# BAYESIAN CONFORMAL RISK CONTROL UNDER COVARIATE SHIFT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Machine learning predictors are rarely deployed on data that match the data they were calibrated on. Conformal prediction and its risk-control generalization promise distribution-free guarantees at deployment, but only under exchangeability or a covariate shift whose likelihood ratio is known exactly—and the estimated-ratio case, the one that occurs in practice, is explicitly open, out of reach of the Bayesian construction available under exchangeability. We reach it from a different starting point: pushing a Dirichlet process prior on the calibration distribution through the likelihood ratio yields the exact Bayesian posterior over the deployed risk. That posterior is exact given the weight function, which at deployment is itself estimated, so we further prove finite-sample upper and lower bounds on the risk the selected threshold actually incurs, in terms of the weight-estimation error and the conditional shift. On synthetic and real covariate shifts, in regression and classification, the realized failure rate then stays close to target where shift-blind selection fails, and what deployment receives is a posterior over the risk of every threshold rather than a single number.

## 1 INTRODUCTION

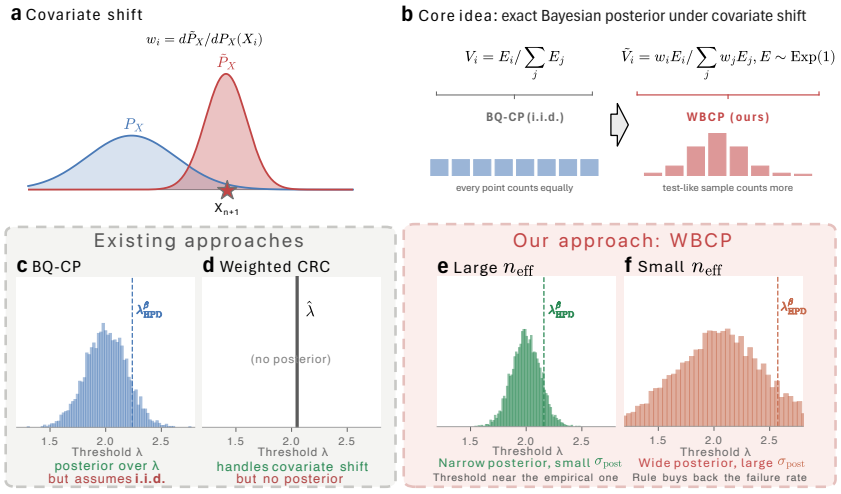



Figure 1: **WBCP overview.** (a) Calibration vs. test covariate distribution, with the likelihood ratio $w_i = d\tilde{P}_X/dP_X(X_i)$ shading the calibration points. (b) The core step: the same Bayesian-bootstrap masses, reweighted by $w_i$, give the exact posterior under shift (Proposition 1)—no Dirichlet is assumed. (c–d) Existing approaches give either a posterior *or* shift-robustness, not both. (e–f) WBCP gives both: a posterior over the deployed risk, hence a threshold selected at a target failure rate, whose spread $\sigma_{\mathrm{post}}$ tracks $n_{\mathrm{eff}}$.

Machine learning systems are increasingly deployed in settings that differ from the ones in which they were calibrated: a poverty-mapping model calibrated on surveyed countries is applied to unsurveyed ones, a land-use classifier trained on one region scores imagery from another. In such settings one wants a guarantee on the loss the deployed predictor will incur—and one wants to know how much that guarantee can be trusted, given that it was computed from a finite calibration set the deployment distribution no longer matches.



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.

Distribution-free methods answer the first question under idealized conditions. Split conformal prediction (Vovk et al., 2005; Papadopoulos et al., 2002) and conformal risk control (Angelopoulos et al., 2024) certify a target risk when calibration and deployment are exchangeable; weighted conformal prediction (Tibshirani et al., 2019) and CRC's covariate-shift extension restore the guarantee when the likelihood ratio between the two covariate distributions is known. The case that occurs in practice—an *estimated* ratio—is explicitly open: the CRC authors "leave the full investigation of weighted conformal risk control with an estimated covariate shift for future research" (Angelopoulos et al., 2024, §4.1). The second question is Bayesian, and Snell & Griffiths (2025) answered it for the exchangeable case: reading conformal prediction as Bayesian quadrature (BQ-CP) yields a posterior distribution over the risk of any threshold, and with it a guarantee conditional on the observed calibration set. That construction, as its authors note, assumes the deployment data are i.i.d. with the calibration data.

The assumption is structural, not incidental. Under exchangeability the Dirichlet law of the quantile spacings is an exact frequentist identity—a consequence of the probability integral transform—so BQ-CP's Bayesian credibility and frequentist confidence coincide by construction. Under covariate shift the identity fails: the transformed ranks are no longer uniform and their law depends on the unknown likelihood ratio. The i.i.d. posterior does not merely lose its guarantee; it loses its derivation.

This paper supplies both halves for the deployed setting. First, we derive the exact posterior over the deployed risk curve: a Dirichlet process prior on the calibration distribution, pushed through the likelihood ratio, induces the normalized weighted Bayesian bootstrap, whose law is a scaled Dirichlet—Dirichlet only at uniform weights, where the exchangeable posterior is recovered. Kish's effective sample size (Kish, 1965) emerges as the unique two-moment concentration rather than being imposed, and the exact posterior samples at the same cost as its Dirichlet approximation.[1] Second, because credibility no longer coincides with confidence, we prove the theorem that replaces the lost identity: a finite-sample two-sided certificate for the deployed risk of the selected threshold. Its leading error is a *product*—weight-estimation error times the score's dependence on the covariates—so either a correct weight profile or a locally adaptive score suffices; an irreducible term prices genuine conditional shift; and the entire cost of the Bayesian selection is one computable mass. The certificate is attained by an explicit two-point construction, and each of its terms is matched to a diagnostic available at prediction time, including the term no unlabeled diagnostic can see.

Experiments on synthetic and real covariate shifts, in regression and classification, compare selection rules at their shared semantics: a target risk at a target failure rate. Where deployment is exchangeable, concentration-based selection suffices and the posterior matches it at a smaller threshold. Where the shift must be estimated, every shift-blind rule fails—the posterior of the exchangeable theory violates its target at rates that *grow* with the calibration size, since more data only concentrates its credibility on the wrong risk curve—while the weighted posterior certifies at the target failure rate, at a premium the certificate predicts.

## 2 BACKGROUND

**Data and shift.** Calibration pairs $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P = P_X \cdot P_{Y|X}$, $i = 1, \ldots, n$; an independent test pair $(X_{n+1}, Y_{n+1}) \sim \tilde{P} = \tilde{P}_X \cdot \tilde{P}_{Y|X}$. Under *pure covariate shift* $\tilde{P}_{Y|X} = P_{Y|X}$ and the shift is carried by the likelihood ratio $w^*(x) = d\tilde{P}_X / dP_X(x)$; we do not assume purity in the certificate, where the deviation is priced. A model $\hat{f}$ is fit on separate training data; nonconformity scores are $\rho_i = s(X_i, Y_i)$ (absolute residuals for regression, APS for classification), and the nested family $\mathcal{C}_\lambda(x) = \{y : s(x, y) \leq \lambda\}$ grows with $\lambda$.

**Losses and risks.** Every method in this paper answers one selection problem: pick the smallest threshold $\lambda$ whose deployed risk stays below a target $\alpha$. We set the problem up once, at the generality of a loss $\ell(\lambda; x, y) \in [0, B_\ell]$, nonincreasing and right-continuous in $\lambda$; the miscoverage loss $\ell = \mathbf{1}\{y \notin \mathcal{C}_\lambda(x)\}$ ($B_\ell = 1$) recovers coverage, and readers focused on coverage may read $\ell$ as this indicator throughout. The deployed risk is $R_{\text{test}}(\lambda) = \mathbb{E}_{\tilde{P}}[\ell(\lambda; X, Y)]$. Given weights

[1] Faster, in fact: one posterior draw is $n+1$ exponentials against the surrogate's $n+1$ gamma variates (2.8 vs. 9.1 ms for 2,000 draws at $n = 200$, NumPy, one CPU core).

$w_i = \hat{w}(X_i)$ with normalization $\tilde{w}_i = w_i / \sum_j w_j$, the weighted empirical risk and its selection are

$$\widehat{R}_w(\lambda) = \sum_{i=1}^{n} \tilde{w}_i \, \ell(\lambda; X_i, Y_i), \qquad \hat{\lambda} = \inf\big\{\lambda : \widehat{R}_w(\lambda) \le \alpha\big\}. \tag{1}$$

For miscoverage, $\hat{\lambda}$ is the weighted quantile of Tibshirani et al. (2019), whose marginal coverage guarantee under weighted exchangeability our method—weighted Bayesian conformal prediction (WBCP)—inherits unchanged (§3.3).

### 2.1 Conformal Risk Control and Its Weighted Extension

Split conformal prediction solves the miscoverage instance under exchangeability: with $\hat{\lambda}$ the conformal quantile of the calibration scores, $\Pr(Y_{n+1} \in \mathcal{C}_{\hat{\lambda}}(X_{n+1})) \ge 1 - \alpha$. Nothing in that construction is tied to the indicator loss, and CRC (Angelopoulos et al., 2024) states the same argument for the whole class: the rule $\hat{\lambda} = \inf\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B_\ell}{n+1} \le \alpha\}$ attains $\mathbb{E}[\ell(\hat{\lambda}; X_{n+1}, Y_{n+1})] \le \alpha$ for every bounded monotone loss. Under covariate shift with a *known* ratio $w^*$, replacing the empirical risk by its $w^*$-weighted version preserves the expectation guarantee (their Proposition 2); for miscoverage this is the weighted conformal quantile of Tibshirani et al. (2019). Two limits of this frequentist line matter here. First, the guarantee is an expectation over calibration draws: it carries no failure-rate parameter, so any single calibration set may yield a threshold whose deployed risk exceeds $\alpha$. Second, it requires the ratio to be known; the estimated case is open (§1).

### 2.2 Conformal Prediction as Bayesian Quadrature

BQ-CP (Snell & Griffiths, 2025) supplies the failure-rate parameter for the exchangeable case by modeling the unknown quantile levels of the calibration losses. With sorted losses $\ell_{(1)} \le \cdots \le \ell_{(n)}$ and $\ell_{(n+1)} = B_\ell$, the masses the deployment distribution places between them satisfy

$$L^+ = \sum_{i=1}^{n+1} V_i \, \ell_{(i)}, \qquad (V_1, \ldots, V_{n+1}) \sim \mathrm{Dir}(1, \ldots, 1), \tag{2}$$

and $L^+$ stochastically dominates the posterior risk of any prior on the state of nature. Selecting the smallest $\lambda$ with $\Pr(L^+ \le \alpha \mid \ell_{1:n}) \ge \beta$ therefore controls the risk at credibility $\beta$ *conditionally on the observed calibration set*. The $\mathrm{Dir}(1, \ldots, 1)$ law is exact under exchangeability—a consequence of the probability integral transform—which is also the construction's boundary: it assumes the deployment data are i.i.d. with the calibration data.

### 2.3 Summary and Prospectus

The frequentist line handles known covariate shift but has no failure-rate parameter; the Bayesian line has the failure-rate parameter but no covariate shift. The next section supplies both: a posterior over the *weighted* risk curve, and the frequentist statement that the lost identity no longer provides (§3.3).

## 3 Our Approach

What we need is a posterior over the deployed risk curve: given the calibration set, how plausible is it that a candidate threshold $\lambda$ keeps $R_{\mathrm{test}}(\lambda)$ below $\alpha$? In the i.i.d. case the spacing identity supplies such a posterior directly (§2.2). Under weighting that route is closed—the spacing law depends on the unknown ratio (Appendix A)—but the covariate-shift model supplies two facts the identity never used: the calibration sample is i.i.d. from a single law $P$, and the deployed distribution is a tilt of $P$ by the ratio $w$—the change of measure $Q \mapsto w\,Q / \int w\,dQ$. This section builds the method from these two facts: §3.1 derives the exact posterior and the decision rule built on it, §3.2 establishes the guarantee that posterior carries conditional on the calibration set, and §3.3 prices what deployment adds to it—an estimated ratio and a possibly impure shift.

### 3.1 THE EXACT POSTERIOR UNDER COVARIATE SHIFT

**A nonparametric posterior on $P$.** Place a $\mathrm{DP}(\tau, G_0)$ prior on $P$ and let $\tau \to 0^+$. The limit is the Bayesian bootstrap (Rubin, 1981): the posterior is supported on the empirical atoms, $P^* = \sum_i V_i\, \delta_{Z_i}$ with $Z_i = (X_i, Y_i)$ and $(V_1, \ldots, V_n) \sim \mathrm{Dir}(1, \ldots, 1)$, and nothing beyond i.i.d. sampling enters. These are the masses of Eq. (2) in a different parametrization: the atom masses of $P^*$ on the sorted scores follow the same $\mathrm{Dir}(1, \ldots, 1)$ law as the order-statistic spacings. The atom-mass parametrization is the one that generalizes—it never invokes the probability integral transform, only the prior on $P$—so covariate shift can now enter as a functional of $P$.

**Pushforward under the change of measure.** The test-score law is a deterministic functional of the calibration law, $T(Q) = w\,Q / \int w\, dQ$. The posterior on the test-score law is therefore the pushforward $T_\#$ of the Bayesian bootstrap; applied to the atoms of $P^*$,

$$\tilde{V}_i = \frac{V_i\, w_i}{\sum_j V_j\, w_j} \;\overset{d}{=}\; \frac{w_i E_i}{\sum_{j=1}^n w_j E_j}, \qquad E_1, \ldots, E_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1), \tag{3}$$

the normalized weighted Bayesian bootstrap (Newton & Raftery, 1994). No approximation enters at either step: the pushforward of an exact posterior is an exact posterior, and the exponential form is an identity rather than a convenient surrogate—substituting the Gamma representation $V_i = E_i / \sum_k E_k$ of $\mathrm{Dir}(1, \ldots, 1)$ into the tilt cancels the normalizer $\sum_k E_k$ between numerator and denominator. Each draw is therefore one candidate deployed score distribution, $\sum_i \tilde{V}_i \delta_{\rho_i}$, obtained at $O(n)$ cost—that of Dirichlet sampling.

**The law of the induced masses. $\tilde{\boldsymbol{V}}$** admits a closed-form characterization:

**Proposition 1** (Exact induced posterior; informal)**.** *The vector $\tilde{\boldsymbol{V}}$ of Eq.* (3) *follows a* scaled Dirichlet *law, with density*

$$p(\tilde{v}_1, \ldots, \tilde{v}_n) = \Gamma(n) \left(\textstyle\prod_i w_i\right)^{-1} \left(\textstyle\sum_j \tilde{v}_j / w_j\right)^{-n} \tag{4}$$

*on the simplex, which is Dirichlet only when the weights are uniform. The unique Dirichlet matching its first two moments at leading order is $\mathrm{Dir}(n_{\mathrm{eff}}\tilde{w}_1, \ldots, n_{\mathrm{eff}}\tilde{w}_n)$ with Kish's effective sample size (Kish, 1965)*

$$n_{\mathrm{eff}} = \left(\textstyle\sum_i w_i\right)^2 \Big/ \textstyle\sum_i w_i^2 : \tag{5}$$

*the effective sample size emerges from the derivation rather than being imposed. (Formal statements and proofs: Appendix A.)*

The i.i.d. theory is contained as a special case: with uniform weights the exponentials normalize to $\mathrm{Dir}(1, \ldots, 1)$ exactly and $n_{\mathrm{eff}} = n$, so the exchangeable posterior of Eq. (2)—with its data-conditional guarantee—is recovered without approximation.

**The decision rule.** For a candidate threshold $\lambda$, define the posterior loss variable

$$L_w^+(\lambda) = \sum_{i=1}^n \tilde{V}_i\, \ell(\lambda; X_i, Y_i) + \tilde{V}_{n+1} B_\ell, \qquad (\tilde{V}_1, \ldots, \tilde{V}_{n+1}) \propto (w_1 E_1, \ldots, w_n E_n, \bar{w} E_{n+1}), \tag{6}$$

which extends Eq. (3) by one atom for the unseen test point—mass $\bar{w} = \mathbb{E}_{\tilde{P}_X}[\hat{w}]$, worst-case loss $B_\ell$, the weighted form of the trailing spacing in Eq. (2)—so that $L_w^+$ equals the $L^+$ of Eq. (2) at uniform weights. The rule selects at credibility $\beta$,

$$\lambda_{\mathrm{HPD}}^{\beta} = \inf\big\{\lambda : \Pr\big(L_w^+(\lambda) \le \alpha \mid \ell_{1:n}\big) \ge \beta\big\}, \tag{7}$$

the one-sided highest-posterior-density rule, and deploys the clamp $\lambda_{\mathrm{dep}} = \max(\hat{\lambda}, \lambda_{\mathrm{HPD}}^{\beta})$, so the inherited frequentist guarantee of Eq. (1) is never weakened; the draws of $\lambda \mapsto L_w^+(\lambda)$ also yield the spread $\sigma_{\mathrm{post}}$. Taking the expected value recovers the previous methods: $\mathbb{E}[L_w^+(\lambda)] = \big(\sum_i w_i \ell_i(\lambda) + \bar{w} B_\ell\big) / \big(\sum_j w_j + \bar{w}\big) + O(1/n_{\mathrm{eff}})$, exactly at uniform weights, and its $\alpha$-crossing is the W-CRC rule—the posterior refines the expectation-based selection rather than replacing it. Algorithm 1 summarizes.

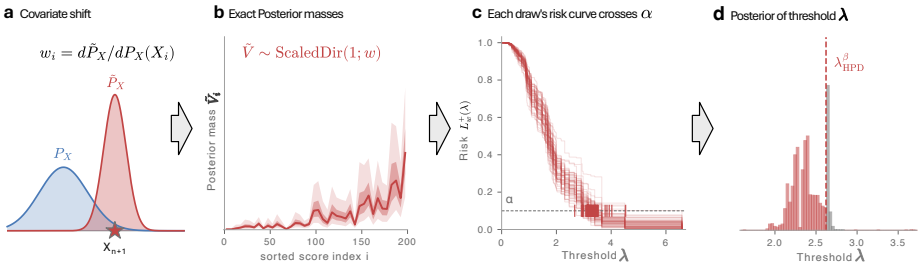



Figure 2: **The WBCP pipeline**, drawn for the miscoverage loss. (a) Covariate shift enters through the likelihood ratio $w_i$. (b) The law of the masses, Eq. (3), which depends on the weights alone: median mass at each calibration score with its central 50% and 90% credible bands. (c) Sampling masses from that law gives one risk curve $L_w^+(\lambda)$ per draw, each crossing the target $\alpha$ at that draw's threshold $\lambda^{(m)}$ (ticks on the $\alpha$ line). (d) The crossings over the $M$ draws—for this single calibration set—form the threshold posterior: a $\beta$ fraction of its mass lies at or below $\lambda^{\beta}_{\text{HPD}}$, and the gray tail $1-\beta$ is the tolerated failure credibility. The trailing test atom of Eq. (6) is omitted for clarity.

**Algorithm 1** WBCP: Weighted Bayesian Conformal Prediction

**Require:** Scores $\{\rho_i\}$, weights $\{w_i\}$, test-point mass $\bar{w}$, loss $\ell$ with bound $B_\ell$, target $\alpha$, credibility $\beta$, draws $M$

1: Sort scores; compute $\tilde{w}_i$, $n_{\text{eff}}$ (Eq. 5), and the empirical selection $\hat{\lambda}$ (Eq. 1)
2: **for** $m = 1$ **to** $M$ **do**
3: Draw $E_1, \ldots, E_{n+1} \sim \text{Exp}(1)$; set $(\tilde{V}_1^{(m)}, \ldots, \tilde{V}_{n+1}^{(m)}) \propto (w_1 E_1, \ldots, w_n E_n, \bar{w} E_{n+1})$, normalized — exact sampler, Eq. (6)
4: $L_w^{+(m)}(\lambda) = \sum_{i \le n} \tilde{V}_i^{(m)} \ell(\lambda; X_i, Y_i) + \tilde{V}_{n+1}^{(m)} B_\ell$ on the score grid; record its $\alpha$-crossing $\lambda^{(m)}$
5: **end for**
6: $\lambda^{\beta}_{\text{HPD}} =$ smallest grid point with $\#\{m : L_w^{+(m)}(\lambda) \le \alpha\}/M \ge \beta$ — Eq. (7)
7: $\lambda_{\text{dep}} = \max(\hat{\lambda}, \lambda^{\beta}_{\text{HPD}})$
8: **return** $\lambda_{\text{dep}}, \sigma_{\text{post}} = \text{std}\{\lambda^{(m)}\}, n_{\text{eff}}, \{\lambda^{(m)}\}$

### 3.2 Data-Conditional Risk Control

This subsection states what the rule guarantees *within* the posited model—conditional on the observed calibration set, with the weight function taken as given; §3.3 will price what deployment adds. The operative point here is that a credible statement cannot be an artifact of the prior, and it is made about the posterior the rule actually samples:

**Theorem 2** (Weighted stochastic dominance)**.** *With $L_w^+$ as in Eq.* (6) *on the exact posterior masses:* $\inf_\pi \Pr(L \le b \mid \ell_{1:n}) \ge \Pr(L_w^+ \le b)$ *for every $b$, the infimum over all priors on the state of nature. (The argument uses no property of the mass law beyond its living on the simplex; the Dirichlet case of Snell & Griffiths (2025) and the scaled Dirichlet of Proposition 1 are both instances.)*

**Corollary 3** (Data-conditional guarantee)**.** *For the $\beta$-credible threshold of Algorithm 1:* $\inf_\pi \Pr\big(L(\lambda_{\text{dep}}) \le \alpha \mid \ell_{1:n}\big) \ge \beta$.

A closed-form companion in Appendix D fixes the scale of the spread: $c \asymp n_{\text{eff}}$ is the *only* concentration matching the frequentist variance of the weighted empirical CDF and quantile (Theorem 16), which is what makes $\sigma_{\text{post}}$ a measurement rather than a modelling choice.

**Remark 1** (Why the i.i.d. case needed no certificate, and this one does)**.** *Under exchangeability with continuous scores,* $\text{Dir}(1, \ldots, 1)$ *is the exact frequentist law of the masses between consecutive order statistics: credibility* equals *confidence, and the i.i.d. construction obtains valid risk control from a posterior for free. Under weighting the identity breaks twice—the spacing law depends on the unknown $w^*$, and the exact posterior is provably not Dirichlet—so how credibility transfers to the deployment distribution is no longer an identity but a theorem (§3.3), priced by exactly the quantities the i.i.d. case sets to zero.*

### 3.3 THE DEPLOYMENT CERTIFICATE

The guarantees above hold under the induced posterior with the true likelihood ratio; the marginal guarantee of weighted conformal prediction is inherited unchanged, since $\lambda_{\mathrm{dep}} \geq \hat{\lambda}$. Deployment supplies an estimated ratio and possibly an impure shift. The following statement, which assumes no posterior model, prices both.

**Assumption 1** (Certificate conditions)**.** (i) Sampling as in §2. (ii) The deployed weight function $\hat{w}$ is *fixed* (estimated on data independent of the scored calibration set and of the evaluated test points) with $B := \operatorname{ess\,sup} \hat{w}/\mathbb{E}_{P_X}[\hat{w}] < \infty$. Pooled density-ratio estimation violates (ii); every experiment uses the compliant covariate-split recipe of §4.

Three error primitives, each with an operational meaning: $\Delta_w = \mathbb{E}_{P_X}|\hat{w} - w^*|$ (weight-estimation error; twice the total variation between the $\hat{w}$-tilt of the calibration covariates and the test covariate law); $\Delta_s = \frac{1}{2}\sup_\lambda \operatorname{ess\,osc}_x \mathbb{E}[\ell(\lambda; X, Y)/B_\ell \mid X = x]$ (non-pivotality: how much the conditional risk curve varies with $x$; zero for a perfectly locally adaptive score); and $\Delta_{\mathrm{shift}}$, the same oscillation between calibration and test conditional laws (zero under pure covariate shift; not estimable without test labels).

**Theorem 4** (Two-sided deployment certificate)**.** *Let Assumption 1 hold, fix $\delta \in (0,1)$ and a grid size $m$, and let $\eta_n(\delta, m)$ be the explicit concentration constant of Appendix B (Hoeffding form; rate constant $B/\sqrt{n}$, refined to $1/\sqrt{n_{\mathrm{eff}}}$ by a Bernstein variant). With probability at least $1-\delta$ over the calibration draw,*

$$\alpha - B_\ell\big(\Delta_w\Delta_s + \Delta_{\mathrm{shift}} + \eta_n + \tilde{w}_{\max}\big) - \widehat{M}_\beta \;<\; R_{\mathrm{test}}(\lambda_{\mathrm{dep}}) \;\leq\; \alpha + B_\ell\big(\Delta_w\Delta_s + \Delta_{\mathrm{shift}} + \eta_n\big), \quad (8)$$

*where $\tilde{w}_{\max} = \max_i \tilde{w}_i$ and the* clamp mass *$\widehat{M}_\beta = \widehat{R}_w(\hat{\lambda}) - \widehat{R}_w(\lambda_{\mathrm{dep}}) \in [0, \alpha]$ is computed, not estimated, at prediction time. All constants are explicit; no step is asymptotic (Appendix B).*

The structure carries the content. The leading error is a *product*: weight error hurts only to the extent that the score is non-pivotal, so either a correct weight profile ($\Delta_w = 0$) *or* a locally adaptive score ($\Delta_s = 0$) suffices under pure covariate shift—a finite-sample either-arm property in the spirit of doubly robust inference (Lei & Candès, 2021). The product term is not an artifact of the proof: a two-point construction attains it exactly, so the leading term is unimprovable in form and constant (Proposition 11, Appendix B.1). The upper (safety) side does not involve $\widehat{M}_\beta$: clamping can only lower risk, and the entire conservatism cost of the Bayesian selection is one computable number, which vanishes as $\beta \downarrow 0.5$.

**Corollary 5** (Instantiations)**.** *(i)* Coverage*: with miscoverage loss, Eq.* (8) *bounds $|\Pr(Y \in \mathcal{C}(X) \mid \mathcal{D}) - (1-\alpha)|$ by the same budget (order-statistic convention shifting $\alpha$ by $\leq 2/n$, absorbed). (ii)* The i.i.d. corner*: uniform-weight deployment gives $\Delta_w = \Delta_{\mathrm{shift}} = 0$, $B = 1$, $\tilde{w}_{\max} = 1/n$: an i.i.d. bounded-loss guarantee of the operational form of Eq.* (2) *with explicit constants—the theorem the lost identity of Remark 1 used to make unnecessary. (iii)* Abstention*: any selection rule measurable in the covariates alone (thresholds on $n_{\mathrm{eff}}$, kernel overlap) preserves the certificate verbatim on the selected subpopulation with its restricted constants; selection never reads scores or labels, so no selection-bias term arises.*

**Remark 2** (What is and is not estimable)**.** *$\tilde{w}_{\max}$ and $\widehat{M}_\beta$ are exact at prediction time. $\eta_n$'s scale is governed by $n_{\mathrm{eff}}$ (Bernstein form), and $\sigma_{\mathrm{post}}$ is its posterior-side counterpart (Theorem 16). $\Delta_s$ can be probed from calibration data; $\Delta_w$ is not identifiable, its collapse detectable one-sidedly through $n_{\mathrm{eff}}$; $\Delta_{\mathrm{shift}}$ is not estimable from unlabeled test data at all. Absence of alarm is not evidence of safety; the certificate says which term is blind, and labeled audits are the only verdict on it.*

## 4 EXPERIMENTS

All experiments compare decision rules at the same specification: deployed risk at most $\alpha$ ($0.4$ in the multilabel setting, matching BQ-CP; $0.1$ elsewhere) with failure rate at most $1 - \beta = 0.05$; sweeping $\beta$ instead, the realized failure rate tracks $1 - \beta$ across the whole dial (Appendix C.6). We report the relative frequency of trials whose realized risk exceeded $\alpha$, with Clopper–Pearson $95\%$ intervals, the mean realised risk, and the size of what is deployed—interval length in regression, prediction-set size in classification, the threshold itself for the multilabel count loss. The ideal rule keeps the

frequency below $5\%$ with the smallest sets. Four rules enter every comparison: **RCPS** (Bates et al., 2021) as published (Hoeffding–Bentkus bound, i.i.d. machinery applied to the calibration scores unchanged—hence blind to the shift), **W-CRC** (the weighted CRC rule; expectation guarantee only), **BQ-CP** (the equal-weight posterior, likewise shift-blind), and **WBCP** ($\beta = 0.95$, Algorithm 1). Likelihood ratios are estimated by a logistic discriminator fit on covariate halves disjoint from the scored calibration and evaluated test sets, so $\hat{w}$ is a fixed function in the sense of Assumption 1(ii). Where the design provides the true ratio, a reference row *WBCP (oracle $w$)* separates the certificate's terms: its distance to nominal is $\eta_n$ plus granularity, and the gap between the two WBCP rows is $\Delta_w$'s share. Full protocols and sweeps: Appendix C.

### 4.1 Synthetic Regression under Covariate Shift

On the heteroskedastic benchmark of Snell & Griffiths (2025), $X \sim U[0, 4]$, $Y \mid X \sim N(0, X^2)$, intervals $[-\lambda, \lambda]$, miscoverage loss; the test covariates tilt toward the noisy end with a closed-form oracle ratio, and weights are estimated. In the exchangeable control the posterior keeps the failure rate below nominal ($2.6\%$) with intervals $3\%$ shorter than RCPS's; under the shift (Table 1) both shift-blind rules fail in most trials while WBCP sits at the nominal rate (per-trial distributions: Figure 3, Appendix C)—on a one-dimensional covariate the weight-estimation premium is negligible.

Table 1: Synthetic heteroskedastic regression under covariate shift, estimated weights ($n = 200$, $\alpha = 0.1$, 10,000 trials).

| Decision rule | Rel. freq. | 95% CI | Mean risk | Mean length |
|---|---|---|---|---|
| BQ-CP | 92.3% | [91.7, 92.8] | 14.4% | 9.35 |
| RCPS | 85.2% | [84.5, 85.9] | 13.3% | 9.67 |
| W-CRC | 42.8% | [41.8, 43.8] | 9.5% | 10.98 |
| WBCP | **5.0**% | [4.5, 5.4] | 5.0% | 13.41 |
| WBCP (oracle $w$) | 6.7% | [6.2, 7.2] | 5.4% | 13.09 |

### 4.2 Synthetic Multilabel Classification under Covariate Shift

On the companion count-loss benchmark, each unit carries $K = 4$ binary sub-outcomes—the loss shape of multilabel false-negative control—except that the sub-outcome scores depend on a covariate, so the conditional risk curve varies across units, and the test distribution tilts toward the hard ones. Table 2 shows $n = 10$ (its original setting) and $n = 250$. The $n$-dependence is the point. At $n = 10$ with $n_{\text{eff}} \approx 7$, no threshold reaches $95\%$ credibility in $93\%$ of trials: WBCP abstains—the covariate-measurable selection of Corollary 5(iii)—and never violates when it does certify. The blind rules certify regardless: BQ-CP violates in $28\%$ of trials, while RCPS's large finite-sample slack happens to absorb the shift. But slack is not control: by $n = 100$ both blind rules violate in essentially every trial (Appendix C), while WBCP certifies in every trial at 7.5–7.9%. The oracle row attributes the residual to $\eta_n$, not weight estimation.

Table 2: Synthetic multilabel classification (count loss, $K = 4$) under covariate shift, estimated weights ($\alpha = 0.4$, 10,000 trials). At $n = 10$ WBCP finds no certifiable threshold in $93\%$ of trials (W-CRC $4\%$, others 0); frequencies are over certified trials. At $n = 250$ every rule always certifies.

| | $n = 10$ | | | $n = 250$ | | |
|---|---|---|---|---|---|---|
| Decision rule | Rel. freq. | 95% CI | Mean $\lambda$ | Rel. freq. | 95% CI | Mean $\lambda$ |
| BQ-CP | 28.2% | [27.3, 29.0] | 2.44 | 100% | [100, 100] | 1.05 |
| RCPS | 5.7% | [5.3, 6.2] | 3.49 | 100% | [100, 100] | 1.16 |
| W-CRC | 13.2% | [12.5, 13.9] | 3.31 | 43.7% | [42.8, 44.7] | 1.92 |
| WBCP | **0.0**% | [0.0, 0.5] | 7.61 | **7.9**% | [7.4, 8.5] | 2.36 |
| WBCP (oracle $w$) | 0.0% | [0.0, 0.9] | 7.94 | 5.6% | [5.1, 6.0] | 2.35 |

### 4.3 PovertyMap: Natural Covariate Shift

PovertyMap-WILDS (Koh et al., 2021; Yeh et al., 2020) regresses an asset-wealth index from satellite imagery; the shift is the one that exists between countries, and the ratio is estimated from embeddings. Nine held-out countries contribute re-splits that pass the overlap screen (Table 3). At the effective

sample sizes real shifts produce ($n_{\text{eff}} \approx 120$–$600$), WBCP's failure rate falls below the nominal $5\%$; every blind or expectation-only rule violates at four to six times nominal.

Table 3: PovertyMap, held-out country, estimated weights (747 screened re-splits across 9 countries).

| Decision rule | Rel. freq. | 95% CI | Mean risk | Mean length |
|---|---|---|---|---|
| BQ-CP | 29.0% | [25.8, 32.5] | 8.8% | 1.53 |
| RCPS | 26.4% | [23.2, 29.7] | 8.5% | 1.54 |
| W-CRC | 18.7% | [16.0, 21.7] | 8.3% | 1.51 |
| WBCP | **1.6%** | [0.9, 2.9] | 5.9% | 1.72 |

### 4.4 Fashion-MNIST: Constructed Covariate Shift

On Fashion-MNIST with APS scores from a multinomial-logistic probe (held-out accuracy $0.84$), resampling the test pool along the softmax-entropy direction gives pure covariate shift with a closed-form ratio; the loss is miscoverage of the prediction set (Table 4). The density-ratio discriminator sees the softmax vector together with the uncertainty statistics a deployed system logs (entropy, top probability, margin); with this recipe the deployed arm matches the oracle arm at near-nominal rates. Restricting the discriminator to the raw softmax vector mis-specifies the tilt: the estimated weights flatten toward uniform rather than blowing up—$n_{\text{eff}}$ stays near $0.8n$ and no overlap diagnostic fires—and the deployed arm inflates to $28.0\%$ with the oracle arm unchanged. This is the $\Delta_w$ premium of Theorem 4 in isolation, realized in its undetectable direction (Remark 2). Tilting in the opposite direction turns the blind rules' failure from invalid into over-conservative (Appendix C).

Table 4: Fashion-MNIST, entropy tilt toward hard inputs, estimated weights (2 tilt strengths $\times$ 200 splits). Mean size is the mean number of classes in the deployed prediction set (of 10).

| Decision rule | Rel. freq. | 95% CI | Mean risk | Mean size |
|---|---|---|---|---|
| BQ-CP | 100% | [99.1, 100] | 15.2% | 2.66 |
| RCPS | 100% | [99.1, 100] | 14.8% | 2.70 |
| W-CRC | 45.0% | [40.1, 50.0] | 9.9% | 3.25 |
| WBCP | **6.5%** | [4.3, 9.4] | 8.7% | 3.40 |
| WBCP (oracle $w$) | 7.0% | [4.7, 10.0] | 8.9% | 3.37 |

## 5 Discussion

The experiments separate three kinds of answer. The expectation guarantee (W-CRC) holds as stated—its mean risk sits at the target—while individual calibration sets violate the target roughly half the time: what an expectation guarantee promises, and no more. The shift-blind rules fail together, and the multilabel experiment shows why apparent exceptions mislead: finite-sample slack can absorb the shift at small $n$, and slack shrinks with $n$ while bias does not (§4). WBCP is the only rule that certifies at the nominal rate—$5.0\%$ synthetic regression, $1.6\%$ PovertyMap, $6.5\%$ Fashion-MNIST—and where the calibration set cannot support certification ($n_{\text{eff}} \approx 7$), it abstains. When the discriminator is mis-specified for the shift the $\Delta_w$ term is visible ($28.0\%$ raw-softmax against $6.5\%$ enriched on Fashion-MNIST), and the oracle–split contrast attributes it entirely to weight estimation. The oracle rows close the accounting: every deviation from the nominal $5\%$ is either oracle-to-nominal distance ($\eta_n$ and granularity, shrinking as the posterior concentrates) or the gap between the WBCP rows ($\Delta_w$)—nothing else. The same accounting covers width: under the tilted law no valid interval can be shorter than $2\lambda^*(\alpha) = 10.6$ in the regression experiment. The blind rules' $9.4$–$9.7$ sit *below* that length—their violation rates are the price—while WBCP's $13.4$ is a $26\%$ finite-sample premium over it (the full $\gamma$-sweep: Figure 4, Appendix C).

**Why $\beta$-credibility is not a frequentist guarantee under estimated weights.** Deployment involves two sources of error, and the posterior models only one. It captures the randomness of the calibration draw—which $n$ points happened to be observed—while taking the estimated weight function as given. When $\hat{w}$ is close to the truth, that is the only error that matters, and the credible level transfers ($6.5\%$ against nominal $5\%$ on Fashion-MNIST). When $\hat{w}$ is wrong, the posterior is a faithful posterior for the wrong population, and the transfer degrades by exactly the certificate's $\Delta_w\Delta_s$—the $28.0\%$ raw-softmax arm. A first-order completion is a sensitivity band: the worst-case HPD threshold over $w/\hat{w} \in [e^{-\varepsilon}, e^{\varepsilon}]$ costs one reweighting pass (Proposition 12), and on the mis-specified arm

above, $\varepsilon = 0.1$ already restores the nominal rate at a 7% set-size premium (Appendix B.2). The full completion—a joint posterior over the calibration law and the ratio—remains future work.

**What the diagnostics can and cannot tell you.** $\sigma_{\text{post}}$ answers one question: how much the selected threshold would move if the calibration set were redrawn (Theorem 16 matches its moments to this variance). It does not say *which* test points will miscover. And it cannot see conditional shift: if $P_{Y|X}$ itself differs between calibration and deployment, every statistic computed from unlabeled test data—$n_{\text{eff}}$, discriminator separability, $\sigma_{\text{post}}$ included—can look normal while the guarantee fails (Remark 2). The only way to check $\Delta_{\text{shift}}$ is to collect labels and audit.

**Limitations.** Four assumptions bound the scope. First, the weight function must be fixed: estimated on data disjoint from the scored calibration points and the evaluated test points (the split of §4). Second, the certificate can be stated but not fully evaluated in the field: $\Delta_w$ cannot be estimated without the true ratio, and $\Delta_{\text{shift}}$ requires labels—the two blind spots discussed above. Third, the loss must be bounded and monotone in the threshold, and heavily tied scores—discrete APS values—inflate the granularity term $\tilde{w}_{\max}$. Fourth, the sampling model of §2 takes the calibration points to be i.i.d.; with autocorrelated data—spatial or temporal—neighbors repeat information, $n_{\text{eff}}$ overstates the effective calibration sample, and our guarantees do not cover that case.

## 6 Related Work

**Conformal prediction.** CP (Vovk et al., 2005) and its split variant (Papadopoulos et al., 2002; Lei et al., 2018) provide distribution-free coverage.

**Weighted and localized CP.** Tibshirani et al. (2019) introduced weighted exchangeability, the foundation WBCP builds upon. Barber et al. (2023) bound the coverage gap under arbitrary non-exchangeability by a weighted sum of total-variation distances—population quantities, one-sided, fixed weights; our certificate differs in being two-sided, product-form, and built from terms with prediction-time estimators; localized variants (Guan, 2023; Lou et al., 2025) remain frequentist point thresholds.

**Risk control.** RCPS (Bates et al., 2021) and quantile risk control (Snell et al., 2023) certify a target risk at a target failure rate, both under i.i.d. sampling. CRC (Angelopoulos et al., 2024) controls expected risk for monotone losses, its covariate-shift extension (Prop. 2) assuming known ratios and leaving the estimated case open (§1); Zecchin et al. (2025) analyze W-CRC's expected set size, also under known ratios—complementary to our risk-deviation certificate. Concurrently, Opoku & Banahene (2026) control risk up to a calibration–deployment mismatch term monitored empirically at deployment, where Theorem 4 derives the bound; selective risk control (Xu et al., 2025) develops the selection dimension Corollary 5(iii) covers. To our knowledge no prior work gives a high-probability, two-sided weighted risk bound that prices estimated weights and conditional shift.

**Doubly robust conformal.** Lei & Candès (2021) and Yang et al. (2024) obtain asymptotic either-nuisance validity for weighted split-CQR and for prediction sets under explainable covariate shift; our product term $\Delta_w\Delta_s$ is a finite-sample form of the same either-arm structure, for arbitrary scores. Qiu et al. (2023) prove that finite-sample-valid sets under *unknown* shift are necessarily uninformative—hence a priced certificate plus covariate-measurable abstention rather than a stronger guarantee.

**Bayesian approaches to CP.** Fong & Holmes (2021) attach Bayes at the predictive-score level; Snell & Griffiths (2025) establish BQ-CP under i.i.d., where credibility and confidence coincide by the spacing identity (Remark 1); we derive the posterior that survives weighting.

## 7 Conclusion

Weighted conformal risk control asks what can be certified when deployment shifts away from calibration and the shift must be estimated. We gave both halves of the answer—the exact posterior over the deployed risk curve, and the bounds that price what deployment adds—and stated the blind spot as precisely: what no unlabeled diagnostic can see, a labeled audit must.

## Reproducibility Statement

All experiments run from fixed seeds on one CPU; no result depends on GPU nondeterminism. The synthetic benchmarks (§4) specify their generative models in closed form in Appendix C, which also states the full protocols for the real-data experiments: data sources, frozen scoring models, the A2-compliant weight-estimation split, the overlap screen, and every sweep behind the reported tables. Complete proofs of all claims are in Appendices A–D. The code implementing Algorithm 1 and reproducing every table and figure is included in the supplementary material and will be released publicly upon publication.

## References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Edwin Fong and Chris C. Holmes. Conformal bayesian computation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023. doi: 10.1093/biomet/asac040.

Ying Jin, Zhimei Ren, and Emmanuel J. Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6): e2214889120, 2023.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021.

Hira L. Koul. *Weighted Empiricals and Linear Models*, volume 21 of *Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, 1992.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

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

Xiayin Lou, Peng Luo, and Liqiu Meng. Geoconformal prediction: A model-agnostic framework for measuring the uncertainty of spatial prediction. *Annals of the American Association of Geographers*, 2025. doi: 10.1080/24694452.2025.2516091.

Michael A. Newton and Adrian E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B*, 56(1):3–26, 1994.

Jeffery Opoku and David Banahene. Promptshift-crc: Drift-aware conformal risk control for foundation models under prompt and domain shift. *arXiv preprint arXiv:2606.15964*, 2026.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

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

Donald B. Rubin. The bayesian bootstrap. *Annals of Statistics*, 9(1):130–134, 1981.

Jake C. Snell and Thomas L. Griffiths. Conformal prediction as bayesian quadrature. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. Outstanding Paper Award.

Jake C Snell, Thomas P Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. In *International Conference on Learning Representations*, 2023.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

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

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 2024.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):2583, 2020.

Matteo Zecchin, Fredrik Hellström, Sangwoo Park, Shlomo Shamai, and Osvaldo Simeone. Generalization and informativeness of weighted conformal risk control under covariate shift. *arXiv preprint arXiv:2501.11413*, 2025.

## DISCLOSURE OF AI ASSISTANCE

Large language models (Claude, Anthropic) were used in the preparation of this paper, in the following capacities. *Writing:* drafting and polishing prose throughout the paper under the authors' direction, including the statements of theorems and the exposition of proofs; all mathematical claims and derivations were checked by the authors. *Research execution:* implementing the experimental pipeline (samplers, decision rules, evaluation scripts) and generating figure templates; the experimental design was directed by the authors, and all reported numbers were checked by the authors against the raw experimental outputs. *Retrieval and discovery:* literature search and analysis complementing references identified by the authors, and verification of the bibliographic metadata of every cited reference. The authors take full responsibility for the entire content of the paper.

# A Derivation of the Weighted Posterior

This appendix states the formal versions of Proposition 1; the complete moment analysis with explicit remainder bounds is given in Appendix D.

**Failure of the PIT route.** In the i.i.d. case, exchangeability implies via the probability integral transform (PIT) that $U_i = F(\rho_i)$ are i.i.d. $\mathrm{Uniform}[0,1]$, whose order-statistic spacings are *exactly* $\mathrm{Dir}(1,\dots,1)$ (BQ-CP Lemma 4.2). Under weighted exchangeability, the joint density $\prod_i w_i(v_i) \cdot g(v_1,\dots,v_n)$ contains the unknown permutation-invariant factor $g$; the joint law of the PIT ranks is non-uniform and depends on $g$, so no fixed spacing distribution exists. The derivation below circumvents this: under covariate shift the *calibration* sample is still i.i.d. (only the test point is shifted), and the shift acts through a *known* likelihood ratio.

**Setup.** Calibration data $Z_i = (X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P$; test point from $\tilde{P} = \tilde{P}_X \times P_{Y|X}$ with known $w(x) = d\tilde{P}_X/dP_X(x) > 0$; $w_i = w(X_i)$. The estimand is the test-score distribution $\tilde{F}(t) = \mathbb{E}_P[w(X)\mathbf{1}\{\rho \le t\}]/\mathbb{E}_P[w(X)]$, a deterministic functional of $P$; the population threshold is $\lambda^* = \tilde{F}^{-1}(1-\alpha)$.

**Theorem 6** (Exact Induced Posterior)**.** *Place a $\mathrm{DP}(\tau, G_0)$ prior on the law of $(X,\rho)$ and let $\tau \to 0^+$, giving the Bayesian bootstrap posterior $Q^* = \sum_i V_i \delta_{Z_i}$, $\boldsymbol{V} \sim \mathrm{Dir}(1,\dots,1)$ (Rubin, 1981). Pushing $Q^*$ forward through the tilt $Q \mapsto w\,Q/\int w\,dQ$ induces the posterior $\tilde{F}^* = \sum_i \tilde{V}_i \delta_{\rho_i}$ with*

$$\tilde{V}_i = \frac{V_i w_i}{\sum_j V_j w_j} \overset{d}{=} \frac{w_i E_i}{\sum_j w_j E_j}, \qquad E_j \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1),$$

*the normalized weighted Bayesian bootstrap (Newton & Raftery, 1994). $\tilde{\boldsymbol{V}}$ follows the scaled Dirichlet law with density $\Gamma(n)\,(\prod_i w_i)^{-1}(\sum_j \tilde{v}_j/w_j)^{-n}$ on the simplex, and is* not *Dirichlet unless all $w_i$ are equal.*

*Proof.* **Step 1 (the tilt applied to $Q^*$).** $Q^*$ is a.s. supported on $\{Z_1,\dots,Z_n\}$ with $V_i > 0$ a.s., and $w_i = w(X_i) > 0$, so $\int w\,dQ^* = \sum_j V_j w_j \in (0,\infty)$ a.s. and the tilt is well defined. Since $w\,Q^* = \sum_i V_i w_i \delta_{Z_i}$,

$$\tilde{F}^* \;=\; \frac{w\,Q^*}{\int w\,dQ^*} \;=\; \sum_i \frac{V_i w_i}{\sum_j V_j w_j}\,\delta_{Z_i},$$

and pushing forward to the score coordinate gives $\tilde{F}^* = \sum_i \tilde{V}_i \delta_{\rho_i}$ with $\tilde{V}_i = V_i w_i / \sum_j V_j w_j$, which is the first identity.

**Step 2 (exponential representation).** Let $E_1,\dots,E_n$ be i.i.d. $\mathrm{Exp}(1)$ and put $V_i = E_i/\sum_k E_k$; then $\boldsymbol{V} \sim \mathrm{Dir}(1,\dots,1)$. Substituting and cancelling the common factor $\sum_k E_k$ from numerator and denominator,

$$\frac{V_i w_i}{\sum_j V_j w_j} = \frac{w_i E_i/\sum_k E_k}{\sum_j w_j E_j/\sum_k E_k} = \frac{w_i E_i}{\sum_j w_j E_j},$$

which gives the stated equality in distribution.

**Step 3 (density on the simplex).** Put $G_i = w_i E_i$, so the $G_i$ are independent with $G_i \sim \mathrm{Exp}(\text{scale } w_i)$, i.e. density $w_i^{-1} e^{-g_i/w_i}$ on $(0,\infty)$, and joint density $(\prod_i w_i)^{-1}\exp(-\sum_i g_i/w_i)$. Let $T = \sum_i G_i$ and $\tilde{V}_i = G_i/T$. The map $(g_1,\dots,g_n) \mapsto (\tilde{v}_1,\dots,\tilde{v}_{n-1},T)$ with $g_i = \tilde{v}_i T$ $(i<n)$ and $g_n = (1-\sum_{i<n}\tilde{v}_i)T$ is a bijection from $(0,\infty)^n$ onto $\Delta^\circ \times (0,\infty)$ with Jacobian determinant $T^{n-1}$. Hence the joint density of $(\tilde{v}_{1:n-1}, T)$ is $(\prod_i w_i)^{-1} T^{n-1}\exp(-T\sum_j \tilde{v}_j/w_j)$. Writing $s = \sum_j \tilde{v}_j/w_j > 0$ and integrating out $T$ using $\int_0^\infty T^{n-1}e^{-Ts}\,dT = \Gamma(n)s^{-n}$ gives the marginal density of $\tilde{\boldsymbol{V}}$ on $\Delta^\circ$,

$$f(\tilde{\boldsymbol{v}}) \;=\; \Gamma(n)\,\Big(\prod_i w_i\Big)^{-1}\Big(\sum_j \tilde{v}_j/w_j\Big)^{-n},$$

as claimed. (This is the scaled Dirichlet law; it depends on $\boldsymbol{w}$ only through ratios, consistent with the scale invariance of Step 2.)

**Step 4 ($\tilde{\boldsymbol{V}}$ is not Dirichlet unless the weights are equal).** On $\Delta^\circ$ the quantity $s = \sum_j \tilde{v}_j / w_j$ is a convex combination of the values $1/w_1, \dots, 1/w_n$, so $\min_j w_j^{-1} \le s \le \max_j w_j^{-1}$ and therefore

$$0 \;<\; \Gamma(n)\Big(\prod_i w_i\Big)^{-1}\big(\max_j w_j^{-1}\big)^{-n} \;\le\; f \;\le\; \Gamma(n)\Big(\prod_i w_i\Big)^{-1}\big(\min_j w_j^{-1}\big)^{-n} \;<\; \infty$$

uniformly on $\Delta^\circ$: $f$ is bounded above and below by positive constants. Suppose $\tilde{\boldsymbol{V}} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, with density proportional to $\prod_i \tilde{v}_i^{\alpha_i - 1}$. If $\alpha_i > 1$ for some $i$, the density tends to 0 as $\tilde{v}_i \downarrow 0$ along $\Delta^\circ$, contradicting the positive lower bound; if $\alpha_i < 1$, it tends to $+\infty$, contradicting the upper bound. Hence $\alpha_i = 1$ for every $i$, so the density is constant on $\Delta^\circ$. But $f$ is constant on $\Delta^\circ$ if and only if $\tilde{\boldsymbol{v}} \mapsto \sum_j \tilde{v}_j / w_j$ is constant there, which holds if and only if $w_1 = \cdots = w_n$. Therefore $\tilde{\boldsymbol{V}}$ is not Dirichlet whenever the $w_i$ are not all equal; when they are, Step 2 gives $\tilde{\boldsymbol{V}} = \boldsymbol{V} \sim \mathrm{Dir}(1, \dots, 1)$ and BQ-CP is recovered exactly. □

**Theorem 7** (Moment Matching: Uniqueness and Impossibility). *For $A \subseteq [n]$ let $\tilde{S}_A = \sum_{i \in A} \tilde{V}_i$, $p_A = \sum_{i \in A} \tilde{w}_i$, and $q_A = n_{\text{eff}} \sum_{i \in A} \tilde{w}_i^2$. Suppose $\max_i \tilde{w}_i \le \kappa / n_{\text{eff}}$ and $n_{\text{eff}} \ge 18\kappa$. Then*

$$\mathbb{E}[\tilde{S}_A] = p_A + \frac{p_A - q_A}{n_{\text{eff}}} + O\big(\kappa^2 n_{\text{eff}}^{-3/2}\big), \qquad \mathrm{Var}(\tilde{S}_A) = \frac{q_A(1-p_A)^2 + (1-q_A)p_A^2}{n_{\text{eff}}} + O\big(\kappa^2 n_{\text{eff}}^{-3/2}\big),$$

*and the identity $q_A(1-p_A)^2 + (1-q_A)p_A^2 - p_A(1-p_A) = (q_A - p_A)(1-2p_A)$ yields:*

1. ***(Uniqueness)** if $|q_A - p_A| \le \varepsilon$ near the quantile boundary, the Dirichlet $\mathrm{Dir}(c\tilde{w}_1, \dots, c\tilde{w}_n)$ matches both moments at leading order iff $c + 1 = n_{\text{eff}}(1 + O(\varepsilon + \kappa^2 n_{\text{eff}}^{-1/2}))$;*
2. ***(Impossibility)** if two sets share a value of $p_A$ but differ in $q_A$—equivalently, if $A \mapsto (q_A - p_A)(1-2p_A)/(p_A(1-p_A))$ is not constant over sets with $p_A \notin \{0, 1/2, 1\}$—then no Dirichlet with mean $\tilde{\mathbf{w}}$ matches all cumulative variances at leading order: Dirichlet variances depend on $A$ only through $p_A$, the exact ones also through $q_A$.*

The impossibility clause is why Algorithm 1 defaults to the exact sampler: it removes the approximation entirely at unchanged cost. The homogeneity quantity $|q_A - p_A|$ is small when weight magnitudes are not systematically aligned with the score ordering near the quantile boundary; when they are (e.g., noise amplitude correlated with distance to the test point), the Dirichlet surrogate degrades while the exact sampler is unaffected.

## B The Deployment Certificate: Statements and Proofs

This appendix proves Theorem 4 and Corollary 5 with all constants explicit and no asymptotic step. Throughout, calibration pairs $(X_i, Y_i)$, $i = 1, \dots, n$, are i.i.d. from $P = P_X \cdot P_{Y|X}$; the test pair is independent from $\tilde{P} = \tilde{P}_X \cdot \tilde{P}_{Y|X}$ with covariate likelihood ratio $w^*(x) = d\tilde{P}_X / dP_X(x)$ (under pure covariate shift $\tilde{P}_{Y|X} = P_{Y|X}$, but the results below do not assume this—the deviation is priced). The loss $\ell(\lambda; x, y) \in [0, B_\ell]$ is nonincreasing and right-continuous in $\lambda$. The deployed weight function $\hat{w}$ is fixed (Assumption 1); since every selection rule below is invariant to rescaling $\hat{w}$, normalise $\mathbb{E}_{\text{cal}}[\hat{w}(X)] = 1$, so $\hat{w} \le B$ a.s. and $\mathbb{E}_{\text{cal}}[w^*(X)] = 1$ automatically.

**Objects.** Normalised conditional risks $r(\lambda \mid x) = \mathbb{E}_P[\ell(\lambda; X, Y) \mid X = x]/B_\ell$ and $r^{\text{te}}(\lambda \mid x)$ (under $\tilde{P}_{Y|X}$), both nonincreasing, right-continuous, $[0,1]$-valued; population risks $R_v(\lambda) = B_\ell\, \mathbb{E}_{\text{cal}}[v(X)\, r(\lambda \mid X)]$ for $v \ge 0$ with $\mathbb{E}_{\text{cal}} v = 1$, and $R_{\text{test}}(\lambda) = B_\ell\, \mathbb{E}_{X^* \sim \tilde{P}_X}[r^{\text{te}}(\lambda \mid X^*)]$; the self-normalised empirical risk $\widehat{R}_w(\lambda) = \sum_i \hat{w}(X_i)\ell(\lambda; X_i, Y_i) / \sum_j \hat{w}(X_j)$, and the selection $\hat{\lambda} = \inf\{\lambda : \widehat{R}_w(\lambda) \le \alpha\}$ (finite whenever $\inf_\lambda \widehat{R}_w \le \alpha$; for miscoverage, always). Error primitives as in §3.3: $\Delta_w = \mathbb{E}_{\text{cal}}|\hat{w} - w^*|$, $\Delta_s = \frac{1}{2}\sup_\lambda \operatorname{ess\,osc}_x r(\lambda \mid x)$, and $\Delta_{\text{shift}} = \sup_\lambda |\mathbb{E}_{X^*}[r^{\text{te}}(\lambda \mid X^*) - r(\lambda \mid X^*)]|$.

**Lemma 8** (Deterministic decomposition with product structure). *For every $\lambda$, $\big|R_{\text{test}}(\lambda) - R_{\hat{w}}(\lambda)\big| \le B_\ell\,(\Delta_{\text{shift}} + \Delta_w \Delta_s)$.*

*Proof.* Write $R_{\text{test}} - R_{\hat{w}} = (R_{\text{test}} - R_{w^*}) + (R_{w^*} - R_{\hat{w}})$. By change of measure, $R_{w^*}(\lambda) = B_\ell\, \mathbb{E}_{\text{cal}}[w^*(X) r(\lambda \mid X)] = B_\ell\, \mathbb{E}_{X^*}[r(\lambda \mid X^*)]$, so the first difference equals $B_\ell\, \mathbb{E}_{X^*}[r^{\text{te}}(\lambda \mid$

$X^*) - r(\lambda \mid X^*)]$, of absolute value at most $B_\ell \Delta_{\text{shift}}$. For the second, fix $\lambda$ and let $c_\lambda$ be the midpoint of the essential range of $r(\lambda \mid \cdot)$, so $|r(\lambda \mid x) - c_\lambda| \le \Delta_s$ a.e. Because $\mathbb{E}_{\text{cal}}[w^* - \hat{w}] = 0$, the constant can be inserted:

$$R_{w^*}(\lambda) - R_{\hat{w}}(\lambda) = B_\ell \, \mathbb{E}_{\text{cal}}\big[(w^* - \hat{w})(X)\,\big(r(\lambda \mid X) - c_\lambda\big)\big],$$

whence $|R_{w^*} - R_{\hat{w}}| \le B_\ell \, \mathbb{E}_{\text{cal}}|w^* - \hat{w}| \cdot \Delta_s = B_\ell \Delta_w \Delta_s$, uniformly in $\lambda$. □

**Lemma 9** (Uniform concentration of the self-normalised weighted risk)**.** *Fix $\delta \in (0,1)$ and an integer $m \ge 1$, and set $\varepsilon = B\sqrt{\log((4m+6)/\delta)/(2n)}$. If $\varepsilon < 1$, then with probability at least $1-\delta$,*

$$\sup_\lambda \big|\widehat{R}_w(\lambda) - R_{\hat{w}}(\lambda)\big| \;\le\; B_\ell \, \eta_n(\delta, m), \qquad \eta_n(\delta, m) := \frac{2\varepsilon + 1/m}{1-\varepsilon}.$$

*Proof.* Work with $G_\lambda = n^{-1}\sum_i \hat{w}(X_i)\,\ell(\lambda; X_i, Y_i)/B_\ell$ and $g_\lambda = \mathbb{E}[\hat{w}(X)\,\ell(\lambda; X, Y)]/B_\ell = R_{\hat{w}}(\lambda)/B_\ell$, both nonincreasing and right-continuous with $g_\lambda \in [0,1]$, and $\overline{W} = n^{-1}\sum_i \hat{w}(X_i)$ with $\mathbb{E}\overline{W} = 1$; note $\widehat{R}_w/B_\ell = G_\lambda/\overline{W}$.

*Grid.* Let $\lambda_j = \inf\{\lambda : g_\lambda \le j/m\}$, $j = 1, \dots, m-1$; by right-continuity $g_{\lambda_j} \le j/m$ and $g_{\lambda_j^-} \ge j/m$. The evaluation set consists of each $\lambda_j$, each left limit $\lambda_j^-$, and the two endpoints: at most $2m$ points.

*Pointwise concentration.* Each summand of $G_\lambda$ (and of each left-limit functional, again a sum of i.i.d. variables) lies in $[0, B]$, so Hoeffding gives $P(|G_u - g_u| \ge \varepsilon) \le 2e^{-2n\varepsilon^2/B^2}$ at each evaluation, and likewise for $\overline{W}$. A union bound over at most $2m+3$ events costs at most $(4m+6)\,e^{-2n\varepsilon^2/B^2} \le \delta$ with the stated $\varepsilon$. Call the good event $\mathcal{E}$.

*Grid to all $\lambda$.* For $\lambda \in [\lambda_{j+1}, \lambda_j)$ (both functions nonincreasing; endpoint cells analogous),

$$G_\lambda - g_\lambda \le G_{\lambda_{j+1}} - g_{\lambda_j^-} \le (G_{\lambda_{j+1}} - g_{\lambda_{j+1}}) + (g_{\lambda_{j+1}} - g_{\lambda_j^-}) \le \varepsilon + \tfrac{1}{m},$$

since $g_{\lambda_{j+1}} \le (j+1)/m$ and $g_{\lambda_j^-} \ge j/m$. The symmetric argument bounds $g_\lambda - G_\lambda$, so $\sup_\lambda |G - g| \le \varepsilon + 1/m$ on $\mathcal{E}$.

*Self-normalisation.* On $\mathcal{E}$, $\overline{W} \ge 1 - \varepsilon > 0$, and with $g_\lambda \le 1$,

$$\Big|\frac{G_\lambda}{\overline{W}} - g_\lambda\Big| \le \frac{|G_\lambda - g_\lambda| + g_\lambda|\overline{W} - 1|}{\overline{W}} \le \frac{(\varepsilon + 1/m) + \varepsilon}{1-\varepsilon}.$$ □

**Remark 3** (Recommended grid; the $n_{\text{eff}}$ rate)**.** *With $m = \lceil\sqrt{n}\rceil$ the rate constant is $B/\sqrt{n}$. Replacing Hoeffding by Bernstein (each summand has variance at most $v = \mathbb{E}_{\text{cal}}[\hat{w}^2]$) gives, at each fixed evaluation, $\varepsilon \le \sqrt{2\log(2/\delta')/n_{\text{eff}}^{\text{pop}}} + \frac{2B}{3n}\log(2/\delta')$ with $n_{\text{eff}}^{\text{pop}} = n/v$; the Kish $n_{\text{eff}}$ of Eq.* (5) *is its plug-in. The concentration scale is thus governed by a weight-*mass *functional with no pointwise weight floor—the reformulation anticipated in Limitation (7). Constants are not optimised.*

**Lemma 10** (Granularity at the selection)**.** *If $\hat{\lambda} < \infty$ then, deterministically, $\alpha - B_\ell\,\tilde{w}_{\max} < \widehat{R}_w(\hat{\lambda}) \le \alpha$ whenever no two calibration points drop their loss discontinuously at the same $\lambda$ (a.s. true for miscoverage with atomless scores); in general the lower bound holds with $B_\ell \tilde{w}_{\max}$ replaced by the realised drop $\widehat{R}_w(\hat{\lambda}^-) - \widehat{R}_w(\hat{\lambda})$, which is computable.*

*Proof.* $\widehat{R}_w(\hat{\lambda}) \le \alpha$ by right-continuity of the nonincreasing step function $\widehat{R}_w$ and the definition of the infimum. For $\lambda < \hat{\lambda}$, $\widehat{R}_w(\lambda) > \alpha$; the drop at $\hat{\lambda}$ is a $\tilde{w}$-weighted sum of per-point loss decrements, each of magnitude at most $B_\ell$ and each carrying normalised weight at most $\tilde{w}_{\max}$, so under the stated genericity it is at most $B_\ell \tilde{w}_{\max}$, giving $\widehat{R}_w(\hat{\lambda}) \ge \widehat{R}_w(\hat{\lambda}^-) - B_\ell \tilde{w}_{\max} > \alpha - B_\ell \tilde{w}_{\max}$. □

*Proof of Theorem 4.* First the empirical selection $\hat{\lambda}$. Conditional on the calibration set, $R_{\text{test}}(\hat{\lambda})$ telescopes as

$$R_{\text{test}}(\hat{\lambda}) = \underbrace{[R_{\text{test}} - R_{\hat{w}}](\hat{\lambda})}_{\text{Lemma 8}} + \underbrace{[R_{\hat{w}} - \widehat{R}_w](\hat{\lambda})}_{\text{Lemma 9}} + \underbrace{\widehat{R}_w(\hat{\lambda})}_{\text{Lemma 10}}.$$

Lemmas 8 and 9 hold *uniformly* in $\lambda$, so their bounds apply at the random $\hat{\lambda}$; on the event of Lemma 9 the three brackets lie in $[-B_\ell(\Delta_w\Delta_s+\Delta_{\text{shift}}), B_\ell(\Delta_w\Delta_s+\Delta_{\text{shift}})]$, $[-B_\ell\eta_n, B_\ell\eta_n]$, and $(\alpha - B_\ell\tilde{w}_{\max}, \alpha]$ respectively; summing endpoints gives the two-sided bound for $\hat{\lambda}$. For the deployed clamp $\lambda_{\text{dep}} = \max(\hat{\lambda}, \lambda^{\beta}_{\text{HPD}})$: the upper (safety) side transfers because $\ell$ is nonincreasing, so $R_{\text{test}}(\lambda_{\text{dep}}) \le R_{\text{test}}(\hat{\lambda})$; for the lower side, telescope at $\lambda_{\text{dep}}$ and use $\widehat{R}_w(\lambda_{\text{dep}}) = \widehat{R}_w(\hat{\lambda}) - \widehat{M}_\beta > \alpha - B_\ell\tilde{w}_{\max} - \widehat{M}_\beta$ by Lemma 10 and the definition of the clamp mass. Both statements hold on the single event of Lemma 9, of probability at least $1-\delta$. □

*Proof of Corollary 5.* (i) With $\ell = \mathbf{1}\{y \notin \mathcal{C}_\lambda(x)\}$, $B_\ell = 1$, $R_{\text{test}}(\lambda) = 1 - \Pr(Y^* \in \mathcal{C}_\lambda(X^*))$ and the selection $\hat{\lambda}$ is the paper's weighted quantile up to the order-statistic convention, whose target differs from $\alpha$ by at most $2/n$ (the $\lceil (1-\alpha)(n+1)\rceil/n$ correction), absorbed into the displayed constants. (ii) With $w^* \equiv 1$ and $\hat{w} \equiv 1$: $\Delta_w = 0$, $\Delta_{\text{shift}} = 0$ (i.i.d. deployment), $B = 1$, $\tilde{w}_{\max} = 1/n$, and Eq. (8) reads $R_{\text{test}}(\lambda_{\text{dep}}) \le \alpha + B_\ell\eta_n$ with failure rate $\delta$—an i.i.d. bounded-loss guarantee of BQ-CP's operational form with explicit constants. We do not claim it reproduces BQ-CP's theorem, which is a posterior-credibility statement; this is its frequentist envelope, the posterior selection's conservatism being priced by $\widehat{M}_\beta$. (iii) Conditioning the test covariate on a $\sigma(X^*, X_{1:n})$-measurable event $A$ with $\Pr(X^* \in A) > 0$ replaces $\tilde{P}_X$ by $\tilde{P}_{X|A}$ (still $\ll P_X$) and leaves $\tilde{P}_{Y|X}$ untouched; every step above used the test law only through $R_{\text{test}}$ and $w^*$, so the bound holds verbatim with the $A$-restricted primitives. If $\hat{w}$ was designed for the unrestricted $w^*$, the restricted $\Delta_w^A$ generally differs from $\Delta_w$: abstention *re-targets* the guarantee rather than preserving its constants. □

### B.1 Tightness of the Product Term

The leading term of Theorem 4 is attained with constant one: no bound of the same form can improve on the product.

**Proposition 11** (The product term is attained)**.** *For every $\delta_w \in (0,1]$, $\delta_s \in (0, \frac{1}{2}]$, and margin $\epsilon > 0$ there is an instance satisfying Assumption 1 with pure covariate shift—indeed with $\tilde{P} = P$, so $\Delta_{\text{shift}} = 0$ and $w^* \equiv 1$—and a fixed weight function $\hat{w}$ with $\Delta_w = \delta_w$ and $\Delta_s = \delta_s$, together with a target $\alpha$, such that the empirical selection satisfies*

$$R_{\text{test}}(\hat{\lambda}) \;\longrightarrow\; \alpha + B_\ell\,\delta_w\delta_s - \epsilon \qquad \textit{almost surely as } n \to \infty.$$

*The deviation from target thus comes within any prescribed margin of the budget's leading term, which is therefore not improvable in form or constant.*

*Proof.* Two covariate values, $P_X\{a\} = P_X\{b\} = \frac{1}{2}$; deployment identical to calibration ($w^* \equiv 1$). Miscoverage loss with three score atoms $s_{\text{lo}} < s_{\text{mid}} < s_{\text{hi}}$: at $x = a$ the score equals $s_{\text{hi}}$ with probability $p_a$ and $s_{\text{lo}}$ otherwise; at $x = b$, $s_{\text{hi}}$ with probability $p_b < p_a$ and $s_{\text{lo}}$ otherwise. Then for $\lambda \in [s_{\text{lo}}, s_{\text{hi}})$ the conditional risks are $r(\lambda \mid a) = p_a$, $r(\lambda \mid b) = p_b$, so $\Delta_s = (p_a - p_b)/2 =: \delta_s$ (the supremum over $\lambda$ is attained on this interval). Take the misspecified weights $\hat{w}(a) = 1 - t$, $\hat{w}(b) = 1 + t$ with $t = \delta_w \in (0,1]$; then $\mathbb{E}[\hat{w}] = 1$, $B = 1 + t < \infty$, and $\Delta_w = \frac{1}{2}t + \frac{1}{2}t = t$. The population $\hat{w}$-weighted risk on the interval is

$$R_{\hat{w}} = \tfrac{1}{2}(1-t)\,p_a + \tfrac{1}{2}(1+t)\,p_b = \underbrace{\tfrac{1}{2}(p_a + p_b)}_{=R_{\text{test}}} - t\cdot\frac{p_a - p_b}{2} = R_{\text{test}} - \Delta_w\Delta_s.$$

Choose $\alpha$ strictly between $R_{\hat{w}}$ and the next lower step of the weighted risk curve (the step at $s_{\text{hi}}$ has value 0; any $\alpha \in (0, R_{\hat{w}})$ with $\alpha$ close to $R_{\hat{w}}$ works after adjusting $p_a, p_b$). By the strong law, $\widehat{R}_w(\lambda) \to R_{\hat{w}}$ a.s. uniformly over the (finite) grid, so eventually the selection $\hat{\lambda}$ equals $s_{\text{lo}}$ whenever $\alpha \ge R_{\hat{w}}$ on that step—choose $\alpha = R_{\hat{w}} + \epsilon_n$ with any fixed margin $\epsilon > 0$ smaller than the gap to the adjacent steps; then $\hat{\lambda} = s_{\text{lo}}$ eventually a.s., and

$$R_{\text{test}}(\hat{\lambda}) = \tfrac{1}{2}(p_a + p_b) = R_{\hat{w}} + \Delta_w\Delta_s = \alpha - \epsilon + \Delta_w\Delta_s \;\xrightarrow[\epsilon\to 0]{}\; \alpha + \Delta_w\Delta_s.$$

The realized deployment risk exceeds the target by exactly the product term (scaling by $B_\ell$ for general bounded losses is immediate). Flipping the sign of $t$ attains the over-conservative side. □

The construction also isolates *why* the product form is the truth of the matter: the misspecified weights transfer mass between covariate groups ($\Delta_w$), and this matters exactly in proportion to how differently the groups score ($\Delta_s$). With a pivotal score ($p_a = p_b$) the same weight error is harmless, as Corollary 5 predicts.

### B.2 A Robust Threshold under a Weight Band

The posterior of §3.1 conditions on the estimated weight function; weight error enters only through the certificate's $\Delta_w\Delta_s$ term, invisibly to $\sigma_{\mathrm{post}}$. A first-order completion prices it with a sensitivity band: take every weight function within a multiplicative factor of the estimate to be plausible, and deploy the worst case over the band. The band $w/\hat{w} \in [e^{-\varepsilon}, e^{\varepsilon}]$ is the marginal sensitivity model of Tan (2006), used for conformal prediction under hidden confounding by Jin et al. (2023); here it enters the posterior rather than the quantile rule.

**Proposition 12** (Robust HPD threshold)**.** *Fix $\varepsilon \geq 0$ and let $\mathcal{W}_\varepsilon = \{w : e^{-\varepsilon} \leq w_i/\hat{w}_i \leq e^{\varepsilon},\ i \leq n{+}1\}$ (the atom included, $\hat{w}_{n+1} = \bar{w}$). For the exceedance loss $\ell(\lambda; x, y) = \mathbf{1}\{s(x,y) > \lambda\}$ define, for each $\lambda$, the bang-bang configuration $w^\varepsilon(\lambda)$: $w_i = \hat{w}_i e^{+\varepsilon}$ for calibration points with $s_i > \lambda$ and for the atom, $w_i = \hat{w}_i e^{-\varepsilon}$ otherwise; let $c_\varepsilon(\lambda) = \Pr(L^+_{w^\varepsilon(\lambda)}(\lambda) \leq \alpha \mid \ell_{1:n})$ and $\lambda^\beta_\varepsilon = \inf\{\lambda : c_\varepsilon(\lambda) \geq \beta\}$. Then, over the grid of calibration scores: (i) $c_\varepsilon(\lambda) = \inf_{w\in\mathcal{W}_\varepsilon} \Pr(L^+_w(\lambda) \leq \alpha \mid \ell_{1:n})$; (ii) $\lambda^\beta_\varepsilon = \sup_{w\in\mathcal{W}_\varepsilon} \lambda^\beta_{\mathrm{HPD}}(w)$; (iii) if the true ratio lies in the band, every data-conditional guarantee of §3.2 stated under the true-ratio posterior holds at $\lambda^\beta_\varepsilon$.*

*Proof.* (i) Fix the exponentials $E_{1:n+1}$ of Eq. (6) and a candidate $\lambda$. Writing $A = \sum_{i:s_i>\lambda} w_i E_i + w_{n+1}E_{n+1}$ and $B = \sum_{i:s_i\leq\lambda} w_i E_i$, the realized loss variable is $L^+_w(\lambda) = A/(A+B)$, increasing in $A$ and decreasing in $B$. Over the product band, $A$ is maximized and $B$ minimized—simultaneously, as they share no coordinate—exactly at $w^\varepsilon(\lambda)$, so $L^+_{w^\varepsilon(\lambda)}(\lambda) \geq L^+_w(\lambda)$ pathwise for every $w \in \mathcal{W}_\varepsilon$, and the event $\{L^+ \leq \alpha\}$ is smallest at the bang-bang configuration; it is attained since $w^\varepsilon(\lambda) \in \mathcal{W}_\varepsilon$. (ii) When $\lambda$ crosses a score, one term moves from $A$ to $B$ with both reweighting factors moving favorably, so $c_\varepsilon$ is nondecreasing and $\lambda^\beta_\varepsilon$ is well defined. For any $w \in \mathcal{W}_\varepsilon$, (i) gives $\Pr(L^+_w(\lambda^\beta_\varepsilon) \leq \alpha \mid \ell_{1:n}) \geq c_\varepsilon(\lambda^\beta_\varepsilon) \geq \beta$, hence $\lambda^\beta_{\mathrm{HPD}}(w) \leq \lambda^\beta_\varepsilon$; conversely, for any grid point $\lambda < \lambda^\beta_\varepsilon$ the member $w^\varepsilon(\lambda)$ itself has credibility below $\beta$ at $\lambda$, so the supremum is attained. (iii) By (ii), $\lambda^\beta_\varepsilon \geq \lambda^\beta_{\mathrm{HPD}}(w^*)$; realized risk is non-increasing in the threshold, so every guarantee at $\lambda^\beta_{\mathrm{HPD}}(w^*)$ transfers. □

The robust threshold costs one prefix-sum pass per $\varepsilon$—the same posterior draws serve every configuration. For general bounded monotone losses the worst case is a level-set tilt at a loss cutoff solving a one-dimensional fixed point; we state the exceedance case, which covers the miscoverage experiments. The band radius $\varepsilon$ is a deployment dial alongside $\alpha$ and $\beta$: it cannot be estimated from unlabeled data (the same epistemic boundary as $\Delta_w$; Remark 2), but its price is computable before deployment.

**The mis-specified arm, rescued.** Table 5 reruns the raw-softmax Fashion-MNIST arm of §4.3—the arm whose deployed failure rate is $28.7\%$ against the nominal $5\%$—with the robust threshold. A band of $\varepsilon = 0.1$ restores the nominal rate at a $7\%$ set-size premium; by $\varepsilon = 0.2$ the arm is conservative. The per-point weight errors are heavy-tailed ($|\log(w^*/\hat{w})|$: median $0.37$, 95th percentile $7.4$), yet a small band suffices: the bang-bang worst case prices the tail without matching it pointwise.

Table 5: Robust HPD threshold on the mis-specified (raw-softmax) Fashion-MNIST arm: failure frequency, mean realized risk, and mean prediction-set size (of 10 classes) against the band radius $\varepsilon$ (2 tilt strengths $\times$ 200 splits, $\alpha = 0.1$, $\beta = 0.95$).

| $\varepsilon$ | 0 | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.5 | 1.0 |
|---|---|---|---|---|---|---|---|---|
| Rel. freq. | 28.7% | 5.2% | 0.8% | 0.2% | 0.0% | 0.0% | 0.0% | 0.0% |
| Mean risk | 9.5% | 8.6% | 7.7% | 7.0% | 6.3% | 5.7% | 3.3% | 1.1% |
| Mean size | 3.30 | 3.42 | 3.54 | 3.66 | 3.78 | 3.89 | 4.43 | 5.20 |

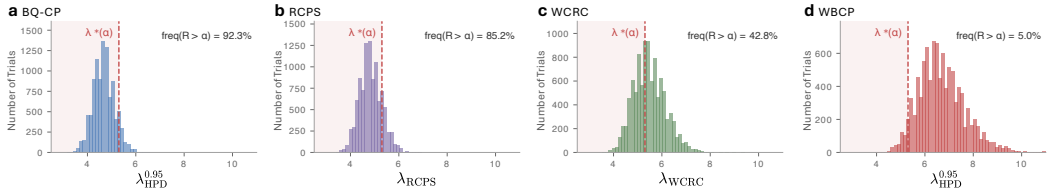



Figure 3: Distribution of each rule's selected threshold across the 10,000 trials of the synthetic regression experiment ($\gamma = 1$, estimated weights; main text Table 1). The population risk under the tilted test law is a known decreasing function of $\lambda$, so the deployed risk exceeds $\alpha$ exactly when the selection falls below $\lambda^*(\alpha)$ (shaded region); the annotated frequency is the relative frequency reported in the table. $\lambda^{0.95}_{\mathrm{HPD}}$ denotes the HPD selection under each Bayesian arm's own posterior—equal-weight in (a), weighted in (d). The blind rules select almost entirely inside the violation region; W-CRC straddles $\lambda^*(\alpha)$, violating in individual trials while its mean risk sits at the target; WBCP places all but 5.0% of its selections on the safe side, at visibly larger thresholds—the price of certification.

# C Experimental Details and Full Sweeps

This appendix records what the main-text tables compress: the full design of each benchmark, the sweeps behind the reported cells, the real-data protocols, and two checks that live entirely here—the numerical verification of tightness and the calibration of the $\beta$ dial.

## C.1 Synthetic Multilabel: Design and Full Sweeps

**Design.** BQ-CP's released experiments (synthetic binomial, $n = 10$; synthetic heteroskedastic; MS-COCO) compare HPD selection against CRC's $(n/(n+1))\widehat{R} + B/(n+1)$ rule and RCPS's Hoeffding UCB at target risk 0.4, failure rate 0.05, 1000 trials—with no shift axis. We adopt that design for comparability and add the axes this paper is about. Covariate $x \sim N(0,1)$; each unit carries $K = 4$ outcomes $T_k \sim \mathrm{Exp}(\mathrm{rate}(x))$ with $x$-dependent rate, and the bounded monotone loss is the exceedance fraction $\ell(\lambda; x) = \frac{1}{K}\#\{k : T_k > \lambda\}$—a scaled count loss with $\Delta_s > 0$ by construction. Shift: test covariates from the exponential tilt $\propto e^{\gamma x}$, $\gamma \in \{0, 1, 2\}$; weights either oracle ($w^*(x) = e^{\gamma x - \gamma^2/2}$) or split-estimated (A2-compliant); $n \in \{10, \ldots, 250\}$, 10,000 trials. Arms: BQ-CP (the equal-weight posterior applied blindly under the shift), W-CRC (CRC Prop. 2's transductive test-point mass form, reducing to the $n+1$ rule at equal weights), and WBCP's deployed rule at $\beta = 0.95$.

**Three regimes, as the budget predicts.** At $\gamma = 0$ (BQ-CP's own setting, replicated): WBCP's violation rate stays below nominal (0.3–2.2%) at every $n$ with no abstention—and here WBCP *is* BQ-CP, by the equal-weight special case—while W-CRC, an expectation-only guarantee, violates in 24–33% of trials by design. At $\gamma = 1$ ($n_{\mathrm{eff}} \approx 6$–100): the shift-blind BQ-CP baseline exhibits the starkest failure in the suite—its violation rate *rises* to 100% as $n$ grows, because more data concentrates its credibility on the wrong risk curve—while WBCP moves from abstention (93% of trials at $n = 10$) to certifying every trial at 7.5–7.9%, and split-estimated weights are nearly indistinguishable from oracle. At $\gamma = 2$ ($n_{\mathrm{eff}} \approx 3$–22): the blind rules certify and fail catastrophically—BQ-CP violates in 91–100% of trials (risk up to 0.74 against 0.4)—while WBCP abstains in every trial up to $n = 100$ and in 75% at $n = 250$, never violating when it does certify: at these $n_{\mathrm{eff}}$ the certificate's budget exceeds the $\alpha$-headroom, and the posterior says so rather than certifying anyway. We report the boundary rather than truncating the sweep before it.

## C.2 Synthetic Regression: Selection Distributions and the $\gamma$-Sweep

Two figures unpack the regression rows of §4. Figure 3 plots each rule's selected threshold across the 10,000 trials at $\gamma = 1$: the relative frequencies of Table 1 are the mass each selection distribution places below $\lambda^*(\alpha)$, so the table's single numbers become visible as distributions. Figure 4 then sweeps the shift strength $\gamma$: panel (a) tracks validity, panel (b) the width every valid rule must pay as the shift grows.

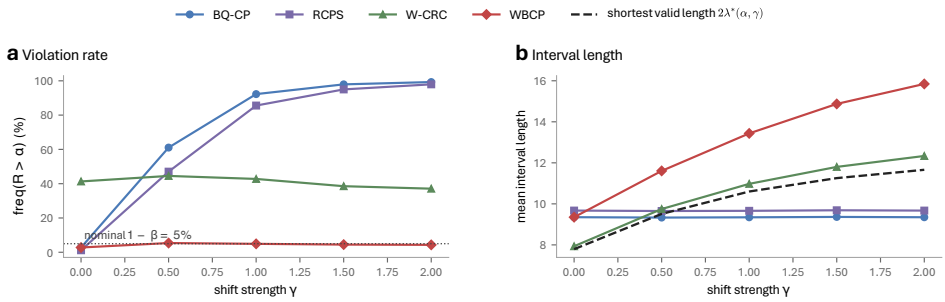



Figure 4: Validity and width against shift strength on the synthetic regression benchmark (estimated weights, 10,000 trials per $\gamma$). (a) WBCP's violation rate stays at the nominal $5\%$ across the entire sweep while the blind rules climb toward $100\%$; W-CRC's expectation semantics hold its mean at target with roughly $40\%$ of individual trials violating. (b) The shortest valid length $2\lambda^*(\alpha,\gamma)$—below it, the population risk exceeds $\alpha$ for any rule—rises with the shift. The blind rules' lengths stay flat and fall below it near $\gamma \approx 0.5$, exactly where their violation rates take off. W-CRC's *mean* length exceeds the valid length, yet its selection distribution straddles $\lambda^*(\alpha)$ trial by trial (Figure 3): mean control pins the center of the distribution at the boundary, not its quantiles. WBCP shifts the whole distribution above the boundary, which is both why it violates in only $5\%$ of trials and why its mean length is largest. At $\gamma = 0$ the two Bayesian arms coincide.

### C.3 Real Data: Protocols

*PovertyMap-WILDS* (Koh et al., 2021) (main text §4.3): scores are absolute residuals of a frozen asset-wealth regressor; per (country, seed) the test pool is the held-out country, the calibration pool the remaining countries, and each of 200 re-splits applies the covariate split of §4 with the discriminator fit on image embeddings. A re-split enters evaluation only if the discriminator AUC is at most $0.95$ and $n_{\text{eff}}/|C_s|$ at least $0.05$—the overlap screen of §4.3; failing cells are reported, not silently dropped (747 of 2,600 pass). *Fashion-MNIST* (main text §4): the dataset is split $50/50$ into probe-training and held-out pools; the probe is PCA-50 + multinomial logistic (held-out accuracy $0.84$). The tilt resamples the held-out pool by accept–reject with acceptance $\sigma(\gamma z)$, where $z$ is the standardized probability-space statistic most correlated with the true-class APS score, sign-aligned so $\gamma > 0$ tilts toward hard inputs; $P(Y \mid X)$ is untouched and the oracle ratio is $a(x)/\mathbb{E}_{\text{cal}}[a]$ in closed form. $\gamma \in \{\pm 2, \pm 8\}$, 200 seeds per cell, calibration capped at 3,000.

### C.4 The Mis-Specified Discriminator, in Detail

(Main text §4.) The raw-softmax discriminator is the mis-specified contrast: it cannot represent the entropy tilt, so the estimated weights flatten toward uniform rather than blowing up ($n_{\text{eff}} \approx 0.8n$, discriminator AUC unremarkable, no overlap diagnostic fires) and the deployed arm violates in $28.0\%$ $[23.7, 32.7]$ of hard-direction splits while the oracle arm is unchanged. This is $\Delta_w$'s undetectable direction: collapse is visible through $n_{\text{eff}}$ (Remark 2); flattening is not. Easy direction: the blind rules over-cover (realized risk $1.7\%$ against the $10\%$ target, half the risk budget spent on oversized sets) while WBCP tracks the target ($9.0\%$). W-CRC pins the mean at target while violating in roughly half the seeds, separating the expectation and failure-rate semantics.

### C.5 Tightness, Numerically

The two-point construction of Proposition 11, simulated at $n = 2 \times 10^5$ over a $4 \times 4$ grid of $(\Delta_w, \Delta_s)$: the realized deviation of the deployed risk from target matches the predicted $\Delta_w \Delta_s - \epsilon$ to four decimal places in all 16 cells. The product term is not an upper-bound artifact; it is the exact price.

### C.6 The Failure Rate Tracks $1 - \beta$ across the Whole Dial

Every experiment above fixes $\beta = 0.95$. The guarantee is a dial, not a point: on the heteroskedastic regression benchmark ($\gamma = 1$, estimated weights, protocol of §4), sweeping $\beta$ and recomputing the deployed rule per trial gives a realized failure frequency that tracks $1 - \beta$ across the range (Table 6). The credibility curve $\lambda \mapsto \Pr(L_w^+(\lambda) \le \alpha \mid \ell_{1:n})$ is $\beta$-free, so the sweep reuses one posterior pass

per trial; the clamp $\hat{\lambda}$ never binds in this regime, and abstention appears only at $\beta = 0.99$ (0.6% of trials).

Table 6: $\beta$-calibration on synthetic heteroskedastic regression under covariate shift, estimated weights ($\alpha = 0.1$, 4,000 trials per row; Clopper–Pearson 95% intervals).

| $\beta$ | Budget $1-\beta$ | Realized failure | 95% CI |
|---|---|---|---|
| 0.50 | 50.0% | 47.9% | [46.3, 49.4] |
| 0.70 | 30.0% | 28.0% | [26.6, 29.4] |
| 0.85 | 15.0% | 14.0% | [13.0, 15.1] |
| 0.90 | 10.0% | 9.8% | [8.9, 10.8] |
| 0.95 | 5.0% | 5.1% | [4.5, 5.9] |
| 0.975 | 2.5% | 2.7% | [2.2, 3.3] |
| 0.99 | 1.0% | 1.3% | [1.0, 1.7] |

# D Full Proofs

## D.1 Notation and Setup

We use the notation from §3.2. Additionally: sorted scores $\rho_{(1)} \le \cdots \le \rho_{(n)}$; normalized weights $\tilde{w}_{(i)}$ sorted by score; cumulative weights $p_j = \sum_{i=1}^{j} \tilde{w}_{(i)}$; effective sample size $n_{\text{eff}} = (\sum w_i)^2 / \sum w_i^2$; quantile level $q = 1 - \alpha$; deterministic index $k^* = \min\{j : p_j \ge q\}$; posterior threshold $\lambda = \rho_{(K)}$ where $K = \min\{j : S_j \ge q\}$ with $S_j = \sum_{i=1}^{j} V_{(i)}$ and $(V_1, \ldots, V_n) \sim \text{Dir}(n_{\text{eff}} \cdot \tilde{w}_{(1)}, \ldots, n_{\text{eff}} \cdot \tilde{w}_{(n)})$.

## D.2 Foundational Lemma: Dirichlet Aggregation

**Lemma 13** (Dirichlet Aggregation). *Let* $(V_1, \ldots, V_n) \sim \text{Dir}(\alpha_1, \ldots, \alpha_n)$ *with* $\alpha_0 = \sum_i \alpha_i$. *For* $A \subseteq [n]$, $S_A = \sum_{i \in A} V_i \sim \text{Beta}(\alpha_0 p_A, \alpha_0(1 - p_A))$ *where* $p_A = \sum_{i \in A} \alpha_i / \alpha_0$. *Thus* $\mathbb{E}[S_A] = p_A$ *and* $\text{Var}(S_A) = p_A(1 - p_A)/(\alpha_0 + 1)$.

*For the WBCP cumulative weight:*

$$\text{Var}(S_j) = \frac{p_j(1 - p_j)}{n_{\text{eff}} + 1}. \tag{9}$$

*Proof.* By the Dirichlet aggregation property, merging components in $A$ gives $(S_A, 1 - S_A) \sim \text{Dir}(\sum_{i \in A} \alpha_i, \sum_{i \notin A} \alpha_i)$, i.e., $S_A \sim \text{Beta}(a, b)$ with $a = \alpha_0 p_A$, $b = \alpha_0(1 - p_A)$. The Beta moments give $\mathbb{E}[S_A] = a/(a + b) = p_A$ and $\text{Var}(S_A) = ab/((a + b)^2(a + b + 1)) = p_A(1 - p_A)/(\alpha_0 + 1)$. For WBCP, $\alpha_0 = n_{\text{eff}}$. □

## D.3 Proof of Theorem 7 (Moment Matching)

Throughout, $\tilde{\boldsymbol{V}}$ is the exact posterior of Theorem 6. Since $\tilde{\boldsymbol{V}}$ is invariant under rescaling of $\boldsymbol{w}$, we normalize $\sum_i \tilde{w}_i = 1$ and write

$$T = \sum_{j=1}^{n} \tilde{w}_j E_j, \qquad \nu = \sum_j \tilde{w}_j^2 = \frac{1}{n_{\text{eff}}}, \qquad m = \max_j \tilde{w}_j,$$

with $E_j$ i.i.d. Exp(1), so that $\tilde{V}_j = \tilde{w}_j E_j / T$. Note $\mathbb{E}[T] = 1$, $\text{Var}(T) = \nu$, and $m \ge \sum_j \tilde{w}_j^2 = \nu$, so $\kappa := m n_{\text{eff}} \ge 1$ always. The hypotheses of the theorem read $m \le \kappa / n_{\text{eff}}$ and $n_{\text{eff}} \ge 18\kappa$, the latter being exactly $1/m \ge 18$.

Write $\xi_j = E_j - 1$, so $\mathbb{E}\xi_j = 0$, $\mathbb{E}\xi_j^2 = 1$, $\mathbb{E}\xi_j^3 = 2$ and $\mathbb{E}\xi_j^4 = 9$, and set $D = T - 1 = \sum_j \tilde{w}_j \xi_j$.

**Lemma 14** (Inverse moments of the normalizer). *For every integer* $p \ge 1$ *with* $pm < 1$,

$$\mathbb{E}\big[T^{-p}\big] \;\le\; (1 - pm)^{-p}.$$

*In particular, under* $1/m \ge 18$ *we have* $\mathbb{E}[T^{-p}] \le (9/5)^8$ *for all* $p \le 8$.

*Proof.* For $p \geq 1$, $T^{-p} = \Gamma(p)^{-1} \int_0^\infty s^{p-1} e^{-sT}\, ds$, and by Tonelli and independence,

$$\mathbb{E}[T^{-p}] = \frac{1}{\Gamma(p)} \int_0^\infty s^{p-1} \prod_j (1 + s\tilde{w}_j)^{-1} ds.$$

The map $x \mapsto \log(1+sx)/x$ is decreasing on $(0,\infty)$, so $\log(1+s\tilde{w}_j) \geq \tilde{w}_j \log(1+sm)/m$ for every $j$; summing over $j$ and using $\sum_j \tilde{w}_j = 1$ gives $\sum_j \log(1+s\tilde{w}_j) \geq \log(1+sm)/m$, i.e. $\prod_j (1+s\tilde{w}_j)^{-1} \leq (1+sm)^{-1/m}$. Substituting $u = sm$,

$$\mathbb{E}[T^{-p}] \leq \frac{m^{-p}}{\Gamma(p)} \int_0^\infty u^{p-1}(1+u)^{-1/m}\, du = \frac{m^{-p}}{\Gamma(p)} B\big(p, \tfrac{1}{m} - p\big) = m^{-p}\, \frac{\Gamma(1/m - p)}{\Gamma(1/m)},$$

the Beta integral converging precisely because $1/m > p$. Finally $\Gamma(\beta - p)/\Gamma(\beta) = \prod_{r=1}^{p} (\beta - r)^{-1} \leq (\beta - p)^{-p}$ with $\beta = 1/m$, so $\mathbb{E}[T^{-p}] \leq m^{-p}(1/m - p)^{-p} = (1 - pm)^{-p}$. For $p \leq 8$ and $m \leq 1/18$ this is at most $(1 - 8/18)^{-8} = (9/5)^8$. □

**Lemma 15** (Moments of centered linear forms)**.** *Let $L = \sum_j a_j \xi_j$ with $\boldsymbol{a} \in \mathbb{R}^n$. Then $\mathbb{E}L = 0$, $\mathbb{E}L^2 = \sum_j a_j^2$, $\mathbb{E}L^3 = 2\sum_j a_j^3$, and $\mathbb{E}L^4 = 3(\sum_j a_j^2)^2 + 6\sum_j a_j^4$. More generally, for even $r$, $\mathbb{E}L^r \leq C_r (\sum_j a_j^2)^{r/2}$ whenever $\max_j |a_j| \leq (\sum_j a_j^2)^{1/2}$.*

*Proof.* The $\xi_j$ are independent and centered, so in expanding $\mathbb{E}L^r$ only terms in which every index appears at least twice survive. For $r = 2, 3$ this leaves $\sum_j a_j^2 \mathbb{E}\xi^2$ and $\sum_j a_j^3 \mathbb{E}\xi^3$. For $r = 4$ it leaves $\sum_j a_j^4 \mathbb{E}\xi^4 + 3\sum_{j \neq k} a_j^2 a_k^2 (\mathbb{E}\xi^2)^2 = 9\sum_j a_j^4 + 3[(\sum_j a_j^2)^2 - \sum_j a_j^4]$, which is the stated identity. The general bound is Rosenthal's inequality. □

*Proof of Theorem 7.* Fix $A \subseteq [n]$ and set $c_j = \tilde{w}_j(\mathbf{1}\{j \in A\} - p_A)$ and $W = \sum_j c_j E_j = \sum_j c_j \xi_j$, the second equality because $\sum_j c_j = p_A - p_A = 0$. Then

$$\tilde{S}_A - p_A = \frac{\sum_{i \in A} \tilde{w}_i E_i - p_A \sum_j \tilde{w}_j E_j}{T} = \frac{W}{T}, \tag{10}$$

an exact identity. By Lemma 15,

$$\mathbb{E}[W^2] = \sum_j c_j^2 = (1-p_A)^2 \sum_{j \in A} \tilde{w}_j^2 + p_A^2 \sum_{j \notin A} \tilde{w}_j^2 = \frac{q_A(1-p_A)^2 + (1-q_A)p_A^2}{n_{\text{eff}}}, \tag{11}$$

using $\sum_{j \in A} \tilde{w}_j^2 = q_A \nu$ and $\sum_{j \notin A} \tilde{w}_j^2 = (1 - q_A)\nu$. Two further quantities are needed: $\mathbb{E}[WD] = \sum_j c_j \tilde{w}_j = \sum_{j \in A} \tilde{w}_j^2 - p_A \nu = \nu(q_A - p_A)$, and, by Lemma 15 applied to $D$, $\mathbb{E}[D^2] = \nu$ and $\mathbb{E}[D^4] = 3\nu^2 + 6\sum_j \tilde{w}_j^4 \leq \nu^2(3 + 6\kappa^2/n_{\text{eff}}) \leq \nu^2(3 + \kappa/3)$, the last step using $\sum_j \tilde{w}_j^4 \leq m^2 \nu$, $m \leq \kappa\nu$ and $n_{\text{eff}} \geq 18\kappa$. Since $0 \leq \tilde{S}_A \leq 1$, (10) gives $|W/T| \leq 1$.

**Mean.** From the exact identity $1/T = 1 - D + D^2/T$,

$$\frac{W}{T} = W - WD + \Big(\frac{W}{T}\Big) D^2.$$

Taking expectations, $\mathbb{E}[W] = 0$ and $\mathbb{E}[WD] = \nu(q_A - p_A)$, so $\mathbb{E}[\tilde{S}_A] = p_A + (p_A - q_A)/n_{\text{eff}} + R_1$ with $R_1 = \mathbb{E}[(W/T)D^2]$. By Cauchy–Schwarz, $|R_1| \leq (\mathbb{E}[(W/T)^2])^{1/2} (\mathbb{E}[D^4])^{1/2}$. For the first factor, $\mathbb{E}[(W/T)^2] \leq (\mathbb{E}[W^4])^{1/2} (\mathbb{E}[T^{-4}])^{1/2} = O(\nu)$ by Lemmas 15 and 14; for the second, $(\mathbb{E}[D^4])^{1/2} \leq \nu(3 + \kappa/3)^{1/2}$. Hence $|R_1| = O(\kappa^{1/2}\nu^{3/2}) = O(\kappa^2 n_{\text{eff}}^{-3/2})$, since $\kappa \geq 1$.

**Variance.** From $W^2 - W^2/T^2 = (W/T)^2(T^2 - 1)$ and $T^2 - 1 = 2D + D^2$,

$$\mathbb{E}\Big[\Big(\frac{W}{T}\Big)^2\Big] = \mathbb{E}[W^2] - 2\,\mathbb{E}\Big[\Big(\frac{W}{T}\Big)^2 D\Big] - \mathbb{E}\Big[\Big(\frac{W}{T}\Big)^2 D^2\Big].$$

The last term is $O(\nu^2)$: by Cauchy–Schwarz it is at most $(\mathbb{E}[(W/T)^4])^{1/2} (\mathbb{E}[D^4])^{1/2}$, and $\mathbb{E}[(W/T)^4] \leq (\mathbb{E}[W^8])^{1/2} (\mathbb{E}[T^{-8}])^{1/2} = O(\nu^2)$ by Lemmas 15 and 14 (this is the largest inverse moment used, and is where $1/m \geq 18 > 8$ is required). For the middle term, expand once more using the exact identity $T^{-2} = 1 - 2D + D^2(2T+1)/T^2$:

$$\mathbb{E}\Big[\Big(\frac{W}{T}\Big)^2 D\Big] = \mathbb{E}[W^2 D] - 2\,\mathbb{E}[W^2 D^2] + \mathbb{E}\Big[\frac{W^2 D^3 (2T+1)}{T^2}\Big].$$

Here $\mathbb{E}[W^2 D] = 2\sum_j c_j^2 \tilde{w}_j$ by the third-moment computation of Lemma 15, whence $|\mathbb{E}[W^2 D]| \le 2m\sum_j c_j^2 = O(\kappa\nu^2)$; $\mathbb{E}[W^2 D^2] = O(\nu^2)$; and the final term is $O(\nu^{5/2})$ by Cauchy–Schwarz with Lemma 14. Therefore $\mathbb{E}[(W/T)^2 D] = O(\kappa\nu^2)$ and

$$\mathrm{Var}(\tilde{S}_A) = \mathbb{E}\Big[\Big(\frac{W}{T}\Big)^2\Big] - \Big(\mathbb{E}\Big[\frac{W}{T}\Big]\Big)^2 = \frac{q_A(1-p_A)^2 + (1-q_A)p_A^2}{n_{\mathrm{eff}}} + O(\kappa^2 n_{\mathrm{eff}}^{-3/2}),$$

the subtracted square being $O(\nu^2)$ by the mean expansion. This proves both displayed expansions.

**(1) Uniqueness.** Under $\mathrm{Dir}(c\tilde{w}_1, \ldots, c\tilde{w}_n)$ the aggregation Lemma 13 gives $S_A \sim \mathrm{Beta}(cp_A, c(1-p_A))$, so $\mathbb{E}[S_A] = p_A$ and $\mathrm{Var}(S_A) = p_A(1-p_A)/(c+1)$ exactly. Matching the variance against the expansion above and using the stated algebraic identity $q_A(1-p_A)^2 + (1-q_A)p_A^2 = p_A(1-p_A) + (q_A - p_A)(1-2p_A)$ gives

$$\frac{p_A(1-p_A)}{c+1} = \frac{p_A(1-p_A) + (q_A-p_A)(1-2p_A)}{n_{\mathrm{eff}}} + O(\kappa^2 n_{\mathrm{eff}}^{-3/2}).$$

Write $\delta = p_A(1-p_A)$, bounded away from 0 at the quantile boundary where $p_A \to q$. If $|q_A - p_A| \le \varepsilon$ then the numerator on the right is $\delta(1 + O(\varepsilon/\delta))$, so $c+1 = n_{\mathrm{eff}}\big(1 + O((\varepsilon + \kappa^2 n_{\mathrm{eff}}^{-1/2})/\delta)\big)$, which is the claim; and $\varepsilon \to 0$, $n_{\mathrm{eff}} \to \infty$ forces $c+1 = n_{\mathrm{eff}}$ exactly in the limit.

**(2) Impossibility.** Any Dirichlet with mean $\tilde{\boldsymbol{w}}$ is $\mathrm{Dir}(c\tilde{\boldsymbol{w}})$ for some $c > 0$, and its cumulative variance $p_A(1-p_A)/(c+1)$ depends on $A$ only through $p_A$. The exact variance (11) depends on $A$ through both $p_A$ and $q_A$. Hence if two sets $A, A'$ satisfy $p_A = p_{A'}$ but $q_A \ne q_{A'}$, no single $c$ matches both. Equivalently, at the level of singletons, matching every $A = \{i\}$ would require $(c+1) = n_{\mathrm{eff}}\, p_i(1-p_i)/\big[p_i(1-p_i) + (q_i - p_i)(1-2p_i)\big]$ to take the same value for all $i$; whenever the map $i \mapsto (q_i - p_i)(1-2p_i)/(p_i(1-p_i))$ is non-constant—which is the generic case as soon as $q_A \ne p_A$ for some $A$ with $p_A \ne 1/2$—no such $c$ exists. Since Dirichlet variances are insensitive to the squared-weight share $q_A$ while the exact ones are not, the mismatch cannot be removed within the Dirichlet family. □

### D.4 Calibration Consistency: Statement and Proof

The two results in this subsection and the next are the closed-form companions of the exact posterior; both are stated under the Dirichlet surrogate, which is where closed forms are available.

**Assumption 2** (Weighted Dirichlet surrogate)**.** $(V_1, \ldots, V_n) \sim \mathrm{Dir}(n_{\mathrm{eff}}\tilde{w}_{(1)}, \ldots, n_{\mathrm{eff}}\tilde{w}_{(n)})$—the two-moment surrogate of Proposition 1, reducing to $\mathrm{Dir}(1, \ldots, 1)$ at uniform weights; statements about the deployed risk use the completed vector of Algorithm 1. The deployed rule never invokes this assumption, sampling the exact posterior instead; where the two differ, at small $n_{\mathrm{eff}}$, the surrogate overstates the spread, so these closed forms err toward conservatism.

**Assumption 3** (Correct weight specification)**.** $w_i \propto d\tilde{P}_X/dP_X(X_i)$, and the local score distribution has a continuous positive density at the quantile—the conditions under which weighted CP states coverage.

**Theorem 16** (Calibration consistency)**.** *Under Assumption 2 with general concentration $c$, and with $p_j = \sum_{i \le j} \tilde{w}_{(i)}$, $q = 1-\alpha$, and $S_j$ the posterior mass of the first $j$ sorted scores: (1) $\mathbb{E}[S_j] = p_j$ exactly for all $c$; (2) the posterior variance of $S_j$ matches the frequentist variance $q(1-q)/n_{\mathrm{eff}}$ of the weighted empirical CDF iff $c = n_{\mathrm{eff}} - 1$ (we use $c = n_{\mathrm{eff}}$; relative difference $\le 1/n_{\mathrm{eff}}$); (3) asymptotically, the posterior quantile variance matches the frequentist sampling variance of the weighted quantile iff $c/n_{\mathrm{eff}} \to 1$ —any other concentration is over- or under-confident.*

*Proof.* **Part 1 (Mean consistency):** For any $c > 0$, the Dirichlet mean is $\mathbb{E}[V_i] = \tilde{w}_i$ (independent of $c$), hence $\mathbb{E}[S_j] = p_j$ exactly for every $j$.

**Part 2 (Variance calibration):** The Dirichlet model gives $\mathrm{Var}(S_j) = p_j(1-p_j)/(c+1)$. Under correct weight specification (Assumption 3), the weighted empirical CDF $\hat{F}_w(t) = \sum_i \tilde{w}_i \mathbf{1}[\rho_i \le t]$ targets the test CDF $F(t) = F_{\mathrm{test}}(t)$, and its leading-order variance is:

$$\mathrm{Var}_{\mathrm{freq}}(\hat{F}_w(t)) \approx F(t)(1-F(t)) \cdot \sum_i \tilde{w}_i^2 = \frac{F(t)(1-F(t))}{n_{\mathrm{eff}}},$$

using $\sum_i \tilde{w}_i^2 = 1/n_{\text{eff}}$ (the standard delta-method computation for the weighted empirical CDF under covariate shift). At the quantile $t = \lambda^*$ where $F(\lambda^*) = q \approx p_{k^*}$:

$$\text{Var}_{\text{freq}} = \frac{q(1-q)}{n_{\text{eff}}}, \quad \text{Var}_{\text{Dir}} = \frac{q(1-q)}{c+1}.$$

Matching requires $c + 1 \approx n_{\text{eff}}$, i.e., $c = n_{\text{eff}} - 1 \approx n_{\text{eff}}$ for large $n_{\text{eff}}$.

**Part 3 (Bernstein–von Mises):** By the delta method (van der Vaart, 1998, Theorem 2.7), the CDF-to-quantile transformation maps variance via division by $f(\lambda^*)^2$; the frequentist variance of the weighted sample quantile follows from the CLT for importance-weighted quantiles (Koul, 1992):

$$\sigma^2_{\text{post}} \approx \frac{q(1-q)}{(n_{\text{eff}}+1)f_q^2}, \quad \sigma^2_{\text{freq}} = \frac{q(1-q)}{n_{\text{eff}} \cdot f^2}.$$

Under correct specification, $f_q \to f$ and $(n_{\text{eff}} + 1)^{-1} \approx n_{\text{eff}}^{-1}$, confirming matching. □

### D.5 Proof of Theorem 2 (Weighted Stochastic Dominance)

*Proof.* The proof follows BQ-CP's Theorem 4.3 structure, adapted for the weighted Dirichlet.

**Step 1 (Worst-case quantile function):** By Proposition B.2 of Snell & Griffiths (2025)—a pure optimization result over quantile functions, independent of distributional assumptions—for quantile positions $t_1, \ldots, t_n$ and losses $\ell_1, \ldots, \ell_n$ bounded by $B_\ell$:

$$\sup_{K \in \mathcal{K}_n} \int_0^1 K(t)dt = \sum_{i=1}^{n} (t_{(i)} - t_{(i-1)}) \cdot \ell_{(i)} + (1 - t_{(n)}) \cdot B_\ell,$$

where $\mathcal{K}_n$ is the set of non-decreasing quantile functions with $K(t_i) = \ell_i$ and $K \le B_\ell$: on $(t_{(n)}, 1]$ the worst admissible quantile function sits at the bound. This depends only on monotonicity and boundedness.

**Step 2 (Marginalization over spacings):** Conditioned on spacings $V_i = t_{(i)} - t_{(i-1)}$ ($i \le n$) and $V_{n+1} = 1 - t_{(n)}$, the worst-case risk is $\sum_{i \le n} V_i \ell_{(i)} + V_{n+1} B_\ell$. Marginalising over the law of the completed mass vector—any law on the simplex, in particular the exact posterior of Proposition 1:

$$\inf_{\pi} \Pr(L \le b \mid \ell_{1:n}) \ge \Pr\Big(\sum_{i \le n} V_i \ell_{(i)} + V_{n+1} B_\ell \le b\Big) = \Pr(L_w^+ \le b).$$

**Step 3 (Validity of substitution):** BQ-CP's proof uses the spacings distribution only in the final marginalization step (their equation (73)). All preceding steps—change of integration order, pushing infimum inside, applying the worst-case bound—are purely analytic and hold for *any* distribution on the simplex. Replacing $\text{Dir}(1, \ldots, 1)$ by the exact posterior of Eq. (3), or by any surrogate for it, is therefore valid; the dominance statement is about the law the rule samples, whichever that is. □