Distribution-Free Risk Certificates under Covariate Shift

Anonymous Authors

Abstract

Predictive systems are commonly trained and validated on one population but deployed on another. Importance weighting corrects the resulting covariate shift asymptotically, yet a point estimate of target risk is not a deployment certificate. We study finite-sample upper certificates for bounded loss when the target-to-source density ratio is bounded and may be estimated from unlabeled data. A single held-out sample yields a simultaneous certificate for a finite family of predictors, allowing a practitioner to choose the most useful model whose target risk is certified below a prescribed tolerance. The guarantee separates three costs: observed weighted loss, statistical uncertainty, and density-ratio error. We also give an empirical-Bernstein refinement, a selection rule, and a deterministic stress test that makes the dependence on sample size, shift severity, and model multiplicity explicit. The result is intentionally simple: every assumption is auditable, the proof fits on one page, and failure modes are visible in the certificate itself.

1 Introduction

Machine-learning evaluation usually treats a held-out source sample as a proxy for future performance. That proxy fails when deployment changes the distribution of inputs. Examples include a medical model transferred between hospitals, a vision model exposed to a new sensor, or a ranking system serving a new region. Dataset shift is therefore not merely a predictive problem; it is a problem of deciding whether the available evidence supports deployment (Quionero-Candela et al., 2009).

Under *covariate shift*, the conditional law of the outcome given the input is stable while the input marginal changes. Importance weighting then identifies the target risk from labeled source data. Classical work develops density-ratio estimators and weighted learning procedures (Sugiyama et al., 2007); weighted conformal methods similarly recover predictive coverage under shift (Tibshirani et al., 2019). In contrast, we focus on a narrower operational question:

> Given several already-trained predictors, can a held-out source sample certify that the chosen predictor's target risk is below a stated tolerance?

The distinction matters. An unbiased or consistent estimator need not be a valid upper bound, and selecting the smallest estimate among many candidates invalidates a pointwise guarantee. Our construction addresses both issues with a simultaneous concentration bound. It makes the following contributions:

- a distribution-free target-risk certificate for bounded losses and an independently estimated, clipped density ratio;
- a uniform version that remains valid after selecting among a finite model family, plus a variance-adaptive refinement; and
- an auditable deployment rule and deterministic stress tests that expose the effects of sample size, overlap, and multiplicity.

The message is deliberately modest. A certificate cannot repair a violation of covariate shift or a lack of overlap. It instead converts explicit assumptions into a finite-sample statement and becomes conservative when those assumptions are weak.

2 Related Work

Dataset shift and importance weighting. Covariate shift is one member of a broader taxonomy of train–test mismatch (Quionero-Candela et al., 2009). Direct density-ratio estimation avoids separately estimating two high-dimensional densities (Sugiyama et al., 2007). Kernel two-sample methods provide useful diagnostics, although failure to reject equality does not establish the covariate-shift assumption (Gretton et al., 2012). Distributionally robust optimization targets performance across a prespecified set of distributions or groups (Sagawa et al., 2020); our objective is instead a confidence bound for one target population.

Distribution-free uncertainty. Conformal prediction gives marginal coverage guarantees under exchangeability (Vovk et al., 2005; Lei et al., 2018). Weighted conformal prediction extends this logic to covariate shift when likelihood ratios are known (Tibshirani et al., 2019). Risk-controlling prediction generalizes coverage to bounded losses (Angelopoulos et al., 2024). Our bound is a compact importance-weighted analogue aimed at post-training model selection rather than construction of prediction sets.

Selective prediction. Abstaining systems trade coverage for conditional accuracy (Geifman and El-Yaniv, 2017). Each abstention threshold can be treated as a candidate in our finite family, so selection over thresholds is covered without reusing the certification sample for training. The guarantee concerns the unconditional bounded loss specified by the

user; conditional risk requires a separate treatment of the random acceptance probability.

3 Problem Setup

Let P be a source distribution and Q a target distribution on $\mathcal{X} \times \mathcal{Y}$. For a predictor $f \in \mathcal{F}$, let $\ell(f(X), Y) \in [0, 1]$ be a measurable loss and define

$$\mathcal{R}_Q(f) = \mathbb{E}_{(X,Y)\sim Q}[\ell(f(X), Y)]\,. \tag{1}$$

The predictors $\mathcal{F}_K = \{f_1, \ldots, f_K\}$ are trained without the certification sample $S = \{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} P$.

Assumption 1 (Covariate shift)**.** $P(Y \mid X) = Q(Y \mid X)$ almost surely, and Q_X is absolutely continuous with respect to P_X.

The corresponding density ratio is $w(x) = dQ_X/dP_X(x)$. By a change of measure, Assumption 1 implies

$$\mathcal{R}_Q(f) = \mathbb{E}_P[w(X)\ell(f(X), Y)]\,. \tag{2}$$

Assumption 2 (Auditable overlap and ratio error)**.** An estimator $\widehat{w} : \mathcal{X} \to [0, B]$, fitted independently of S, satisfies

$$\mathbb{E}_{X\sim P_X} |\widehat{w}(X) - w(X)| \le \varepsilon_w \tag{3}$$

for declared constants $B < \infty$ and $\varepsilon_w \ge 0$.

Independence can be obtained with a three-way split: predictor training, ratio estimation using unlabeled source and target inputs, and certification. The error budget ε_w may come from domain knowledge, a separate validation study, or a high-probability bound for a particular ratio estimator. Treating an unvalidated estimate as though $\varepsilon_w = 0$ is not justified.

4 Simultaneous Certificates

For each candidate define the estimated weighted risk

$$\widehat{R}_k = \frac{1}{n}\sum_{i=1}^n \widehat{w}(X_i)\ell(f_k(X_i), Y_i). \tag{4}$$

The following result is the core certificate.

Theorem 1 (Finite-family target-risk certificate)**.** *Under Assumptions 1–2, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over S, simultaneously for every $k \in \{1, \ldots, K\}$,*

$$\mathcal{R}_Q(f_k) \le U_k := \min\left\{1,\ \widehat{R}_k + \varepsilon_w + B\sqrt{\frac{\log(K/\delta)}{2n}}\right\}. \tag{5}$$

Consequently, the inequality remains valid for any data-dependent index $\widehat{k} = \widehat{k}(S)$ taking values in $\{1, \ldots, K\}$.

Proof. Fix k and condition on the fitted predictors and ratio estimator. Put $Z_{ik} = \widehat{w}(X_i)\ell(f_k(X_i), Y_i) \in [0, B]$. Hoeffding's inequality gives

$$\mathbb{P}\left(\mathbb{E}_P[Z_{ik}] - \widehat{R}_k > B\sqrt{\frac{\log(K/\delta)}{2n}}\right) \le \frac{\delta}{K}.$$

Moreover, boundedness of the loss and equation (3) imply

$$\mathbb{E}_P[w\ell_k] \le \mathbb{E}_P[\widehat{w}\ell_k] + \mathbb{E}_P\,|w - \widehat{w}| \le \mathbb{E}_P[Z_{ik}] + \varepsilon_w.$$

Apply a union bound over k and use equation (2). Clipping the upper bound at one is valid because the loss lies in $[0, 1]$. Since all K inequalities hold on one event, substituting the selected index $\widehat{k}(S)$ preserves the claim. □

The radius scales as $B\sqrt{\log(K/\delta)/n}$. Thus poor source–target overlap is costly through B, while searching a larger finite family has only logarithmic cost. The additive ε_w term does not vanish with more certification labels; better labels cannot compensate for an unvalidated ratio estimator.

4.1 Variance-adaptive refinement

Hoeffding's bound ignores that the observed weighted losses may have low variance. Let

$$V_k = \frac{1}{n-1}\sum_{i=1}^n (Z_{ik} - \widehat{R}_k)^2, \qquad n \ge 2.$$

An empirical-Bernstein inequality (Maurer and Pontil, 2009) yields the alternative certificate

$$\rho_k^{\text{EB}} = \sqrt{\frac{2V_k \log(2K/\delta)}{n}} + \frac{7B\log(2K/\delta)}{3(n-1)},$$
$$U_k^{\text{EB}} = \min\left\{1,\ \widehat{R}_k + \varepsilon_w + \rho_k^{\text{EB}}\right\}. \tag{6}$$

The same union-bound argument makes equation (6) simultaneous. A predeclared choice of the smaller of equation (5) and equation (6) requires allocating failure probability to both bounds; one may replace δ by $\delta/2$ in each and take their minimum.

4.2 Certified model selection

Suppose a_k is a utility score computed without S (for example, target-input coverage for an abstaining predictor), and τ is the maximum acceptable target risk. Define

$$\widehat{k} \in \operatorname*{arg\,max}_{k:U_k \le \tau} a_k, \tag{7}$$

returning "no certified model" if the feasible set is empty. By theorem 1, any returned model satisfies $\mathcal{R}_Q(f_{\widehat{k}}) \le \tau$ with probability at least $1 - \delta$.

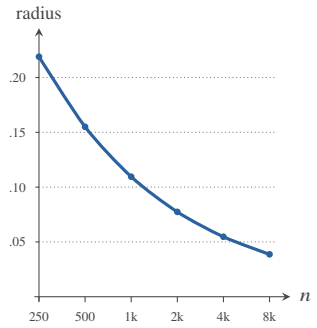

Figure 1: Deterministic evaluation of the Hoeffding radius in equation (5) for $B = 2$, $K = 20$, and $\delta = 0.05$. The curve is not experimental data; it displays the exact bound as n varies.

Table 1: Exact values of equation (5) as n and ε_w vary.

n	radius	U_k ($\varepsilon_w = 0$)	.015	.050
500	.155	.228	.243	.278
2000	.077	.150	.165	.200
8000	.039	.112	.127	.162

5 Deterministic Stress Tests

This section evaluates the algebra of the certificate rather than claiming an empirical comparison. Every table entry is obtained by direct substitution into equation (5); no random trial or hidden dataset is involved. This is useful because the dominant tradeoffs can be inspected before collecting labels.

5.1 Sample size and systematic ratio error

Fix $K = 20$, $\delta = 0.05$, $B = 2$, and an observed weighted loss $\widehat{R}_k = 0.073$. Table 1 reports $U_k = \widehat{R}_k + \varepsilon_w +$radius (all values are below one). Quadrupling n halves the statistical radius, as predicted, but it leaves the ratio-error term unchanged.

5.2 Overlap and model multiplicity

For $n = 2000$ and $\delta = 0.05$, table 2 isolates the statistical radius. Doubling the ratio cap doubles uncertainty, whereas multiplying the model count by fifty has a smaller, logarithmic effect. Large B is therefore a warning about unsupported target regions, not merely a nuisance tuning parameter.

5.3 What a deployment decision looks like

Consider a risk tolerance $\tau = 0.15$ and the first column of table 1. With $n = 500$, a candidate having weighted loss .073 cannot be certified even if the ratio is known exactly. At $n = 2000$, the bound is approximately .150 before ratio error and becomes infeasible for any positive ε_w. At $n = 8000$, an error budget as large as .038 remains feasible. The calculation gives a concrete data-acquisition target: gather more certification labels, improve overlap, validate the ratio more tightly, or decline deployment.

Table 2: Hoeffding radius $B\sqrt{\log(K/\delta)/(2n)}$.

	B		
K	1	2	4
2	.030	.061	.121
20	.039	.077	.155
200	.046	.091	.182
1000	.050	.100	.199

6 Practical Diagnostics

The theorem's probability statement is only as meaningful as its assumptions. Before issuing a certificate, we recommend the following audit.

Check support, not only average fit. Inspect the distribution of $\widehat{w}(X)$ on held-out source inputs and the fraction clipped at B. Frequent clipping suggests that the numerical guarantee is tied to a poor approximation of w; this should increase ε_w or stop the analysis. Compare source and target representations using domain-relevant plots and two-sample diagnostics (Gretton et al., 2012).

Probe conditional stability. Covariate shift asserts equality of $P(Y \mid X)$ and $Q(Y \mid X)$, which cannot in general be verified from unlabeled target inputs. Small, carefully sampled target label audits are more informative than elaborate marginal tests. If conditional shift is plausible, add a separately justified sensitivity budget $\varepsilon_{\text{cond}}$ to the right side of equation (5).

Preserve the holdout. The model family, loss, clipping level, and bound variant must be fixed before examining certification outcomes. If the sample is reused iteratively, the simple union bound no longer describes the search. Fresh data, nested splitting, or tools for adaptive data analysis are then needed.

Choose the loss to match the harm. For classification error, $\ell = \mathbf{1}\{f(X) \neq Y\}$ is immediate. Cost- sensitive or subgroup-weighted harms can also be normalized to $[0, 1]$. A low average loss does not imply low risk for every subgroup; certify each prespecified group separately and include those certificates in the multiplicity budget.

7 Limitations and Broader Impact

Three limitations are fundamental. First, the result assumes covariate shift; it does not detect changes in the outcome mechanism. Second, the declared ratio-error budget is external to the theorem. A practitioner who sets it optimistically can produce a mathematically correct calculation with an invalid premise. Third, bounded importance weights trade bias against variance: clipping stabilizes the estimate but can enlarge ε_w.

Algorithm 1: Split, certify, and select

1. Split all available data before fitting: use labeled source data to train $f_1, \ldots, f_K$; independent unlabeled source/target inputs to fit $\widehat{w}$; and a labeled source sample S only for certification.
2. Declare B, ε_w, δ, and the acceptable risk τ. Clip the fitted ratio into $[0, B]$ and document how ε_w was obtained.
3. On S, compute $Z_{ik} = \widehat{w}(X_i)\ell(f_k(X_i), Y_i)$, $\widehat{R}_k$, and either U_k from equation (5) or a multiplicity-adjusted combination with equation (6).
4. Return the feasible candidate with greatest independently measured utility, as in equation (7); otherwise abstain from deployment.

Figure 2: The certification data are touched only in Step 3. Adding candidates after inspecting their certification losses invalidates the stated value of K.

The framework can support safer deployment by making uncertainty and abstention explicit. It can also create false reassurance if a single population-average certificate is used to dismiss subgroup harms or if lack of overlap is hidden by aggressive clipping. Any deployment report should therefore publish the target population definition, loss, split protocol, weight diagnostics, B, ε_w, candidate count, and failure probability. High-impact decisions still require domain review and ongoing monitoring.

8 Conclusion

Importance weighting identifies target risk under covariate shift, but deployment requires uncertainty around that estimate. A held-out sample and a bounded density-ratio estimate yield a simultaneous finite-sample certificate whose three terms have clear interpretations: measured weighted loss, sampling uncertainty, and ratio error. Because the event is uniform over a declared finite family, the certificate survives model selection. Its conservatism is useful information: failure to certify identifies whether the next effort should target more labels, better overlap, better ratio validation, or a decision not to deploy.

Acknowledgments

Omitted for anonymous review.

References

Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Risk-controlling prediction sets. *Journal of the Royal Statistical Society: Series B*, 86(4):973–996, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

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.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, 2009.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

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

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

A Additional Derivations

A.1 A bound with a random ratio-error guarantee

Assumption 2 states a deterministic error budget after conditioning on the fitted ratio estimator. Often the ratio procedure instead supplies

$$\mathbb{P}\big(\mathbb{E}_{P_X}|\widehat{w}(X) - w(X)| \le \varepsilon_w\big) \ge 1 - \delta_w. \tag{8}$$

If ratio fitting data and certification data are independent, combine equation (8) with theorem 1 using a union bound. With probability at least $1 - \delta_w - \delta_c$, simultaneously for all k,

$$\mathcal{R}_Q(f_k) \le \widehat{R}_k + \varepsilon_w + B\sqrt{\frac{\log(K/\delta_c)}{2n}}. \tag{9}$$

This form prevents the confidence used to validate the ratio from disappearing inside a nominally deterministic constant.

A.2 Bias introduced by clipping

Let $\widetilde{w}(x) = \min\{w(x), B\}$ and abbreviate $\ell_f = \ell(f(X), Y)$. Even with the true ratio available, clipping changes the estimand. Since $0 \le \ell_f \le 1$,

$$\mathcal{R}_Q(f) - \mathbb{E}_P[\widetilde{w}(X)\ell_f] = \mathbb{E}_P[(w(X) - B)_+\ell_f] \tag{10}$$

$$\le \mathbb{E}_P[(w(X) - B)_+]. \tag{11}$$

Thus the clipping tail $\mathbb{E}_P[(w - B)_+]$ is a valid contribution to ε_w. The expression also shows why reporting only the cap B is insufficient: the amount of target mass represented by clipped observations matters.

A.3 Prespecified groups

For groups $G_1, \ldots, G_m \subseteq \mathcal{X}$ with $Q_X(G_j) > 0$, conditional group risk can be written as a ratio,

$$\mathcal{R}_{Q,j}(f) = \frac{\mathbb{E}_P[w(X)\mathbf{1}\{X \in G_j\}\ell(f(X), Y)]}{\mathbb{E}_P[w(X)\mathbf{1}\{X \in G_j\}]}. \tag{12}$$

Certifying this quantity requires simultaneous control of both numerator and denominator; plugging a noisy denominator into equation (5) is not valid. A conservative construction uses an upper confidence bound for the numerator and a strictly positive lower confidence bound for the denominator, with failure probability allocated across all mK pairs. Small target group mass makes the denominator bound weak, correctly reflecting limited evidence.

B Reproducibility Checklist

The following information is sufficient to reproduce every numerical value in the main text and should accompany an empirical application.

- **Bound:** equation (5), with natural logarithms.
- **Stress-test constants:** $K = 20, \delta = .05, B = 2, \widehat{R}_k = .073$, and $n \in \{500, 2000, 8000\}$.
- **Rounding:** compute in full precision and round to three decimals only for display.
- **Figure constants:** the same K, δ, B, with $n \in \{250, 500, 1000, 2000, 4000, 8000\}$.
- **Randomness and compute:** none; the tables and figure are deterministic evaluations of a closed-form expression.
- **Empirical extension:** disclose data provenance, all three split sizes, candidate construction, ratio-estimation method, clipping diagnostics, the derivation of ε_w, and every selection made after viewing certification outcomes.

C Notation Reference

Table 3: Symbols used throughout the paper.

Symbol	Meaning
P, Q	source and target joint distributions
P_X, Q_X	source and target input marginals
w	target-to-source density ratio
$\widehat{w}$	independently fitted, clipped ratio
B	upper bound on $\widehat{w}$
ε_w	declared $L^1(P_X)$ ratio-error budget
S, n	certification sample and its size
$\mathcal{F}_K, K$	finite candidate family and its size
ℓ	user-declared loss in $[0, 1]$
$\mathcal{R}_Q$	target-population risk
$\widehat{R}_k$	estimated weighted risk of candidate k
U_k	simultaneous upper risk certificate
δ	total certification failure probability