathrm{corr}})\,.$$

Thus both target mixture components retain the same mean shift μ, while the correlated component becomes increasingly prevalent as $\|\mu\|_2$ grows. This component has unit marginal variances and pairwise correlation 0.40; the conditional score law $R(X, Y) \mid X$ remains unchanged.

D.2 CALIFORNIA HOUSING SAMPLING DESIGN

This appendix describes how the California Housing records are converted into labeled source data $\mathcal{D}_{\mathrm{src}}$ and unlabeled target covariates $\mathcal{D}_{\mathrm{tar}}$ under controlled covariate shift. The data set contains $20{,}640$ census block groups. The outcome Y is the median house value, measured in units of $100{,}000$ dollars, and the covariate vector $X \in \mathbb{R}^8$ contains median income, median house age, average rooms, average bedrooms, population, average occupancy, latitude, and longitude. All covariates and outcomes are taken directly from the original data set. In particular, no response is simulated, modified, or used to determine whether a record is sampled.

The source and target covariate distributions are constructed by assigning different sampling weights to the observed records. In the formulas below, x denotes the covariate vector on the scale used to form these sampling weights. Define the low-occupancy score

$$\ell(x) = 0.45\{-x_{\mathrm{AveOccup}}\} + 0.55\max\{-x_{\mathrm{AveOccup}}, 0\}.$$

Larger values of $\ell(x)$ correspond to records with lower average occupancy.

The experiment considers three settings: the Additive setting, the Variance-tail setting, and the Correlation-interaction setting. Let s index one of these three settings, and let $d_s(x)$ denote its direction of the shift, which is specified in the corresponding subsubsection below. For

$$\gamma \in \{0, 0.5, 1, \ldots, 5\},$$

define the source and target sampling weights by

$$\omega_{\mathrm{src}}(x) = e^{-0.50\ell(x)}, \qquad \omega_{\mathrm{tar}}^{(s,\gamma)}(x) = \omega_{\mathrm{src}}(x)e^{0.70\gamma d_s(x)}.$$

The factor $e^{-0.50\ell(x)}$ defines the common source construction used in all three settings and down-weights records with large low-occupancy scores. The target construction starts from the same factor and adds the setting-specific tilt $e^{0.70\gamma d_s(x)}$. Before normalization within a sampling pool, the target-to-source sampling-weight ratio is therefore

$$\frac{\omega_{\mathrm{tar}}^{(s,\gamma)}(x)}{\omega_{\mathrm{src}}(x)} = e^{0.70\gamma d_s(x)}.$$

Hence $\gamma = 0$ gives identical source sampling weights $\omega_{\mathrm{src}}(x)$ and target sampling weights $\omega_{\mathrm{tar}}^{(s,0)}(x)$, whereas increasing γ makes the target sample increasingly favor records with large values of $d_s(x)$. The source sampling weight $\omega_{\mathrm{src}}(x)$ does not change with either s or γ.

For each Monte Carlo replication, the full data set is randomly partitioned into four disjoint pools of 5,160 records:

point-prediction model fitting, labeled source calibration,
unlabeled target calibration, prediction on new target observations.

The two source pools use ω_{src}, while the two target pools use $\omega_{\mathrm{tar}}^{(s,\gamma)}$ for the selected setting and grid value.

A sample of size 1000 is drawn without replacement from each source pool. Specifically, if $\mathcal{P}_t$ is the set of records remaining before source draw t, then record $i \in \mathcal{P}_t$ is selected with conditional probability

$$\mathbb{P}(i \text{ is selected at draw } t \mid \mathcal{P}_t) = \frac{\omega_{\mathrm{src}}(x_i)}{\sum_{k\in\mathcal{P}_t}\omega_{\mathrm{src}}(x_k)},$$

after which record i is removed from the pool. Thus the point-prediction model fitting sample and the labeled source sample $\mathcal{D}_{\mathrm{src}}$ are distinct but follow the same weighted source construction.

A sample of size 1000 is drawn independently with replacement from each target pool. For a target pool $\mathcal{P}$, every draw selects record $i \in \mathcal{P}$ with probability

$$\mathbb{P}_{s,\gamma}(i \mid \mathcal{P}) = \frac{\omega_{\mathrm{tar}}^{(s,\gamma)}(x_i)}{\sum_{k\in\mathcal{P}}\omega_{\mathrm{tar}}^{(s,\gamma)}(x_k)}.$$

The unlabeled target sample $\mathcal{D}_{\mathrm{tar}}$ and the independent new target sample $\mathcal{D}_{\mathrm{tar}}^{\mathrm{new}}$ therefore follow the same setting-specific target construction but are drawn independently from disjoint pools.

The fitted regression function $\widehat{\mu}$ is fitted using XGBoost only on the point-prediction model fitting sample. The absolute-residual score R used for calibration and evaluation is

$$R(X_i^{\mathrm{src}}, Y_i^{\mathrm{src}}) = |Y_i^{\mathrm{src}} - \widehat{\mu}(X_i^{\mathrm{src}})|, \qquad R(X_j^{\mathrm{tar,new}}, Y_j^{\mathrm{tar,new}}) = \left|Y_j^{\mathrm{tar,new}} - \widehat{\mu}(X_j^{\mathrm{tar,new}})\right|.$$

The conformal methods receive the labeled source sample $\mathcal{D}_{\mathrm{src}}$ and the unlabeled target covariates $\mathcal{D}_{\mathrm{tar}}$. New target outcomes $Y_j^{\mathrm{tar,new}}$ are not used during point-prediction model fitting or calibration; they are reserved for the final coverage calculation and the infeasible Oracle benchmark.

Only the direction of the shift $d_s(x)$ changes across the three real-data settings; the source sampling weight $\omega_{\mathrm{src}}(x)$, the difficulty grid for γ, and the sampling-pool construction remain fixed.

D.2.1 ADDITIVE SETTING.

For the Additive setting, $s = \mathrm{add}$ and

$$d_{\mathrm{add}}(x) = \ell(x).$$

Hence the target-to-source sampling-weight ratio is

$$\frac{\omega_{\mathrm{tar}}^{(\mathrm{add},\gamma)}(x)}{\omega_{\mathrm{src}}(x)} = e^{0.70\gamma\ell(x)}.$$

Increasing γ therefore makes the target sample increasingly favor census block groups with larger $\ell(x)$, corresponding to lower average occupancy.

D.2.2 VARIANCE-TAIL SETTING.

For the Variance-tail setting, $s = \mathrm{var}$ and the direction of the shift is

$$d_{\mathrm{var}}(x) = \ell(x) + 0.55\log[1 + \max\{\ell(x), 0\}],$$

with the two components placed on the common scale used in the sampling construction. Consequently,

$$\frac{\omega_{\mathrm{tar}}^{(\mathrm{var},\gamma)}(x)}{\omega_{\mathrm{src}}(x)} = e^{0.70\gamma d_{\mathrm{var}}(x)}.$$

The first term preserves the additive low-occupancy shift, while the second term makes the target sample increasingly favor the positive tail of that direction as γ grows.

D.2.3 CORRELATION-INTERACTION SETTING.

For the Correlation-interaction setting, $s = \mathrm{corr}$ and
$d_{\mathrm{corr}}(x) = \ell(x) + 0.55\max\{-x_{\mathrm{AveOccup}}, 0\}\left[0.70\max\{-x_{\mathrm{Population}}, 0\} + 0.30|x_{\mathrm{AveRooms}}|\right]$,
again with the components placed on the common scale used in the sampling construction. Thus

$$\frac{\omega_{\mathrm{tar}}^{(\mathrm{corr},\gamma)}(x)}{\omega_{\mathrm{src}}(x)} = e^{0.70\gamma d_{\mathrm{corr}}(x)}.$$

Increasing γ preserves the low-occupancy shift while additionally favoring census block groups in which low occupancy occurs jointly with low population or an atypical average-room value.