High Probability Risk Control Under Covariate Shift

Duarte C. Almeida DUARTECALADOALMEIDA@TECNICO.ULISBOA.PT
Instituto Superior Técnico, Universidade de Lisboa
Instituto de Telecomunicações
Feedzai

João Bravo JOAO.BRAVO@FEEDZAI.COM
Feedzai

Jacopo Bono JACOPO.BONO@FEEDZAI.COM
Feedzai

Pedro Bizarro PEDRO.BIZARRO@FEEDZAI.COM
Feedzai

Mário A. T. Figueiredo MARIO.FIGUEIREDO@TECNICO.ULISBOA.PT
Instituto Superior Técnico, Universidade de Lisboa
Instituto de Telecomunicações

Editor: Khuong An Nguyen, Zhiyuan Luo, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

Distribution-free uncertainty quantification is an emerging field, which encompasses risk control techniques in finite sample settings with minimal distributional assumptions, making it suitable for high-stakes applications. In particular, high-probability risk control methods, namely the *learn then test* (LTT) framework, use a calibration set to control multiple risks with high confidence. However, these methods rely on the assumption that the calibration and target distributions are identical, which can pose challenges, for example, when controlling label-dependent risks under the absence of labeled target data. In this work, we propose a novel extension of LTT that handles covariate shifts by directly weighting calibration losses with importance weights. We validate our method on a synthetic fraud detection task, aiming to control the false positive rate while minimizing false negatives, and on an image classification task, to control the miscoverage of a set predictor while minimizing the average set size. The results show that our approach consistently yields less conservative risk control than existing baselines based on rejection sampling, which results in overall lower false negative rates and smaller prediction sets.

Keywords: High-probability risk control, covariate shift, learn then test, distribution-free uncertainty quantification, conformal prediction.

1. Introduction

Machine learning is increasingly used to automate high-risk and high-stakes decisions (*e.g.*, in healthcare or finance). These decisions are often associated with specific risks representing statistical measures of inaccuracy that must be controlled to meet safety, regulatory, quality, or other standards. For example, fraud detection systems should be properly tuned to detect fraudulent transactions while avoiding hindering legitimate economic activity.

Distribution-free uncertainty quantification is an emerging field that encompasses risk control techniques in finite sample settings with minimal distributional assumptions (Angelopoulos and Bates, 2023). A cornerstone method of this family is *conformal prediction*

(CP) (Shafer and Vovk, 2008), which calibrates a set predictor to control the miscoverage probability. Letting $\mathcal{X}$ and $\mathcal{Y}$ denote the feature and label spaces, respectively, CP uses a calibration set $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n} \subset \mathcal{X} \times \mathcal{Y}$ to generate prediction sets $\mathcal{C}(X_{n+1})$ for a new test data point (X_{n+1}, Y_{n+1}). Assuming only exchangeability of $\{(X_i, Y_i)\}_{i=1}^{n+1}$ (its joint distribution is permutation-invariant), CP provides a *coverage/validity* guarantee: $\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$. However, CP does not guarantee the stronger *conditional validity* property $\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} = x) \geq 1 - \alpha, \forall x \in \mathcal{X}$. While this property cannot be attained in general, relaxations such as group-conditional and class-conditional coverage offer guarantees within specific feature-label subsets (Barber et al., 2021; Gibbs et al., 2025) . Furthermore, *calibration-set validity* is always achievable, as concentration bounds for the calibration-set conditional coverage level exist (Vovk, 2012).

Under the same exchangeability assumption, CP can be generalized to control other risks beyond miscoverage, via the calibration of some (possibly multidimensional) parameter λ. If the risk can be expressed as the expectation of a lower-bounded loss function $L(\cdot, \cdot; \lambda) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which is coordinate-wise non-increasing with respect to λ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, *conformal risk control* (CRC) can be applied to compute $\hat{\lambda}$ that guarantees $\mathbb{E}[L(X_{n+1}, Y_{n+1}; \hat{\lambda})] \leq \alpha$ (Angelopoulos et al., 2024).

The probability and expectation in the aforementioned guarantees are taken over both the new test point and the calibration set. To rigorously control the risk, the calibration procedure must be applied each time a new test point is introduced. If calibration is only done periodically, there may be time frames where the risk exceeds the desired threshold due to an anomalous calibration sample, which could be problematic in scenarios where risk control should be as strict as possible. An alternative approach is to perform *high-probability risk control* (HPRC) (Angelopoulos et al., 2025; Bates et al., 2021), which bounds the probability of risk violations occurring due to a "bad" calibration set below some threshold δ. Methods for this purpose operate under slightly stronger distributional assumptions, namely, that the calibration set is i.i.d. according to the target distribution.

When models are deployed over data from a *target* distribution that may differ from that of the training data (the *source* distribution), a significant performance degradation may occur (Kouw and Loog, 2019). This is the case, *e.g.,* when a fraud detection system is applied to transactions from a new geographical setting, or if an image classifier trained on a particular type of image (*e.g.*, photographs) is used on other kinds of image (*e.g.*, drawings). Furthermore, the risk control techniques mentioned above cannot be applied over source calibration data, as distribution shifts violate important underlying assumptions (e.g., exchangeability and i.i.d.). While labeled target data are often unavailable, due to slow or costly labeling processes, unlabeled target data could still be used for risk control.

In this work, we introduce a novel adaptation of high-probability risk control methods, specifically of the *learn then test* (LTT) approach (Angelopoulos et al., 2025), that addresses covariate shift between the source and target distributions in the presence of unlabeled target data. Making use of recent advancements in hypothesis testing, we prove how calibration losses can be weighted by importance weights to achieve risk control. We experimentally validate our approach in two settings. The first is a synthetic fraud detection scenario in which a model trained on a source distribution is calibrated to control the false positive rate (FPR) on a target distribution while minimizing the false negative rate (FNR). The second setting involves an image classification task, where a model trained on

a source distribution is deployed on a target domain representing a specific data subpopulation. Here, the covariate shift assumption may not hold, and the goal is to calibrate a set predictor to control the miscoverage probability while minimizing the average set size, as considered by Park et al. (2022). Our results show that our approach is less conservative than the existing covariate shift-adapted HPRC baseline in both cases (i.e., smaller FNR and average set size), while achieving the desired risk control in the synthetic setting and closing the risk gap in the second scenario.

Related Work. This work is part of a broader research avenue aimed at adapting risk control methods to handle distribution shifts. Notably, Tibshirani et al. (2019) proposed a weighting scheme for quantile calculation in CP that uses importance weights to preserve the original coverage guarantee under covariate shift. Barber et al. (2023) examined the impact of both data-independent and dependent weighting schemes on the coverage gap, providing important theoretical guidelines for designing such schemes to bridge the gap under arbitrary violations of the exchangeability assumption. This approach has been extended to conformal risk control by Farinhas et al. (2024). In the context of high-probability risk control, Park et al. (2022) introduced rejection sampling to align the calibration and target distributions under covariate shift while aiming specifically for the control of the miscoverage risk. This approach was later extended to other use cases by Zollo et al. (2024); although different from our method, it will be fully explained in later sections for completeness.

2. Background: High-Probability Risk Control

In this section, we present a review of some foundational principles of HPRC. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the feature and label spaces, respectively, and let $\mathbb{P}$ be a probability measure on a suitable measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F})$. For convenience, we will also use the notation $\mathbb{P}$ to refer to the corresponding distribution. Furthermore, consider a calibration set $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, where $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, $\forall i \in \{1, \dots, n\}$. Let $L(\cdot, \cdot; \lambda) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable loss parameterized by λ, and define $R_{\mathbb{P}}(\lambda) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[L(X, Y; \lambda)]$ as the risk (expected loss) under $\mathbb{P}$. For conciseness, we include the subscript $\mathbb{P}$ in the expectation defining the risk only when the underlying probability measure is not clear from the context.

A core HPRC approach is *learn then test* (LTT) (Angelopoulos et al., 2025). Given a predefined subset Λ of the parameter space, LTT uses a calibration set to generate a subset $\hat{\Lambda} \subseteq \Lambda$ such that the risk is uniformly controlled over all its elements, i.e.,

$$\mathbb{P}\Big(\sup_{\lambda \in \hat{\Lambda}} R(\lambda) \leq \alpha\Big) \geq 1 - \delta,$$

for some confidence level $\delta \in (0, 1)$, where the randomness is over the calibration set. Once $\hat{\Lambda}$ is determined, a single parameter can be selected based on additional optimality criteria. For instance, in general detection (binary classification) systems, one might control the false positive rate and select the parameter in $\hat{\Lambda}$ that minimizes the false negative rate (or vice versa). Moreover, this approach does not require the risk or loss to be monotonic with respect to λ, which may even be any general mathematical object.

Hypothesis tests serve as the main building blocks of this procedure; specifically, for each value of $\lambda \in \Lambda$, the following null hypothesis is considered:

$$\mathcal{H}_0(\lambda, \alpha) : R(\lambda) > \alpha.$$

If a rejection rule over the calibration set bounds the probability of a false rejection below $1-\delta$, then rejecting the null hypothesis based on it provides $1-\delta$ confidence that λ controls the risk below α. Such procedures can be designed using valid p-values.

Definition 1 (p-value) *(Angelopoulos et al., 2025) Let $\mathcal{H}_0$ be a null hypothesis. A statistic p is said to be a valid p-value for $\mathcal{H}_0$ if it is super-uniform under $\mathcal{H}_0$, i.e.,*

$$\mathbb{P}_{\mathcal{H}_0}\big(p \leq \delta\big) \leq \delta, \ \forall \delta \in [0,1].$$

Thus, rejecting $\mathcal{H}_0$ if $p \leq \delta$ ensures that a false rejection occurs with probability at most δ.

The following theorem establishes how to generate p-values from concentration inequalities (such as Hoeffding's inequality).

Theorem 2 *(Bates et al., 2021) Let $g(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$ be a function satisfying the condition*

$$\mathbb{P}\big(\hat{R}(\lambda) \leq t\big) \leq g\big(t, R(\lambda)\big), \ \forall t \in \mathbb{R}, \tag{1}$$

where $\hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} L(X_i, Y_i; \lambda)$ represents the empirical risk over the calibration set and $\mathbb{E}[L(X_i, Y_i; \lambda)] = R(\lambda)$, $\forall i \in \{1, \ldots, n\}$. Then, $g\big(\hat{R}(\lambda), \alpha\big)$ is a valid p-value for the null hypothesis $\mathcal{H}_0(\lambda, \alpha) : R(\lambda) > \alpha$.

In some cases, the distribution of the empirical risk can be exactly specified, making it possible to replace potentially loose non-parametric bounds with an exact expression on the r.h.s. of Equation (1), leading to more powerful p-values. For instance, if the loss function is almost surely supported on $\{0,1\}$, then $n\hat{R}(\lambda)$ follows a binomial distribution. The 0-1 loss scenario is ubiquitous, arising when controlling the miscoverage or the FPR.

Theorem 3 (Clopper-Pearson p-value) *(Clopper and Pearson, 1934; Park et al., 2022) Let $L(\cdot,\cdot;\lambda) : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be a measurable loss function parametrized by λ. Let $\hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} L(X_i, Y_i; \lambda)$ be the empirical risk computed over an i.i.d. calibration set such that $\mathbb{E}[L(X_i, Y_i; \lambda)] = R(\lambda)$, $\forall i \in \{1, \ldots, n\}$. Then, the statistic*

$$p(\lambda, \alpha) = F_{\mathrm{Bin}(n,\alpha)}\big(n\hat{R}(\lambda)\big),$$

where $F_{\mathrm{Bin}(n,\alpha)}$ is the cumulative distribution function of the binomial distribution with n trials and success probability α, is a valid p-value for $\mathcal{H}_0(\lambda, \alpha) : R(\lambda) > \alpha$.

Obtaining $\hat{\Lambda}$ requires ensuring that all its elements control the risk below α with $1-\delta$ confidence, rather than being $(1-\delta)$-confident for each $\lambda \in \hat{\Lambda}$. In hypothesis testing terminology, this is referred to as controlling the *family-wise error rate* (FWER) (Angelopoulos et al., 2025) at level δ, which is defined as

$$\mathrm{FWER}\big(\hat{\Lambda}\big) = \mathbb{P}\big(\exists \lambda \in \hat{\Lambda} : \mathcal{H}_0(\lambda, \alpha) \text{ holds}\big).$$

Such approaches are called FWER-controlling and consist of strategies to aggregate p-values generated by testing the set of hypotheses $\{\mathcal{H}_0(\lambda, \alpha) : \lambda \in \Lambda\}$. If Λ is discrete, one possible approach is *fixed-sequence testing* (FST) (Angelopoulos et al., 2025). In FST, hypotheses

are tested in a predefined order, and all corresponding values of λ in $\hat{\Lambda}$ are gathered until the first non-rejection at level δ occurs. The risk-controlling subset is thus defined by $\hat{\Lambda} = \{\lambda'_i\}_{i=1}^{k^*-1}$, where

$$k^* = \min\{k : p(\lambda'_i, \alpha) \leq \delta,\ \forall i < k\},$$

for a predefined ordering $(\lambda'_1, ...\lambda'_{|\Lambda|})$ of Λ. For this method to be useful, safer values for λ should be tested first. When the monotonicity relationship between λ and $R(\lambda)$ is unclear, *split fixed-sequence testing* (SFST) (Angelopoulos et al., 2025; Laufer-Goldshtein et al., 2023) can be used. In this approach, the calibration set is divided into two disjoint subsets: one for defining an ordering of Λ and another for applying FST.

LTT can also be extended to simultaneously control multiple risks by considering a p-value for each individual risk and taking the maximum of all p-values. Additionally, when p-values are almost surely monotonically non-increasing or non-decreasing in λ for a fixed calibration set $\mathcal{D}$, it is possible to consider an interval $[\lambda_-, \lambda_+]$ as Λ. In this case, the iterative nature of FST/SFST can be avoided by directly computing the risk-controlling half-subset $\hat{\Lambda}$ using any standard root-finding algorithm (Bates et al., 2021).

3. High-Probability Risk Control under Covariate Shift

Consider the problem of asserting risk control over a target distribution $\mathbb{P}_{\text{target}}$, from which only an *unlabeled* sample $\mathcal{T} = \{X_i\}_{i=1}^{n_t}$ is available. Standard risk control methods generally cannot be applied in this setting if the loss function depends on the label Y, as it usually does. If we possess a *labeled* sample $\mathcal{S} = \{(X_i, Y_i)\}_{i=1}^{n_s}$ drawn from a possibly different source distribution $\mathbb{P}_{\text{source}}$, we should ask if the application of these methods on $\mathcal{S}$ yields the desired statistical confidence guarantees.

To test $\mathcal{H}_0(\lambda, \alpha) : R(\lambda) > \alpha$ at a significance level δ, a general strategy is to use p-values, which require that $\mathbb{E}[L(X_i, Y_i; \lambda)] = R(\lambda)$ for any data point (X_i, Y_i) in the calibration set. Therefore, without further assumptions, we are only guaranteed to control the source risk

$$R_{\mathbb{P}_{\text{source}}}(\lambda) = \mathbb{E}_{(X,Y)\sim\mathbb{P}_{\text{source}}}[L(X, Y; \lambda)].$$

For the risk to be controlled over the target distribution, the relation $R_{\mathbb{P}_{\text{target}}}(\lambda) \leq R_{\mathbb{P}_{\text{source}}}(\lambda)$ should hold, which is a strong assumption that may not be verified in general.

3.1. The Covariate Shift Assumption and Existing HPRC Methods

Fortunately, it is possible to make use of the source dataset $\mathcal{S}$ under some further assumptions on $\mathbb{P}_{\text{source}}$ and $\mathbb{P}_{\text{target}}$. In particular, the *covariate shift* assumption renders this problem tractable by assuming equal feature-conditional probability measures.

Definition 4 (Covariate shift) *(Quiñoreo-Candela et al., 2009) Two probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on some measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F})$ are said to differ by a covariate shift if the corresponding feature-conditional probability measures coincide, i.e.,*

$$d\mathbb{P}(y \mid x) = d\mathbb{Q}(y \mid x), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

Furthermore, under mild regularity conditions, we can define the notion of importance weight between two probability measures (also known as the Radon-Nikodym derivative).

Definition 5 (Importance weight) *(Resnick, 1999) Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures defined on a measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F})$ such that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ ($\mathbb{Q} \ll \mathbb{P}$), i.e.,*

$$\mathbb{P}(A) = 0 \Rightarrow \mathbb{Q}(A) = 0, \quad \forall A \in \mathcal{F}.$$

Define the importance weight function as

$$w(x, y) = \frac{d\mathbb{Q}}{d\mathbb{P}}(x, y), \quad \forall (x, y) \in supp(\mathbb{P}).$$

Then, it holds that

$$\mathbb{Q}(\mathcal{A}) = \int_{\mathcal{A}} w(x, y)\, d\mathbb{P}(x, y),\ \forall \mathcal{A} \in \mathcal{F}.$$

Under the covariate shift assumption, the importance weight function depends only on the marginal distributions over $\mathcal{X}$ (Yu and Szepesvári, 2012),

$$w(x, y) = \frac{d\mathbb{P}_{\text{target}}}{d\mathbb{P}_{\text{source}}}(x, y) = \frac{d\mathbb{P}_{\text{target},X}}{d\mathbb{P}_{\text{source},X}}(x) := w(x),$$

where $\mathbb{P}_{\text{source},X}$ and $\mathbb{P}_{\text{target},X}$ denote the induced source and target marginals over $\mathcal{X}$. Moreover, the importance weight can be estimated only from unlabeled source target data several well-established methods (You et al., 2019; Sugiyama et al., 2007; Huang et al., 2006).

To correct covariate shift between the source and target distributions, Park et al. (2022) propose aligning the calibration set with the target distribution via *rejection sampling* (Robert and Casella, 2004), a classical technique to generate samples from a *target* distribution using another (so-called *proposal*) distribution. The definition and correctness of this procedure are established by the following theorem.

Theorem 6 (Rejection Sampling) *(Robert and Casella, 2004) Let $\mathbb{P}$ and $\mathbb{Q}$ be the target and proposal, respectively, probability measures defined on a measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F})$ such that $\mathbb{Q} \ll \mathbb{P}$. Define $w = \frac{d\mathbb{Q}}{d\mathbb{P}}$ as the corresponding importance weight function, assumed to be bounded above by some constant $B < \infty$. Let $(X, Y) \sim \mathbb{P}$, and define (X', Y') as the pair (X, Y) accepted under the event $U \leq w(X, Y)/B$, where $U \sim Uniform[0, 1]$ is independent of (X, Y); that is, $(X', Y') = (X, Y) \mid w(X)/B \leq U$. Then, $(X', Y') \sim \mathbb{Q}$.*

Given an importance weight function $w : \mathcal{X} \to \mathbb{R}$ and an upper bound $B \geq \sup\{w(x),\ x \in \mathcal{X}\}$, both estimated from unlabeled source and target data, rejection sampling can be applied to the labeled source set $\mathcal{S}$ to produce a new i.i.d. sample $\mathcal{S}'$ from $\mathbb{P}_{\text{target}}$, assuming these estimates are accurate. This allows the application of standard risk control procedures as if $\mathcal{S}'$ was directly drawn from the target distribution.

Applying rejection sampling as described, the expected size of $\mathcal{S}'$ (the set of non-rejected samples) is inversely proportional to B, since

$$\mathbb{E}\left[\sum_{i=1}^{n_s} \mathbf{1}\left\{U_i \leq \frac{w(X_i)}{B}\right\}\right] = \frac{n_s}{B},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. If B is large, the retained sample may become too small, yielding overly conservative p-values that fail to identify useful risk-controlling values of λ. This can degrade performance with respect to complementary metrics; for instance, a low FPR may come at the cost of a higher false negative rate and vice versa.

3.2. HPRC under Covariate Shift via Importance Weighting

In this section, we propose a different approach to HPRC under covariate shift, addressing the limitations of rejection sampling. We discuss a new method to control risks over the joint distribution (Section 3.2.1) and extend it to conditional risks (Section 3.2.2).

3.2.1. Controlling Risks over the Joint Feature/Label Distribution

We first propose an alternative way to deal with covariate shift when controlling risks that are the expectation of a loss $L(X, Y; \lambda)$ over the joint feature and label distribution induced by $\mathbb{P}_{\text{target}}$ (e.g., miscoverage). In such cases, these can be written as expectations of the importance-weighted loss $w(X)L(X, Y; \lambda)$ over $\mathbb{P}_{\text{source}}$ (Quiñoreo-Candela et al., 2009):

$$R(\lambda) = \mathbb{E}_{(X,Y)\sim\mathbb{P}_{\text{target}}}[L(X, Y; \lambda)] = \mathbb{E}_{(X,Y)\sim\mathbb{P}_{\text{source}}}[w(X)L(X, Y; \lambda)]. \tag{2}$$

In this way, we can perform risk control using the *entirety* of the labeled source data by considering the sample of importance-weighted losses $\{w(X_i)L(X_i, Y_i; \lambda)\}_{i=1}^{n_s}$, where $(X_i, Y_i) \sim \mathbb{P}_{\text{source}}$, $\forall i \in [n_s]$. We formalize this approach in the next theorem.

Theorem 7 *Let $\mathbb{P}_{source}$ and $\mathbb{P}_{target}$ be probability measures on $(\mathcal{X}\times\mathcal{Y}, \mathcal{F})$ such that $\mathbb{P}_{target} \ll \mathbb{P}_{source}$, and assume they differ by a covariate shift. Let $w = \frac{d\mathbb{P}_{target}}{d\mathbb{P}_{source}}$ be the corresponding importance weight function, and let $L(\cdot, \cdot; \lambda) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable loss parametrized by λ. Then, any p-value for $\mathcal{H}_0'(\lambda, \alpha) : \mathbb{E}_{\mathbb{P}_{source}}[w(X)L(X, Y; \lambda)] > \alpha$ is also a p-value for $\mathcal{H}_0(\lambda, \alpha) : \mathbb{E}_{\mathbb{P}_{target}}[L(X, Y; \lambda)] > \alpha$.*

Proof By Equation (2), the conditions in $\mathcal{H}_0(\lambda, \alpha)$ and $\mathcal{H}_0'(\lambda, \alpha)$ are equivalent. Thus, for any valid p-value $p(\lambda, \alpha)$ for $\mathcal{H}_0'(\lambda, \alpha)$, we have $\mathbb{P}_{\mathcal{H}_0(\lambda,\alpha)}(p(\lambda, \alpha) \leq \delta) = \mathbb{P}_{\mathcal{H}_0'(\lambda,\alpha)}(p(\lambda, \alpha) \leq \delta) \leq \delta$, $\forall \delta \in [0, 1]$. ■

It is important to recognize that, while this method retains all the source data, the distribution of the weighted losses may be more challenging to handle. For instance, the introduction of weights can destroy desirable properties of the original loss that allow the use of very tight p-values—e.g., it may break the 0–1 structure and significantly broaden the range of possible values. Nevertheless, recent advances in testing by betting provide highly variance-adaptive p-values that work very well in practice, such as the Waudby-Smith–Ramdas (WSR) p-value (Waudby-Smith and Ramdas, 2024).

Theorem 8 (WSR p-value) *(Bates et al., 2021; Waudby-Smith and Ramdas, 2024) Let $L(\cdot, \cdot; \lambda) : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be a measurable loss parametrized by λ and let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}$ be a calibration set such that*

$$\mathbb{E}[L(X_i, Y_i; \lambda)] = \mathbb{E}[L(X_i, Y_i; \lambda) \mid L(X_{i-1}, Y_{i-1}; \lambda), \ldots, L(X_1, Y_1; \lambda)] = R(\lambda),\ \forall i \in \{1, \ldots, n\}.$$

Define the following statistics:

$$\hat{\mu}_i(\lambda) = \frac{1/2 + \sum_{j=1}^{i} L(X_j, Y_j; \lambda)}{1 + i}, \quad \hat{\sigma}_i^2(\lambda) = \frac{1/4 + \sum_{j=1}^{i}(L(X_j, Y_j; \lambda) - \hat{\mu}_j(\lambda))^2}{1 + i},$$

$$\nu_i(\lambda) = \min\left\{1, \sqrt{\frac{2\log(1/\delta)}{n\,\hat{\sigma}_{i-1}^2(\lambda)}}\right\}.$$

Furthermore, define the capital process $\{\mathcal{K}_i(\lambda,\alpha)\}_{i=1}^n$ *as*

$$\mathcal{K}_i(\lambda,\alpha) = \prod_{j=1}^{i}\big(1-\nu_j(\lambda)(L(X_j,Y_j;\lambda)-\alpha)\big).$$

Then, the statistic

$$p(\lambda,\alpha) = \left(\max_{i\in\{1,\dots,n\}} \mathcal{K}_i(\lambda,\alpha)\right)^{-1}$$

is a p-value for $\mathcal{H}_0(\lambda,\alpha) : R(\lambda) > \alpha$.

Although this p-value assumes that the loss lies between 0 and 1, it can be readily extended to our setting. If L is supported on $[l,u]$, then the importance-weighted loss wL is supported on $[l',u']$, where $l' = \min(0,Bl)$ and $u' = \max(0,Bu)$. It is then sufficient to test the equivalent hypothesis

$$\mathcal{H}_0(\lambda,\alpha) : \frac{R(\lambda)-l'}{u'-l'} > \frac{\alpha-l'}{u'-l'},$$

using the rescaled importance-weighted losses $\left\{\frac{w(X_i)L(X_i,Y_i;\lambda)-l'}{u'-l'}\right\}_{i=1}^{n_s}$.

Due to the sequential nature of the procedure, the order of the calibration data can affect the outcome when the source dataset is a mixture of subsources (e.g., different types of images). If such subsources are processed in bulk and the early part of the capital process is computed over samples from a subsource for which the risk is controlled, the capital process may surpass δ^{-1} prematurely, leading to a rejection even though the risk is *not* controlled over the entire source distribution. To solve this, the data can be randomized B times, yielding B capital processes $\left\{\{\mathcal{K}_i^{(b)}(\lambda,\alpha)\}_{i=1}^n\right\}_{b=1}^B$, which can be averaged into a new process $\mathcal{K}_i(\lambda,\alpha) := B^{-1}\sum_{b=1}^B \mathcal{K}_i^{(b)}(\lambda,\alpha)$ (Waudby-Smith and Ramdas, 2024).

3.2.2. Controlling Conditional Risks

Many risks are defined over distributions other than the joint feature and label distribution. For example, the FPR of a binary classifier f, $\text{FPR} = \mathbb{P}(f(X)=1 \mid Y=0) = \mathbb{E}[\mathbf{1}\{f(X)=1\} \mid Y=0]$, is evaluated over the conditional distribution of X given $Y=0$. Under covariate shift between $\mathbb{P}_{\text{source}}$ and $\mathbb{P}_{\text{target}}$, it follows that

$$d\mathbb{P}_{\text{target},\,X\,|\,y}(x \mid y) = w(x)\,\frac{d\mathbb{P}_{\text{source},\,Y}}{d\mathbb{P}_{\text{target},\,Y}}(y)\; d\mathbb{P}_{\text{source},\,X\,|\,y}(x \mid y),$$

showing that covariate shift does not hold if the source and target class priors are different. In general, this reweighting procedure does not directly apply to risks defined over distributions other than the joint. However, it can be adapted for a broad class of risks, as shown in the following theorem.

Theorem 9 *Consider the null hypothesis* $\mathcal{H}_0(\lambda,\alpha) : \mathbb{E}_{\mathbb{P}_{target}}[L(X,Y;\lambda) \,|\, (X,Y)\in\mathcal{A}(\lambda)] > \alpha$, *where* $\mathcal{A}(\lambda)\in\mathcal{X}\times\mathcal{Y}$ *is a non-zero probability set and* $L(\cdot,\cdot;\lambda) : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ *is a measurable loss parametrized by* λ. *Then, any p-value for* $\mathcal{H}_0'(\lambda,\alpha) : \mathbb{E}_{\mathbb{P}_{source}}[L'(X,Y;\lambda)] > \alpha$ *is also a p-value for* $\mathcal{H}_0(\lambda,\alpha)$, *where*

$$L'(X,Y;\lambda) = w(X)\Big(L(X,Y;\lambda)\mathbf{1}\{(X,Y)\in\mathcal{A}(\lambda)\} + \alpha\mathbf{1}\{(X,Y)\notin\mathcal{A}(\lambda)\}\Big).$$

Proof Developing the condition in $\mathcal{H}_0(\lambda, \alpha)$, we have:

$$\begin{aligned}
& \mathbb{E}_{\mathbb{P}_{\text{target}}}[L(X,Y;\lambda) \mid (X,Y) \in \mathcal{A}(\lambda)] > \alpha \\
\Leftrightarrow \quad & \mathbb{E}_{\mathbb{P}_{\text{target}}}[L(X,Y;\lambda)\mathbf{1}\{(X,Y) \in \mathcal{A}(\lambda)\}] > \alpha \mathbb{P}_{\text{target}}((X,Y) \in \mathcal{A}(\lambda)) \\
\Leftrightarrow \quad & \mathbb{E}_{\mathbb{P}_{\text{target}}}[L(X,Y;\lambda)\mathbf{1}\{(X,Y) \in \mathcal{A}(\lambda)\}] > \alpha(1 - \mathbb{P}_{\text{target}}((X,Y) \notin \mathcal{A}(\lambda))) \\
\Leftrightarrow \quad & \mathbb{E}_{\mathbb{P}_{\text{target}}}[L(X,Y;\lambda)\mathbf{1}\{(X,Y) \in \mathcal{A}(\lambda)\}] + \alpha \mathbb{P}_{\text{target}}((X,Y) \notin \mathcal{A}(\lambda)) > \alpha \\
\Leftrightarrow \quad & \mathbb{E}_{\mathbb{P}_{\text{target}}}[L(X,Y;\lambda)\mathbf{1}\{(X,Y) \in \mathcal{A}(\lambda)\} + \alpha \mathbf{1}\{(X,Y) \notin \mathcal{A}(\lambda)\}] > \alpha.
\end{aligned}$$

The result then follows from applying Theorem 7 to the loss $L(X,Y;\lambda)\mathbf{1}\{(X,Y) \in \mathcal{A}(\lambda)\} + \alpha \mathbf{1}\{(X,Y) \notin \mathcal{A}(\lambda)\}$. ■

This result shows that any conditional risk can be reformulated as a risk over the joint feature-label distribution for the purpose of hypothesis testing. In the setting of a binary classifier $f(\cdot;\lambda) : \mathcal{X} \to \{0,1\}$, many common classification risks fall within this family. For instance, the FPR corresponds to taking $L(X,Y;\lambda) = \mathbf{1}\{f(X;\lambda) = 1\}$ and $\mathcal{A}(\lambda) = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : y = 0\}$. Similarly, for the false discovery rate, we have $L(X,Y;\lambda) = \mathbf{1}\{Y = 0\}$ and $\mathcal{A}(\lambda) = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : f(x;\lambda) = 1\}$.

4. Results and Discussion

4.1. Experimental Setup

4.1.1. Tasks and Datasets

We evaluate our method on two tasks. The first consists in controlling the FPR of a transaction fraud detection model below $\alpha = 0.05$ with confidence $(1 - \delta) = 0.95$. The model has the form $f(x;\lambda) = \mathbf{1}\{s(x) > \lambda\}$, where $s : \mathcal{X} \to [0,1]$ is a trained score function that captures the likelihood of a transaction being fraudulent. We apply *fixed sequence testing* (FST) over a set $\Lambda = \{s_{(i/1000)}\}_{i=1}^{1000}$, where $s_{(q)}$ denotes the sample q-quantile of the score distribution on source data. We test Λ in decreasing order and choose the smallest risk-controlling λ to minimize the FNR.

In this first task, we simulate a situation in which new unlabeled transaction data becomes available and previously collected labeled data are used to ensure risk control. For that purpose, we partition the Bank Account Fraud (BAF) dataset (Jesus et al., 2022) into three domains based on the credit risk of the client performing the transaction (`low`, `medium`, and `high`). Each domain is treated as the target in turn, with the remaining two combined to form the source. All domains are designed to differ solely in terms of covariate shift. Details on the partitioning procedure are provided in Appendix A. Furthermore, we use 70% of each domain's data for model training and the remaining 30% for risk control. We use LightGBM classifiers (Ke et al., 2017), trained with 5-fold cross-validation, and tune hyperparameters to maximize the AUROC using Optuna's implementation of TPE (Watanabe, 2023; Akiba et al., 2019). The hyperparameter grid is specified in Appendix F.

In the second task, we control the miscoverage of an image set predictor of the form $f(x;\lambda) = \{y \in \mathcal{Y} : s(x,y) > \lambda\}$ below 0.10 with confidence 0.95, where $s : \mathcal{X} \times \mathcal{Y} \to [0,1]$ is a trained score function such that $s(x,y)$ estimates the posterior probability of class y for an image x. We perform FST over $\Lambda = \{s^*_{(i/1000)}\}_{i=1}^{1000}$, where $s^*_{(q)}$ denotes the empirical

q-quantile of the true-class scores on the source data. Here, Λ is tested in increasing order and the largest risk-controlling $\lambda \in \Lambda$ is selected to minimize the average set size.

We use the DomainNet dataset (Peng et al., 2019), which consists of 6 domains: `clipart`, `real`, `infograph`, `painting`, `sketch`, and `quickdraw`. We consider each domain as target and *all* domains as sources, simulating a setting where a model is trained on broad data but is deployed on some (potentially unknown) subpopulation. We use the same ResNet model as Park et al. (2022) and use the original test and validation splits for risk control.

4.1.2. Estimation of Importance Weights

To estimate importance weights, we apply *kernel mean matching* (KMM), which minimizes the *maximum mean discrepancy* (MMD) between empirical kernel mean embeddings of the target distribution and the reweighted source distribution (Quiñoreo-Candela et al., 2009). Given a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ associated with a universal kernel k, KMM estimates the weights at the source datapoints $\{w(x_j) : x_j \in \mathcal{S}\}$ that minimize

$$\left\| \frac{1}{n_t} \sum_{x_i \in \mathcal{T}} k(x_i, \cdot) - \frac{1}{n_s} \sum_{x_j \in \mathcal{S}} w(x_j)\, k(x_j, \cdot) \right\|_{\mathcal{H}}^2$$

subject to $w(x_j) \geq 0, \forall x_j \in \mathcal{S}$, and $|\frac{1}{n_s} \sum_{x_j \in \mathcal{S}} w(x_j) - 1| \leq \epsilon$. Following Huang et al. (2006), we set $\epsilon = 1 - 1/\sqrt{n_s}$.

We consider a Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$, using one-hot encodings for categorical variables when computing distances. We set σ to the median of pairwise distances between all the points in the source and target datasets, following Sugiyama et al. (2009). Additionally, the maximum admissible importance weight is set at 10,000. To accelerate this procedure, we use the *very fast* KMM (VFKMM) algorithm proposed by Chandra et al. (2016), averaging importance weights computed across bootstrap samples of size 1000 from the source dataset, with the number of bootstrap samples set to ensure that each point is sampled at least once with probability 0.9999. Moreover, we consider the maximum estimated importance weight as an estimate for the upper bound B.

The second image classification task poses challenges due to the high dimensionality of the input, affecting the stability of importance weight estimates and increasing runtime. These challenges can be attenuated by considering a lower-dimensional feature transformation $h : \mathbb{R}^n \to \mathbb{R}^d$ with $d \ll n$, such that X and Y are conditionally independent given $h(X)$. One such transformation is $h(x) = (\mathbb{P}(Y = y_1 \mid X = x), \ldots, \mathbb{P}(Y = y_d \mid X = x))$, where $\mathcal{Y} = \{y_1, \ldots, y_d\}$ is the label space (Stojanov et al., 2019). This motivates our choice to use the model's predicted class scores as features for importance weight computation.

Park et al. (2022) propose a method to account for uncertainty in the estimation of importance weights. In short, the observations are binned into K equal-mass bins $\{B_i\}_{i=1}^K$ according to an estimate of the importance weights. Then, a δ-upper confidence bound

$$w^+(x) = \frac{\mathbb{P}^+_{\text{target}}(X \in B(x)) + E}{\left(\mathbb{P}^-_{\text{source}}(X \in B(x)) - E\right)_+}$$

can be obtained for $w(x)$, where $B(x)$ is the bin containing x, $\mathbb{P}^+$ and $\mathbb{P}^-$ denote $\delta/2$ Clopper-Pearson upper and lower-confidence bounds, respectively, and E is a predefined

smoothness constant to account for the histogram approximation. However, there are no theoretical guidelines for choosing K and E. Thus, we opt for more established procedures and proceed under the assumption that the estimates are accurate.

4.1.3. Baselines and Evaluation Procedure

We compare our proposed importance-weighted LTT method based on the WSR p-value (Waudby-Smith and Ramdas, 2024) (LTT-IW) against two baselines: LTT with rejection sampling and the Clopper-Pearson p-value (Clopper and Pearson, 1934) (LTT-RS), and LTT without importance weights (LTT) (i.e., directly controlling the source risk). Variations of our method for different p-values can be found in Appendix B.

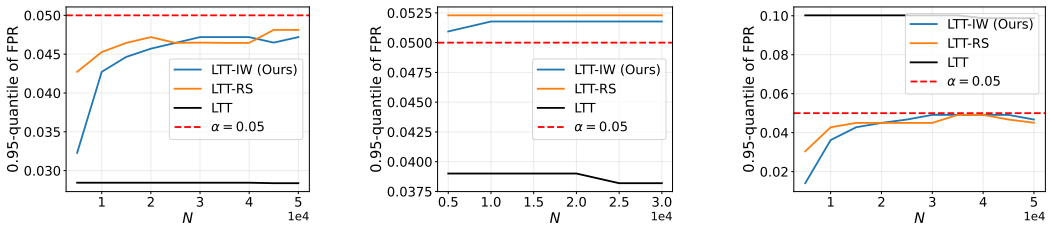
At the beginning of the procedure, we estimate the importance weights using the source and target splits allocated for risk control. We then run 1,000 iterations of LTT, each time over a different subsample of source data drawn without replacement. In the first task, we use the entire target dataset to evaluate the resulting target risk, while in the second baseline (where the target domain is contained in the source domain), we only use the part that was not sampled. To evaluate sample efficiency, we vary the source sample size, but always use the full source and target datasets to get the most accurate importance weight estimates. Finally, we estimate $(1-\delta)$-quantiles of the risk estimates computed over all iterations to assess if risk control is achieved. To evaluate how conservative the methods are, we also report the mean FNR and average set size obtained over all runs. In addition, Appendix C contains a brief analysis of the computational cost of the proposed method.

4.2. Results

Figure 1 reports the 0.95-quantile of the FPR for different values of the source sample size N. Ignoring covariate shift (LTT) can either make risk control overly conservative (for target domains `low` and `medium`), or outright invalid, as is the case of target domain `high`, where the risk is controlled at twice the intended level. Both weighted variants (LTT-IW and LTT-RS) control the FPR in nearly all cases, with the exception being in `medium`, where a residual risk gap of ≈ 0.0025 likely stems from errors in the estimated importance weights. Figure 2 further indicates that LTT-IW consistently achieves a lower average FNR than LTT-RS, showing that our method of direct importance weighting yields a less conservative and more stable risk control than rejection sampling in this case.

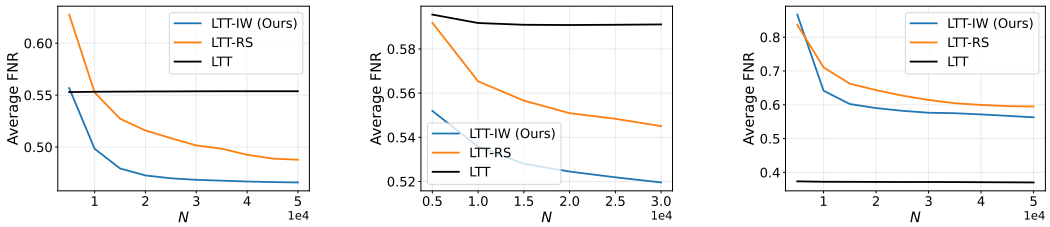
We perform a similar analysis for the second task. Figure 3 shows that the risk is only controlled when `real` or `quickdraw` serve as the target, indicating that the covariate-shift assumption does not hold. In fact, we are actually controlling the risk over a distribution $\mathbb{P}_{\text{aligned}}$ that matches the target feature marginal, but maintains the source feature-conditional distribution of the labels. With high probability, the true target risk can exceed α by at most the total-variation distance $d_{\text{TV}}\big(\mathbb{P}_{\text{target}}, \mathbb{P}_{\text{aligned}}\big)$ (Angelopoulos et al., 2024).

Nevertheless, we see that using importance weights (LTT-IW and LTT-RS) can bring the effective risk level closer to α: for the `clipart`, `infograph`, `painting`, and `sketch` cases, it bridges the risk gap, while making the procedure less conservative for `quickdraw`. In the `real` domain, however, the effect is minimal for LTT-IW, and LTT-RS deviates further from the target level. While both methods perform comparably on miscoverage, our method (LTT-IW) yields smaller average set sizes compared to LTT-RS (Figure 4). We



(*a*) `low` (*b*) `medium` (*c*) `high`

Figure 1: 0.95 FPR quantiles (vertical axis) vs. source sample size (horizontal axis) for each target domain in BAF (`low`, `medium` and `high` panels). Ignoring covariate shift (LTT) results in overly conservative or overly invalid risk control.



(*a*) `low` (*b*) `medium` (*c*) `high`

Figure 2: Average FNR (vertical axis) vs. source sample size (horizontal axis) for each target domain in BAF (`low`, `medium` and `high` panels). LTT-IW consistently results in lower FNR compared to LTT-RS.

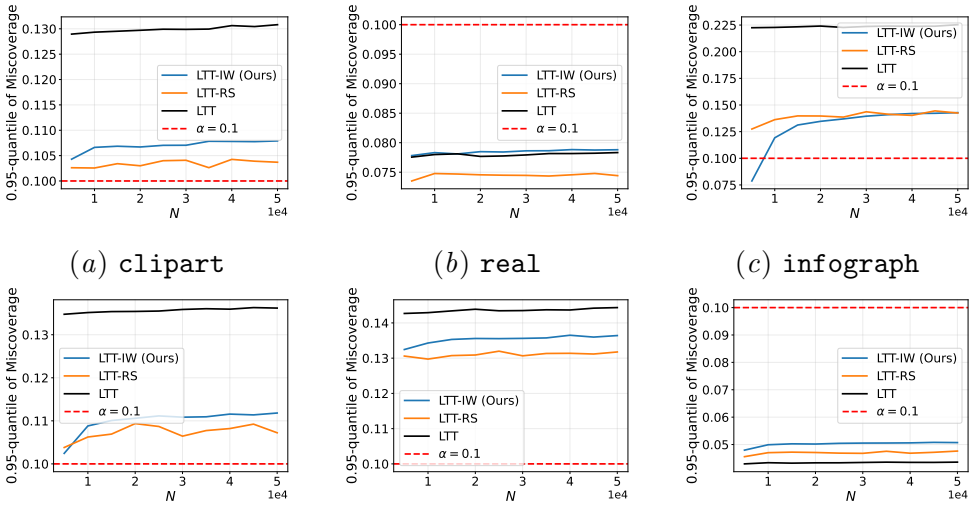
notice that, on real-world data where the covariate shift assumption may be violated, higher sample efficiency makes the procedure less conservative, but may exacerbate risk violations.

5. Conclusion and Future Work

In this work, we showed that using importance-weighted losses is a viable approach to tackle high-probability risk control under covariate shift. The experimental results show that our method outperforms the rejection-sampling baseline in terms of auxiliary performance measures. Although caveats remain, as these approaches rely on the covariate-shift assumption, the results show that the use of importance weights can narrow the risk gap, bringing the risk closer to the prescribed level even if the covariate-shift assumption is violated.

While the results presented show LTT-IW to be less conservative than LTT-RS in general, there may be individual cases where the opposite happens. An interesting research avenue consists of automatically selecting the better method. Further work could also extend the LTT-IW framework to support other risk functionals and develop strategies to increase p-value power (*e.g.*, via variance-reduction techniques). Appendix D discusses some limitations of one such approach.

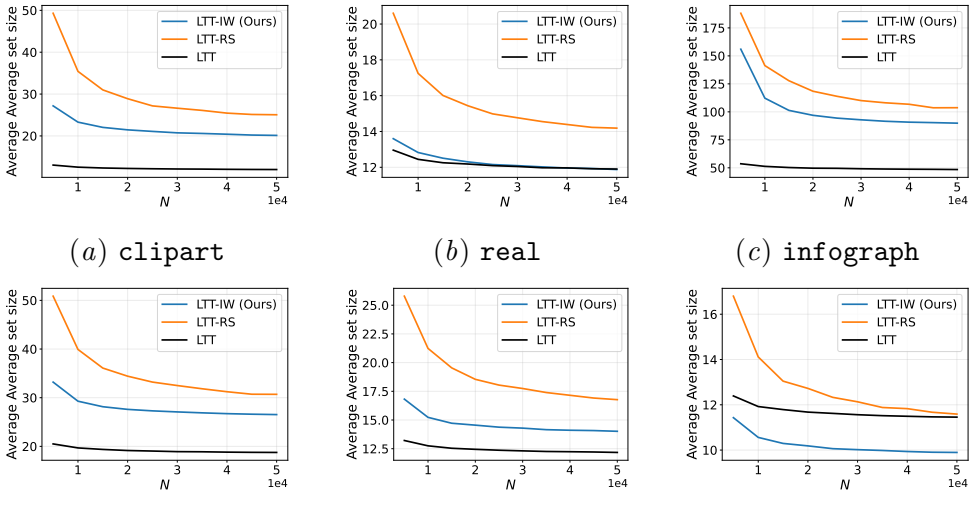
Both methods herein considered assume accurate importance weight estimates. The upper-confidence bounds provided by Park et al. (2022) address this uncertainty but assume the knowledge of hyperparameters for which there are no tuning guidelines, as well as



(a) `clipart` (b) `real` (c) `infograph`

(d) `painting` (e) `sketch` (f) `quickdraw`

Figure 3: 0.90 miscoverage quantiles (vertical axis) vs. source sample size (horizontal axis) for each target domain in DomainNet. The use of HPRC methods bridges the risk gap for `clipart`, `infograph`, `painting` and `sketch`.



(a) `clipart` (b) `real` (c) `infograph`

(d) `painting` (e) `sketch` (f) `quickdraw`

Figure 4: Mean average set size (vertical axis) vs. source sample size (horizontal axis) for each target domain in DomainNet. LTT-IW achieves smaller average set sizes than LTT-RS, but with increased risk violations due to concept shift.

Lipschitz-continuous densities, which may not be appropriate for dealing with mixed categorical and numerical feature spaces. Future work could relax this assumption or develop alternatives to account for uncertainty. Furthermore, these methods assume a known upper bound B on the importance weights. Appendix E presents a sensitivity analysis on B,

along with a discussion of alternative self-normalized approaches. The design of bound-free methods is also a promising direction for future research.

References

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623—-2631, 2019.

A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends Machine Learning*, 16(4):494—-591, 2023.

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. In *The 12th International Conference on Learning Representations (ICLR)*, 2024.

A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 19 (2):1641–1662, 2025.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10 (2):455–482, 2021.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–43, 2021.

S. Chandra, A. Haque, L. Khan, and C. Aggarwal. Efficient sampling-based kernel mean matching. In *IEEE Interna. Conference on Data Mining (ICDM)*, pages 811–816, 2016.

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

A. Farinhas, C. Zerva, D. Ulmer, and A. F. T. Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

I. Gibbs, J. J. Cherian, and E. J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2025.

P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2003.

J. Huang, A. Gretton, Ka. Borgwardt, B. Schölkopf, and A. Smola. Correcting sample selection bias by unlabeled data. In *Neural Information Processing Systems*, pages 601–608. MIT Press, 2006.

S. Jesus, J. Pombal, D. Alves, A. Cruz, P. Saleiro, R. P. Ribeiro, J. Gama, and P. Bizarro. Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation. In *Neural Information Processing Systems*, 2022.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Neural Information Processing Systems*, volume 30. Curran Associates, 2017.

W. Kouw and M. Loog. An introduction to domain adaptation and transfer learning, 2019. URL https://arxiv.org/abs/1812.11806.

I. Kuzborskij and C. Szepesvári. Efron-Stein PAC-Bayesian inequalities, 2020. URL https://arxiv.org/abs/1909.01931.

B. Laufer-Goldshtein, A. Fisch, R. Barzilay, and T. S. Jaakkola. Efficiently controlling multiple risks with Pareto testing. In *Intern. Conf. on Learning Representations*, 2023.

S. Park, E. Dobriban, I. Lee, and O. Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022.

X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *International Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019.

J. Quiñoreo-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

S. Resnick. *A Probability Path*. Birkhäuser, 1999.

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 2nd edition, 2004.

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

P. Stojanov, M. Gong, J. Carbonell, and K. Zhang. Low-dimensional density datio dstimation for covariate shift correction. In *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3449–3458, 2019.

M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Neural Information Processing Systems*, 2007.

M. Sugiyama, T. Kanamori, T. Suzuki, S. Hido, J. Sese, I. Takeuchi, and L. Wang. A density-ratio framework for statistical data processing. *Information and Media Technologies*, 4(4):962–987, 2009.

R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas. Conformal prediction under covariate shift. In *Neural Information Processing Systems*, 2019.

V. Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490, 2012.

S. Watanabe. Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance, 2023. URL https://arxiv.org/abs/2304.11127.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1): 1–27, 2024.

K. You, X. Wang, M. Long, and M. Jordan. Towards accurate model selection in deep unsupervised domain adaptation. In *36th International Conference on Machine Learning (ICML)*, pages 7124–7133, 2019.

Y. Yu and C. Szepesvári. Analysis of kernel mean matching under covariate shift. In *29th International Coference on Machine Learning (ICML)*, pages 1147–1154, 2012.

T. P. Zollo, T. Morrill, Z. Deng, J. Snell, T. Pitassi, and R. Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models. In *International Conference on Learning Representations*, 2024.

Appendix A. Dataset Generation

In this appendix, we provide details on the generation of domains from the BAF dataset. During preprocessing, missing numerical values are imputed with the mean and standardized using z-score normalization, while categorical features are imputed with the mode. To eliminate any temporal drift, the `month` feature is excluded.

To generate the disjoint covariate-shifted datasets, we employ a strategy similar to that of Huang et al. (2006). Let $c(x)$ be the value of the feature `credit_risk_score` of sample x, and let $q_{\text{low}}, q_{\text{med}}, q_{\text{high}}$ be the empirical 0.25, 0.50, 0.75 quantiles of $c(x)$, respectively. Sampling proceeds in stages: each observation is included in the low-risk domain with probability $p_{\text{low}}(x) = \exp\left(-\frac{(c(x)-q_{\text{low}})^2}{2\sigma_{\text{low}}^2}\right)$; if not selected, it enters the medium-risk domain with probability $p_{\text{med}}(x)$ defined analogously using q_{med} and σ_{med}; any remaining sample is assigned to the high-risk with probability $p_{\text{high}}(x)$, defined similarly. The bandwidths $\{\sigma_{\text{low}}, \sigma_{\text{med}}, \sigma_{\text{high}}\}$ are optimized to maximize the largest importance weight between any pair of domains while ensuring the expected dataset sizes lie in $[5 \times 10^4, 1.5 \times 10^5]$.

Since the sampling procedure is label-independent, the resulting domains differ solely by covariate shift. Let S_i be the indicator that a sample is assigned to domain i, and let $\mathbb{P}_i$, $\mathbb{P}_j$ be the corresponding distributions. Then, the importance weight is

$$\frac{d\mathbb{P}_i}{d\mathbb{P}_j}(x, y) = \frac{\mathbb{P}(S_i = 1 \mid x)}{\mathbb{P}(S_j = 1 \mid x)} \frac{\mathbb{P}(S_j = 1)}{\mathbb{P}(S_i = 1)},$$

which is a function of x alone, confirming the covariate shift assumption. Under the sequential scheme, we have $\mathbb{P}(S_{\text{low}} = 1 \mid x) = p_{\text{low}}(x)$, $\mathbb{P}(S_{\text{med}} = 1 \mid x) = (1 - p_{\text{low}}(x))\, p_{\text{med}}(x)$, and $\mathbb{P}(S_{\text{high}} = 1 \mid x) = (1 - p_{\text{low}}(x))\,(1 - p_{\text{med}}(x))\, p_{\text{high}}(x)$. Marginal selection probabilities can be estimated via Monte Carlo as $\mathbb{P}(S_k = 1) \approx n^{-1} \sum_{i=1}^{n} \mathbb{P}(S_k = 1 \mid x_i)$, for $k \in \{\text{low}, \text{med}, \text{high}\}$, and used to estimate dataset sizes.

Appendix B. P-value ablations

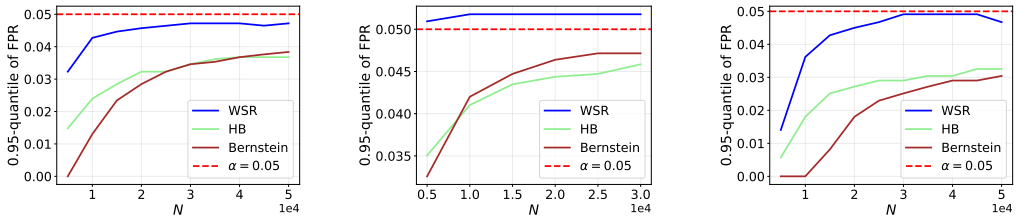
In this appendix, we compare the use of the WSR p-value against other possible p-values. We consider the Hoeffding-Bentkus p-value for testing $\mathcal{H}_0(\lambda, \alpha)$ (Angelopoulos et al., 2025). Since this approach only works for losses supported in $[0, 1]$, we consider the rescaled hypothesis $\mathcal{H}_0(\lambda, \alpha) : \frac{R(\lambda)-l'}{u'-l'} > \frac{\alpha-l'}{u'-l'}$ (see Section 3), yielding the following p-value:

$$p(\lambda, \alpha) = \min\left(\exp\left\{-nh\left(\frac{\hat{R}_w(\lambda) - l'}{u' - l'}, \frac{\alpha - l'}{u' - l'}\right)\right\}, eF_{\text{Bin}\left(n, \frac{\alpha-l'}{u'-l'}\right)}\left(\left\lceil n\frac{\hat{R}_w(\lambda) - l'}{u' - l'}\right\rceil\right)\right),$$

where $R_w(\lambda) = n_s^{-1}\sum_{i=1}^{n_s} w(X_i)L(X_i, Y_i; \lambda)$ and $h(a, b) = a\log(a/b)+(1-a)\log((1-a)/(1-b))$. We also consider the application of Bernstein's inequality for this hypothesis test (Bates et al., 2021). In particular, we reject $\mathcal{H}_0(\lambda, \alpha)$ if

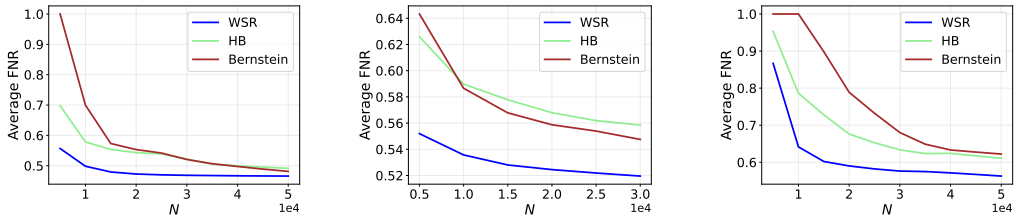
$$\frac{\hat{R}_w(\lambda) - l'}{u' - l'} + \frac{\hat{\sigma}_w(\lambda)}{u' - l'}\sqrt{\frac{2\log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)} \leq \frac{\alpha - l'}{u' - l'},$$

where $\hat{\sigma}_w(\lambda) = \sqrt{(n_s - 1)^{-1}\sum_{i=1}^{n_s}(L(X_i, Y_i; \lambda) - \hat{R}_w(\lambda))^2}$ denotes the empirical importance-weighted risk standard deviation. For brevity, we report only the BAF results, where the covariate shift assumption is guaranteed to hold. Figures 5 and 6 replicate the analysis of Section 4. Results show that the WSR p-value is the least conservative of the three, bringing the FPR the closest to α and achieving the lowest average FNR.



(a) `low` (b) `medium` (c) `high`

Figure 5: 0.95 FPR quantiles (vertical axis) vs. source sample size (horizontal axis) for each target domain in BAF. WSR yields 0.95 FPR quantiles closest to 0.05.



(a) `low` (b) `medium` (c) `high`

Figure 6: Mean FNR (vertical axis) vs. source sample size (horizontal axis) for each target domain in BAF. Out of the three p-values, WSR achieves the lowest FNR values.

Appendix C. Computational Considerations

We now conduct a brief analysis of the computational cost of LTT-IW. All experiments were run on a machine with a 14-core CPU and 20-core GPU Apple M4 chip, 24GB of RAM, and a 512GB SSD. We first analyze KMM, which dominates runtime due to the need to solve multiple quadratic programs. Table 1 presents runtime, peak memory usage, and source/target dataset sizes (n_s, n_t) for each target domain in both datasets.

LTT can be made lightweight by precomputing model scores once before running the procedure. To evaluate scalability, we measure the average runtime and peak memory usage across all LTT runs on the `low` domain of BAF, varying the size of the subsampled calibration set N. Figure 7 shows that both metrics grow roughly linearly with N.

Table 1: Execution time, peak memory usage, and dataset size statistics for KMM.

(a) BAF

Domain	**Time (min)**	**Mem. (GB)**	n_s	n_t
`low`	9.28	1.303	56,196	16,224
`medium`	6.03	1.493	31,537	40,883
`high`	9.45	1.296	57,107	15,313

(b) DomainNet

Domain	**Time (min)**	**Mem. (GB)**	n_s	n_t
`clipart`	33.23	4.597	176,743	14,604
`real`	46.20	4.790	176,743	52,041
`infograph`	35.56	4.603	176,743	15,582
`painting`	38.25	4.635	176,743	21,850
`sketch`	37.22	4.630	176,743	20,916
`quickdraw`	46.75	4.719	176,743	51,750

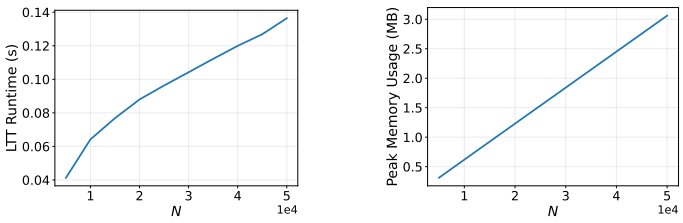


Figure 7: Runtime (left panel) and peak memory usage (right panel) of LTT-IW as a function of calibration subset size on the `low` domain of BAF. Both scale approximately linearly.

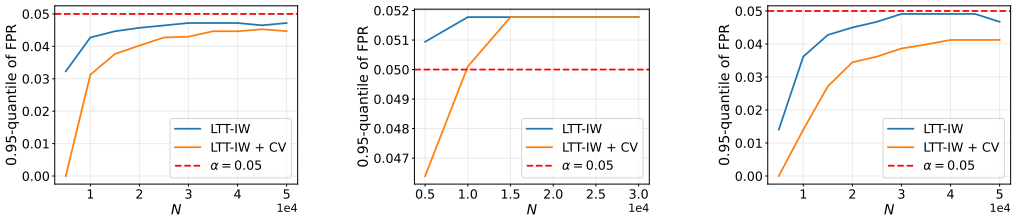
Appendix D. Variance Reduction Techniques

High-variance losses can inflate p-values, reducing the power to detect risk-controlling configurations. One classical approach to mitigate this is to use a control variate $T(X, Y; \lambda)$ (Glasserman, 2003), which yields the following adjusted loss with the same expectation:

$$L_{\text{cv}}(X, Y; \lambda) = w(X)L(X, Y; \lambda) + \eta(T(X, Y; \lambda) - \mathbb{E}[T(X, Y; \lambda)]),$$

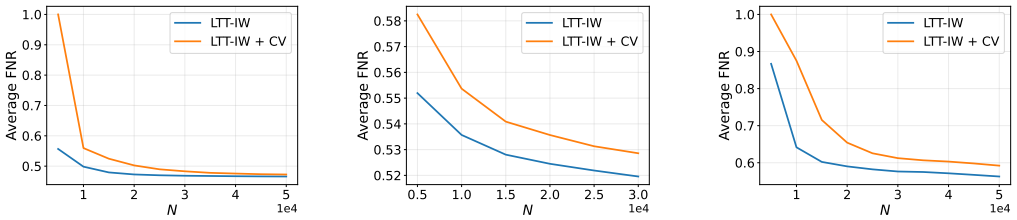
where $\eta = -\text{Cov}[w(X)L(X, Y; \lambda), T(X, Y; \lambda)]/\text{Var}[T(X, Y; \lambda)]$ is chosen to to minimize the variance of L_{cv}. We choose $T(X, Y; \lambda) = w(X)$, since $\mathbb{E}_{\mathbb{P}_{\text{source}}}[w(X)] = 1$ (You et al., 2019). Figures 8 and 9 show the results for LTT-IW with and without control variates. To preserve the i.i.d. nature of the data, we use half of the source data to estimate η and perform risk control on the other half. The introduction of control variates yields more conservative

results; the variance reduction obtained may be outweighed by the fewer data used for hypothesis testing. Future work could explore more sample-efficient variance reduction techniques or smarter ways to allocate data for variance reduction. It is important to note that the range of the loss changes from $[0, B]$ to $[\min(\eta B, 0) - \eta, \max((1+\eta)B, 0) - \eta]$. For $\eta > 0$, it expands from B to $(1+\eta)B$, meaning variance reduction comes at the cost of a wider range, which may reduce p-value power.



(a) `low` (b) `medium` (c) `high`

Figure 8: 0.95 FPR quantiles (vertical axis) vs. source sample size (horizontal axis) for each target domain in BAF. Control variates yield more conservative risk levels.



(a) `low` (b) `medium` (c) `high`

Figure 9: Mean FNR (vertical axis) vs. source sample size (horizontal axis) for each target domain in BAF. Using control variates results in higher FNR.

Appendix E. Sensitivity to the Importance Weight Upper Bound

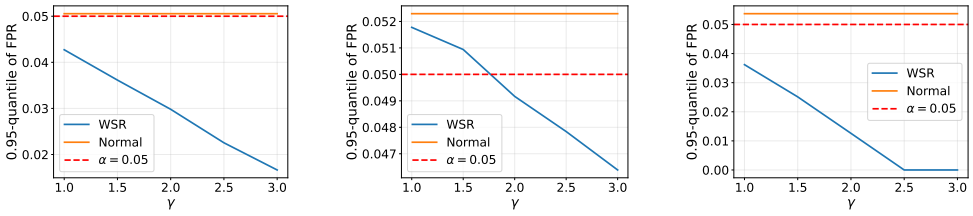
This appendix reports a sensitivity analysis regarding the choice of the importance weight upper bound B, comparing the WSR p-value to the (asymptotically valid) normal p-value (Angelopoulos et al., 2025):

$$p(\lambda, \alpha) = \Phi\left(\frac{n_s^{-1}\sum_{i=1}^{n_s} w(X_i)L(X_i, Y_i; \lambda) - \alpha}{\sqrt{\hat{\sigma}_w^2/n_s}}\right),$$

which eliminates the need to specify B. We consider the BAF dataset and 10,000 calibration points, setting $B = \gamma\hat{B}$ for $\gamma \in \{1, 1.5, 2, 2.5, 3.0\}$, where $\hat{B}$ is the sample maximum importance weight. As shown in Figures 10 and 11, LTT-IW becomes increasingly conservative with larger γ, while the normal p-value remains unaffected, as expected.

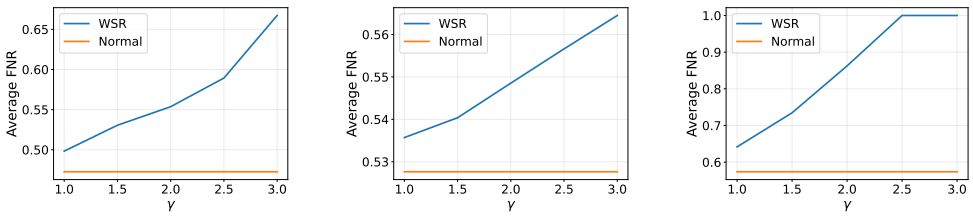
Kuzborskij and Szepesvári (2020) propose a self-normalized high-probability lower bound on the true risk. Let L be a loss supported in $[0, 1]$; then, with probability at least $1 - (n_s + 1)e^{-x}$, for $x \geq 2$ and $y \geq 0$, we have

$$R_{\mathbb{P}_{\text{target}}}(\lambda) \geq \frac{N_x(n_s)}{n_s}\left(\frac{\sum_{i=1}^{n_s} w(X_i)L(X_i, Y_i; \lambda)}{\sum_{i=1}^{n_s} w(X_i)} - \sqrt{2(2V_{\text{W}} + y)\left(1 + \ln\left(\sqrt{1 + 2V_{\text{W}}/y}\right)\right)x}\right),$$



(a) `low` (b) `medium` (c) `high`

Figure 10: 0.95 FPR quantiles (vertical axis) vs. source sample size (horizontal axis) for each target domain in BAF as a function of the multiplicative factor γ.



(a) `low` (b) `medium` (c) `high`

Figure 11: Mean FNR (vertical axis) vs. source sample size (horizontal axis) for each target domain in BAF as function of the multiplicative factor γ.

where $N_x(n_s) = \left(n_s - \sqrt{2xn_s\,\mathbb{E}[w^2(X)]}\right)_+$ and $V_{\mathrm{W}} = \frac{1}{N_x^2(n_s)} \sum_{i=1}^{n_s} \left(w^2(X_i) + \mathbb{E}[w^2(X)]\right)$. From here, an upper bound can be derived for hypothesis testing by setting $L' = 1 - L$, identifying the l.h.s. with $1 - \mathbb{E}[L(X, Y; \lambda)]$ and solving the inequality for $\mathbb{E}[L(X, Y; \lambda)]$. However, it assumes that $\mathbb{E}[w^2(X)]$ is known. Future work could address this limitation, as well as investigate optimal choices of y.

Appendix F. LightGBM Parameter Grid

Table 2: Hyperparameter grid used for LightGBM

Parameter	**Suggestion type**	**Range**
learning rate	log-uniform	[0.01, 0.50]
max. number of leaves	integer	[10, 201]
max. depth	integer	[−1, 21]
min. points per leaf	integer	[20, 101]
data subsample fraction	uniform	[0.5, 1.0]
feature subsample fraction	uniform	[0.5, 1.0]
boosting iterations	integer	[50, 1000]
L_2 regularisation constant	log-uniform	[10, 10 000]
negative-class subsample fraction	uniform	[0.01, 0.5]
early-stopping rounds	integer	[20, 100]