Anytime-Valid Conformal Risk Control

Bror Hultberg
Uppsala University
bror.hultberg@it.uu.se

Dave Zachariah
Uppsala University
dave.zachariah@it.uu.se

Antônio H. Ribeiro
Uppsala University & SciLifeLab
antonio.horta.ribeiro@it.uu.se

Abstract

Prediction sets provide a means of quantifying the uncertainty in predictive tasks. Using held out calibration data, conformal prediction and risk control can produce prediction sets that exhibit statistically valid error control in a computationally efficient manner. However, in the standard formulations, the error is only controlled on average over many possible calibration datasets of fixed size. In this paper, we extend the control to remain valid with high probability over a cumulatively growing calibration dataset at any time point. We derive such guarantees using quantile-based arguments and illustrate the applicability of the proposed framework to settings involving distribution shift. We further show that our guarantees are asymptotically tight, in the sense that the risks converge to the target level from below, and that the rate of our correction cannot be improved. Finally, we demonstrate the practical performance of our methods through both simulations and real-world numerical examples.

1 Introduction

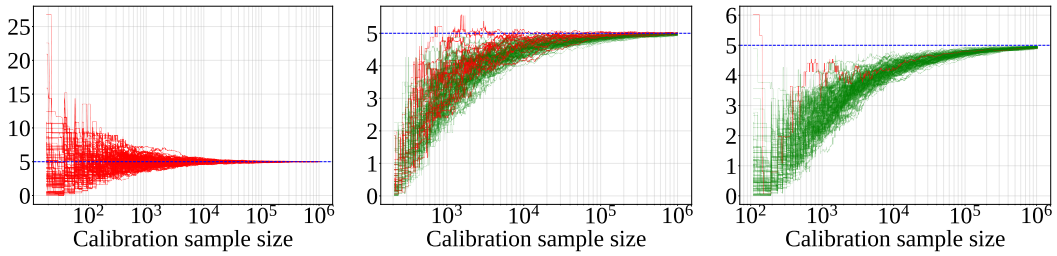
Standard prediction problems focus on a point prediction $f(X)$ of an outcome variable Y. A prediction *set* $\mathcal{C}(X)$, which aims to cover Y, is often more informative as the set size reflects uncertainty of the outcome. Thus $\mathcal{C}(X)$ could be a variable interval or a collection of labels depending on the domain of the outcome. Split conformal prediction is a widely used framework for constructing prediction sets. Using a held out calibration set $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ (Vovk et al., 2005; Angelopoulos & Bates, 2023), the framework generates a prediction set $\mathcal{C}_n(X)$ that ensures a miscoverage rate no greater than α for the next observation (X_{n+1}, Y_{n+1}), i.e.,

$$\mathbb{P}\big(Y_{n+1} \notin \mathcal{C}_n(X_{n+1})\big) \leq \alpha. \tag{1}$$

While elegant and computationally simple, the achievable miscoverage control (1) has certain limitations.

One main limitation about the procedure is that it is a data-marginalized control, meaning that the probability is evaluated *jointly* over the calibration set $\mathcal{D}_n$ and the out-of-sample test point (X_{n+1}, Y_{n+1}). Consequently, for any realized dataset $\mathcal{D}_n$, the miscoverage rate can vary widely from α (Vovk, 2012; Marques F., 2025; Zwart, 2025). For instance, when $n = 500$ and $\alpha = 0.1$ there is a 48% chance of drawing $\mathcal{D}_n$ such that the rate exceeds the specified α, see Figure 1a. This limitation is well known, and *sample-conditional* (or PAC-style) guarantees address it by ensuring that the miscoverage rate of the realized prediction set is at most α with a given probability over the draw of $\mathcal{D}_n$, say, $1 - \delta = 90\%$ (Vovk, 2012; Bates et al., 2021; Duchi, 2025), see Figure 1b. Such guarantees, however, are stated for a calibration set of a single, fixed size n.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Standard method (b) Fixed-time valid, $\delta = 10\%$. (c) Anytime-valid, $\delta = 10\%$.

Figure 1: Miscoverage rates of prediction sets $\{\mathcal{C}_n(X)\}$ versus calibration sample size n. Each line corresponds to a particular draw of calibration data and the miscoverage rates are evaluated conditional on this data at each n. A **green** line indicates that *all* the rates fall below $\alpha = 5\%$, while **red** indicates a failure. Details are given in Section 5.

In many circumstances data collection is a cumulative process such that n in $\mathcal{D}_n$ increases and yields a sequence of prediction sets $\{\mathcal{C}_n(X)\}$. However, (1) and its fixed-time sample-conditional counterparts only control the miscoverage rate for an individual $\mathcal{C}_n(X)$ rather than all the rates for the *entire* sequence of prediction sets with high probability. In fact, by the law of the iterated logarithm and under mild conditions, these corrections lead to miscoverage rates that exceed α for infinitely many n almost surely (Proposition 4.12), as reflected by the many red trajectories in Figure 1b. Overall, we want to ensure that the miscoverage rates approach α as n increases, while remaining below it for all n, with high probability (Figure 1c).

While miscoverage control is important in certain applications, there are other relevant types of error control for $\{\mathcal{C}_n(X)\}$ that we want to consider, such as false negative rate control in multilabel classification. The chosen error metric is the *risk* of a prediction set (Angelopoulos et al., 2024).

In this paper, we propose an approach that jointly addresses all these limitations. Our contributions are as follows:

1. We show that one can construct a sequence of prediction sets $\{\mathcal{C}_n(X)\}$ for which the risks will be no greater than α and approach it from below with high probability.
2. We show that, under mild conditions, the risks converge to α from below at rate $O(\sqrt{n^{-1}\log\log n})$, and we prove that this rate cannot be improved. Any correction term that decays faster fails to achieve anytime-valid risk control almost surely.
3. Finally, we specialize our results to the case of test-time distribution shifts and illustrate them with numerical experiments.

2 Problem Formulation

A prediction set is a subset of the domain of the outcome variable, i.e., $\mathcal{C}(X) \subseteq \mathcal{Y}$. Throughout the paper, the index $n \in \mathbb{N}$ denotes the number of calibration points collected so far, and $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ denotes the calibration set of size n. We assume that the calibration points $(X_1, Y_1), (X_2, Y_2), \ldots$ and test point (X, Y) are drawn independently and identically. We let $\mathcal{C}_n(X)$ be a function of the calibration data $\mathcal{D}_n$.

2.1 Risk of a Prediction Set

The prediction error of $\mathcal{C}(X)$ is quantified by a loss function $\ell(\mathcal{C}(X), Y)$. The *risk* of the *given* prediction set $\mathcal{C}_n(X)$ is defined as:

$$\mathbb{E}[\ell\left(\mathcal{C}_n(X), Y\right) \mid \mathcal{D}_n], \tag{2}$$

which is a sample-conditional risk.

Example 2.1. In regression and single-label classification problems, the miscoverage event is a common loss function $\ell(\mathcal{C}(X), Y) = \mathbf{1}\{Y \notin \mathcal{C}(X)\}$. In this case, (2) is the miscoverage rate of the given prediction set, i.e.,

$$\mathbb{P}(Y \notin \mathcal{C}_n(X) \mid \mathcal{D}_n), \tag{3}$$

Example 2.2. In multilabel classification problems, where $\mathcal{Y} = \{1, \ldots, K\}$ denotes the labels, the fraction of true labels missed by the prediction set $\ell(\mathcal{C}(X), Y) = 1 - |\mathcal{C}(X) \cap Y|/|Y|$ is a suitable loss function. Then (2) is the false negative rate of the given prediction set.

We shall assume that the chosen loss function is *monotone* in prediction set size, i.e.,

$$\ell(\mathcal{C}(X), Y) \geq \ell(\mathcal{C}'(X), Y), \quad \forall \mathcal{C} \subseteq \mathcal{C}' \subseteq \mathcal{Y}$$

and *bounded* in $[0, B]$.

We can also consider the risk under known distribution shifts. That is, when calibration data is drawn as $(X_i, Y_i)_{i=1}^n \sim P$ but the test point is drawn from different distribution $(X, Y) \sim P^*$ so that the test-time risk is

$$\mathbb{E}_{P^*}[\ell\left(\mathcal{C}_n(X), Y\right) \mid \mathcal{D}_n]. \tag{4}$$

It is assumed that P^* is absolutely continuous with respect to P and the Radon–Nikodym derivative

$$\omega(x, y) = \frac{\mathrm{d}P^*(x, y)}{\mathrm{d}P(x, y)} \tag{5}$$

is a known importance weight.

2.2 Conformal Risk Control

We consider a family of prediction sets $\mathcal{C}_\lambda : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ indexed by $\lambda \in \Lambda$, where Λ is totally ordered and the family is nested such that

$$\mathcal{C}_\lambda(X) \subseteq \mathcal{C}_{\lambda'}(X), \quad \lambda \leq \lambda'$$

and all X. It is assumed that there exists a $\lambda_{\max}$ such that $\ell(\mathcal{C}_{\lambda_{\max}}(X), Y) \leq \alpha$, for all (X, Y). Note that $\mathcal{C}_{\lambda_{\max}}(X)$ may equal $\mathcal{Y}$ and is therefore uninformative. Throughout, $\mathcal{C}_\lambda(X)$ denotes a generic prediction set parameterized by λ, and $\mathcal{C}_{\lambda_n}(X)$ denotes the prediction set obtained by setting $\lambda = \lambda_n$ using calibration data of size n; we use $\mathcal{C}_n(X)$ as shorthand for $\mathcal{C}_{\lambda_n}(X)$ when the dependence on λ_n is clear from the context.

Example 2.3. In regression problems, we may use score function, e.g., $s(X, y) = |y - f(X)|$ given a trained model $f(X)$, and then set $\mathcal{C}_\lambda(X) = \{y : s(X, y) \leq \lambda\}$.

Example 2.4. In multilabel classification problems, where $f : \mathcal{X} \to [0, 1]^K$, we may use $\mathcal{C}_\lambda(X) = \{k : f_k(X) \geq 1 - \lambda\}$, for $\lambda \in [0, 1]$.

In conformal prediction and risk control, we specify a *tolerated* risk level α and set a data-dependent threshold λ_n using the calibration set $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$. Specifically, choose the smallest λ subject to a constraint:

$$\boxed{\lambda_n = \inf\left\{\lambda : R_n(\lambda) \leq \alpha - \gamma_n\right\},} \tag{6}$$

where $R_n(\lambda) = \frac{1}{n}\sum_{i=1}^n \ell\left(\mathcal{C}_\lambda(X_i), Y_i\right)$ is the empirical risk and γ_n is any *correction* term. Intuitively, the correction term γ_n acts as a safety margin. It inflates the threshold to account for statistical uncertainty in the empirical risk estimate.

Example 2.5. When the risk is the miscoverage rate, the standard correction term is (Angelopoulos & Bates, 2023):

$$\gamma_n = \frac{\lceil (1 - \alpha)(n + 1) \rceil}{n} - (1 - \alpha). \tag{7}$$

This choice only ensures that the miscoverage rate of prediction sets is below α *when averaged over all n-sized calibration sets*, i.e., marginal risk control: $\mathbb{E}[\mathbb{P}(Y \notin \mathcal{C}_{\lambda_n}(X) | \mathcal{D}_n)] \leq \alpha$.
More recently, Duchi (2025) showed that the correction term

$$\gamma_n = \frac{4 \log \frac{1}{\delta}}{3n} + \sqrt{\left(\frac{4}{3n} \log \frac{1}{\delta}\right)^2 + \frac{2\alpha(1 - \alpha)}{n} \log \frac{1}{\delta}}, \tag{8}$$

where $\delta \in (0, 1)$, ensures that $\mathcal{C}_{\lambda_n}(X)$ has a sample-conditional miscoverage rate no greater than α with probability at least $1 - \delta$, i.e., $\mathbb{P}\big(\, \mathbb{P}(Y \notin \mathcal{C}_{\lambda_n}(X) \mid \mathcal{D}_n) \leq \alpha \,\big) \geq 1 - \delta$. This strengthens the miscoverage control of the prediction set.

Remark 2.6. If the set inside the infimum in (6) is empty, we define $\lambda_n = \lambda_{\max}$. We assume that the empirical risk $R_n(\lambda)$ is right-continuous in λ, meaning that for every λ and n,

$$\lim_{\epsilon \to 0^+} R_n(\lambda + \epsilon) = R_n(\lambda).$$

Right-continuity is standard in the conformal risk control literature (Angelopoulos et al., 2024). It guarantees that the infimum in (6) is attained, i.e., $R_n(\lambda_n) \leq \alpha - \gamma_n$. Without it, the selected threshold could violate the constraint. For the standard construction $\mathcal{C}_\lambda(X) = \{y : s(X, y) \leq \lambda\}$ with the miscoverage loss, $\ell(\mathcal{C}_\lambda(X), Y) = \mathbf{1}\{s(X, Y) > \lambda\}$ is right-continuous in λ, so the assumption holds automatically.

2.3 Risk Control with Anytime Validity

In this paper, we consider cumulative data collection processes such that n in $\mathcal{D}_n$ increases and yields a sequence of prediction sets $\{\mathcal{C}_{\lambda_n}(X)\}$. This motivates the following definition.

Definition 2.7. Let ℓ be a loss function and α a tolerated level of risk. For any given $\delta \in (0, 1)$, a method producing a sequence of prediction sets $\{\mathcal{C}_{\lambda_n}(X)\}$ is said to achieve *anytime-valid risk control*, if

$$\boxed{\mathbb{P}\Big(\forall n \in \mathbb{N} : \ \mathbb{E}[\ell\left(\mathcal{C}_{\lambda_n}(X), Y\right) \mid \mathcal{D}_n] \ \leq \ \alpha\Big) \ \geq \ 1 - \delta.} \tag{9}$$

That is, with a probability of at least $1 - \delta$, we have a sequence of prediction sets for which *all* the risks are no greater than the tolerance α. Additionally, we say that the risk control is *asymptotically tight* if the risks approach the specified level α from below.

In (9), the inner expectation is taken over the test point $(X, Y) \sim P$ (or $(X, Y) \sim P^*$ under distribution shift) with the realized calibration set $\mathcal{D}_n$ held fixed, whereas the outer probability is taken over the entire infinite sequence of calibration data $(X_1, Y_1), (X_2, Y_2), \ldots$. A similar notion of anytime-valid risk control is considered by Xu et al. (2024).

The main question we aim to address is how to construct a *sequence* of prediction sets $\{\mathcal{C}_{\lambda_n}(X)\}$ using (6) that achieves anytime valid risk control and is asymptotically tight.

3 Related Work

Conformal prediction and risk control. Conformal prediction has had a significant impact in machine learning and overviews are provided in (Vovk et al., 2005; Angelopoulos & Bates, 2023; Angelopoulos et al., 2025). Vovk (2012) studied miscoverage rate of $\mathcal{C}_{\lambda_n}(X)$ when *conditioned* on $\mathcal{D}_n$ and a distributional characterization is provided in (Hulsman, 2022; Marques F., 2025), which can be used to improve miscoverage control (Zwart, 2025) as alternative to (8). Duchi (2025) showed that using (8) when the scores are pairwise distinct also yields a lower bound,

$$\alpha - \frac{1}{n} - 2\gamma_n \leq \mathbb{P}(Y_{n+1} \notin \mathcal{C}_n(X_{n+1}) \mid \mathcal{D}_n) \leq \alpha,$$

that holds with probability of at least $1 - 2\delta$. Miscoverage control was generalized to marginal and sample-conditional risk control in (Angelopoulos et al., 2024) and (Bates et al., 2021), respectively. Sample-conditional miscoverage control under distribution shifts was recently achieved in (Pournaderi & Xiang, 2026). Compared to Pournaderi & Xiang (2026), who also consider distribution shift, our guarantees are anytime-valid, apply to general bounded loss functions, and incorporate the importance weights explicitly in the correction term, resulting in bounds of a different nature. The two approaches rely on different proof techniques, a weighted DKW inequality versus sub-gamma concentration. As a special case of our framework, fixed-time sample-conditional bounds under covariate shift can be obtained, and we provide explicit instances in Appendix C.

Anytime-valid concentration bounds. A large body of work on sample-conditional conformal prediction under the i.i.d. setting relies on classical concentration inequalities, see, for instance (Vovk, 2012; Duchi, 2025; Bates et al., 2021). While most prior work relies on concentration inequalities for a fixed sample size, such tools are not always well suited to sequential settings. Anytime-valid concentration bounds, which hold uniformly over time, provide a natural alternative and have been studied for several decades, dating back to at least (Darling & Robbins, 1967). These bounds have

been refined and extended over the years (Howard et al., 2021; Howard & Ramdas, 2022; Waudby-Smith & Ramdas, 2024; Waudby-Smith et al., 2024). In certain settings, asymptotic rates on the order of $O\left(\sqrt{n^{-1}\log\log n}\right)$ are achieved, and this asymptotic rate is known to be optimal (Duchi & Haque, 2024). In this paper, we adapt the quantile construction of Howard & Ramdas (2022) to the conformal risk control framework of Angelopoulos et al. (2024), obtaining anytime-valid guarantees via sub-gamma concentration arguments following Howard et al. (2021, 2020). What is new is the reduction to a mean-zero bounded process at an oracle threshold, the resulting closed-form correction, the tightness and rate-optimality results and the importance-weighted extension.

Sequential conformal prediction and risk control. A related line of work studies conformal prediction in sequential settings, where coverage is guaranteed on average over time (Gibbs and Candès, 2021). Most closely related is Ge et al. (2025), who obtain a guarantee of the form (9), but only up to a fixed, known horizon T, as their focus is on regret minimization, whereas we require no pre-specified horizon. Xu et al. (2024) achieve anytime-valid risk control, also for actively labeled data, by running a betting-based e-process for each candidate threshold. Their thresholds are defined implicitly through the wealth processes, and their analysis concerns the growth rate of these processes, so how fast the risk approaches α is not explicit. In contrast, our correction γ_n is closed-form and plugs directly into (6), so this rate is explicit (Proposition 4.6) and of optimal order among corrections of this form (Proposition 4.12). Our construction also extends to covariate shift with known likelihood ratios. Gauthier et al. (2025) use e-values to construct prediction sets that are simultaneously valid across sequentially arriving data batches, rather than uniformly valid as the calibration set grows within a single batch.

4 Main Results

We now turn to setting the correction term in (6) such that the resulting sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ achieves anytime-valid risk control and is asymptotically tight. The threshold sequence $\{\lambda_n\}$ will be non-increasing and, when necessary, we use a running minimum. Due to space limits, the proofs are relegated to Appendix B.

4.1 Anytime-Valid Conformal Risk Control

Intuition. It suffices to control the gap between R_n and the population risk $R(\lambda) = \mathbb{E}[\ell(\mathcal{C}_\lambda(X), Y)]$ at the deterministic *oracle threshold* $\hat{\lambda}(\alpha) = \inf\{\lambda : R(\lambda) \le \alpha\}$. If $R_n(\hat{\lambda}(\alpha)) > \alpha - \gamma_n$, then $\lambda_n \ge \hat{\lambda}(\alpha)$ by monotonicity, so the risk of $\mathcal{C}_{\lambda_n}(X)$ is at most α. Thus, γ_n only needs to be a deviation bound, valid simultaneously for all n, for an average of n i.i.d. losses in $[0, B]$ with mean α (if R is continuous at $\hat{\lambda}(\alpha)$; see Appendix B otherwise), whose accumulated variance is at most $v = \alpha(B-\alpha)n$. We let $m \in \mathbb{N}$, $c \ge 0$ and $\delta \in (0,1)$. For notational convenience, define the following two functions

$$h_{m,\delta}(v) \;=\; 2\,\log\left(\log_2\left(\frac{\max\{v,m\}}{m}\right)+1\right) + \;\log(\pi^2/(6\delta)),$$

and

$$f_{c,m,\delta}(v) \;= 1.44\sqrt{v\,h_{m,\delta}(v)} \;+\; 2.42\,c\,h_{m,\delta}(v).$$

The function $f_{c,m,\delta}$ is a Bernstein-type bound with a variance term and a range term with scale c, where $h_{m,\delta}$ replaces the factor $\log(1/\delta)$ of a fixed-time bound.

Theorem 4.1. *Let ℓ be an arbitrary loss function bounded in $[0, B]$, monotone in λ, and right-continuous in λ. Let $c_\alpha = \max\{(2\alpha - B)/3, 0\}$. Construct the prediction sets according to (6) and set the correction term as*

$$\gamma_n = f_{c_\alpha,\alpha(B-\alpha)m^*,\delta}\big(\alpha(B-\alpha)n\big)/n, \tag{10}$$

where

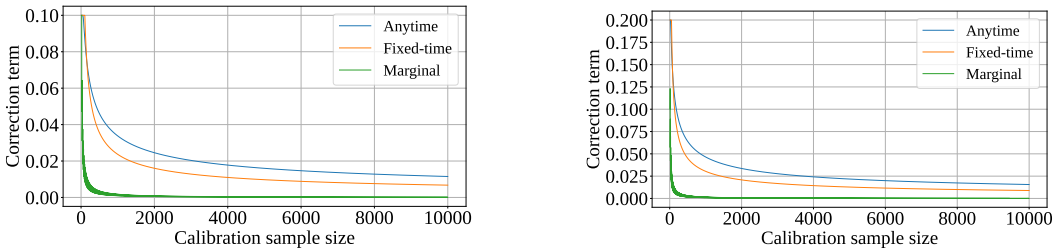
$$m^* \;=\; \min\Big\{\, m' \in \mathbb{N} : f_{c_\alpha,\alpha(B-\alpha)m',\delta}\big(\alpha(B-\alpha)m'\big)/m' \le \alpha \,\Big\}.$$

Then the resulting sequence of prediction sets $\{\mathcal{C}_{\lambda_n}(X)\}$ achieves anytime-valid risk control (9).

Corollary 4.2. *In conformal prediction, where ℓ is the miscoverage loss, anytime valid miscoverage rate control is achieved by setting $B = 1$ in (10).*

The proofs, provided in Appendix B, are based on constructing suitable sub-gamma processes, applying a time-uniform concentration bound, and carefully handling potential discontinuities in the population risk.

Remark 4.3. A smaller correction term γ_n yields smaller prediction sets (i.e. higher efficiency). Note, however, that smaller prediction sets do not automatically imply valid risk control, so efficiency is meaningful only when paired with a validity guarantee. We plot the correction terms (7), (8), and (10) against n in Figure 2, illustrating the cost to strengthen marginal to fixed-time and anytime validity in terms of sample efficiency (visible as the vertical gap between the curves).



(a) $\alpha = 10\%, \delta = 10\%$ (b) $\alpha = 20\%, \delta = 10\%$

Figure 2: Correction term γ_n in (6) versus n. Using (7), (8), and (10), guarantees marginal, fixed-time and anytime validity, respectively.

Remark 4.4. For $n < m^*$, we have $\gamma_n > \alpha$. Then (6) yields the infimum of an empty set which, by definition, results in $\lambda_n = \lambda_{\max}$ and an uninformative prediction set at that sample n. The burn-in m^* is a design choice. The proof of Theorem 4.1 remains valid if m^* in (10) is replaced by any $m \in \mathbb{N}$ fixed in advance, with $\alpha(B-\alpha)n$ replaced by $\alpha(B-\alpha)\max\{n, m\}$. No choice of m yields informative sets before $n = m^*$, and $m < m^*$ only increases γ_n for $n \geq m^*$. Choosing $m > m^*$ delays informative sets but tightens them for all $n \geq m$, as less of the error budget δ is spent on small n. Hence, there is no uniformly optimal choice. If prediction sets are only needed from some calibration size $n_0 \geq m^*$ onward, $m = n_0$ is a natural choice.

Remark 4.5. There are alternative ways to construct valid sequences of correction terms. For further details, we refer to Remark A.4.

Assuming that the empirical risk R_n does not exhibit large jumps, we can show that the choice of correction terms γ_n is tight up to a controlled error. For a right-continuous function l, we define the jump at λ by $J(l, \lambda) = \lim_{\epsilon \to 0^+} l(\lambda - \epsilon) - l(\lambda)$.

Proposition 4.6. *In the setting of Theorem 4.1, let $\{\lambda_n\}$ be non-increasing (using a running minimum if necessary) and assume that $\alpha \leq B/2$ and $\alpha \leq \sup_{\lambda \in \Lambda} R(\lambda)$, where $R(\lambda) = \mathbb{E}[\ell(\mathcal{C}_\lambda(X), Y)]$. Furthermore, let $d_n = \sup_{\lambda \in \Lambda} J(R_n, \lambda)$. If d_n converges to 0, then the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ is asymptotically tight with the risk approaching α at the rate $O\left(\max\left\{d_n, B\sqrt{n^{-1}\log\log n}\right\}\right)$. In particular, the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ satisfies*

$$\mathbb{P}\big(\forall n : \alpha - k_n \leq \mathbb{E}[\ell\left(\mathcal{C}_{\lambda_n}(X), Y\right) | \mathcal{D}_n] \leq \alpha\big) \geq 1 - 2\delta,$$

where $k_n = d_{g(n)} + \gamma_{g(n)} + \gamma'_{g(n)}(\delta_{\lfloor \log_2 n \rfloor + 1})$, $g(n) = 2^{\lfloor \log_2 n \rfloor}$, $\delta_i = \delta/(i(i+1))$, and

$$\gamma'_n(\delta) = \frac{B}{n}\log\frac{1}{\delta} + \sqrt{2\frac{\alpha(B-\alpha)}{n}\log\frac{1}{\delta} + \left(\frac{B}{n}\log\frac{1}{\delta}\right)^2}.$$

Remark 4.7. The condition $d_n \to 0$ is satisfied in several settings. For instance, if R_n is continuous, then $d_n = 0$. Alternatively, let $L_i(\lambda) = \ell(\mathcal{C}_\lambda(X_i), Y_i)$ and assume that for any $\lambda \in \Lambda$ and i, $\mathbb{P}\left(J(L_i, \lambda) > 0\right) = 0$. Then $d_n \overset{\text{a.s.}}{\leq} B/n$ (Lemma 1, Angelopoulos et al., 2024). This can be interpreted as follows: for any fixed λ, the function L_i is almost surely continuous. In other words, L_i might have discontinuities, but their locations are random rather than fixed. For example, in

conformal prediction with almost surely distinct scores, every jump of R_n has size exactly $1/n$, so that $d_n \leq 1/n$. The condition may fail if the score distribution has atoms. The clipped score $\min\{|y - f(x)|, C\}$ places positive mass at C, in which case d_n converges to a positive constant. Thus, the condition concerns the distribution of the jump locations rather than continuity of the loss or the score function. In practice, d_n can be computed directly from the calibration data, since R_n is already evaluated when solving (6). Moreover, the proof only uses the size of the jump at the selected threshold, so d_n may be replaced by the smaller, also computable quantity $J(R_n, \lambda_n)$.

4.2 Anytime-Valid Risk Control under Distribution Shift

We generalize Theorem 4.1 to the setting of conformal risk control under test-time distribution shift. The calibration data is drawn from a distribution P, which may differ from distribution P^* for the test data. The goal is to control the test-time risk (4), assuming $\omega(x, y)$ is given (5).

Theorem 4.8. *Let ℓ be an arbitrary loss function bounded in $[0, B]$, monotone in λ, and right-continuous in λ. Define $b_{m,\delta}(v) = 1.44\sqrt{v h_{m,\delta}(v)}$, i.e., $b_{m,\delta} = f_{0,m,\delta}$. Define the importance weights $\omega_i := \omega(X_i, Y_i)$ using (5) and $W_n = \sum_{i=1}^n \omega_i^2$, and fix $m > 0$ before observing the calibration data. Construct the prediction sets according to (6), using the importance-weighted loss $\omega(x, y)\,\ell(\mathcal{C}_\lambda(x), y)$, i.e., with empirical risk $R_n^\omega(\lambda) = \frac{1}{n}\sum_{i=1}^n \omega_i\,\ell(\mathcal{C}_\lambda(X_i), Y_i)$, with the correction term as*

$$\gamma_n = B\left(1 - \frac{1}{n}\sum_{i=1}^n \omega_i\right) + \frac{b_{m,\delta}(\max\{B^2 W_n, m\})}{n}.$$

Then the resulting sequence of prediction sets $\{\mathcal{C}_{\lambda_n}(X)\}$ achieves anytime-valid risk control (9) under distribution shift (4). That is, with probability at least $1 - \delta$, we obtain a sequence of prediction sets such that

$$\mathbb{E}_{P^*}[\ell(\mathcal{C}_{\lambda_n}(X), Y) \mid \mathcal{D}_n] \;\leq\; \alpha$$

holds for all n.

Remark 4.9. For $\alpha \leq B/2$, we have $c_\alpha = 0$, so Theorems 4.1 and 4.8 use the same boundary $b_{m,\delta}$ and differ mainly in the variance proxy. Without distribution shift, the losses at the oracle threshold lie in $[0, B]$ and have mean α, so the relevant (lower) tail of their average is sub-Gaussian with the small variance $\alpha(B - \alpha)$. Under distribution shift, the weighted losses may be unbounded, and their variability is controlled through the observed weights, which leads to the variance proxy $B^2 W_n$. For further technical details, we refer to Remark B.2.

Remark 4.10. Since P^* is absolutely continuous with respect to P, we have $\mathbb{E}_P[\omega_i] = 1$. By the strong law of large numbers, the term $B\big(1 - \frac{1}{n}\sum_{i=1}^n \omega_i\big)$ therefore tends to zero almost surely, with no further assumption on the weights.

Proposition 4.11. *In the distribution-shift setting of Theorem 4.8, let $\{\lambda_n\}$ be non-increasing (using a running minimum if necessary) and assume that $\alpha \leq \sup_{\lambda\in\Lambda} \mathbb{E}_{P^*}[\ell(\mathcal{C}_\lambda(X), Y)]$. Let $d_n = \sup_{\lambda\in\Lambda} J(R_n^\omega, \lambda)$. If d_n converges to 0 and $\sqrt{\frac{W_n \log\log W_n}{n^2}} \to 0$, where the latter holds almost surely whenever $\mathbb{E}_P[\omega^2] < \infty$, then the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ is asymptotically tight. In particular, the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ satisfies*

$$\mathbb{P}\big(\forall n : \alpha - k_n \leq \mathbb{E}_{P^*}[\ell\left(\mathcal{C}_{\lambda_n}(X), Y\right) | \mathcal{D}_n] \leq \alpha\big) \geq 1 - 2\delta,$$

where $k_n = d_{g(n)} + \gamma_{g(n)} + \gamma'_{g(n)}$, $g(n) = 2^{\lfloor \log_2 n \rfloor}$, and $\gamma'_n = 2^{j_n}/n - \gamma_n$, where

$$j_n = \min\left\{j \geq 0 : \; \frac{2^j}{n} \geq \gamma_n + \frac{b_{m,\delta_{n,j}}(\max\{B^2 W_n, m\})}{n}\right\}, \qquad \delta_{n,j} = \frac{\delta_{\lfloor \log_2 n \rfloor + 1}}{(j+1)(j+2)},$$

and $\delta_i = \frac{\delta}{i(i+1)}$.

4.3 Optimality of the Rate

Proposition 4.6 shows that the risks of our procedure converge to α at rate $O(\sqrt{n^{-1}\log\log n})$ when d_n vanishes sufficiently fast. The correction term (10) is of order $\sqrt{n^{-1}\log\log n}$, and we now show that any such term in (6) that vanishes faster than this rate violates the risk constraint for infinitely many n almost surely.

Proposition 4.12 (Rate optimality). *Let ℓ be the miscoverage loss, $\alpha \in (0,1)$, and let $\Lambda \subseteq \mathbb{R}$ be an interval. Suppose that the population risk $R(\lambda) = \mathbb{E}[\ell(\mathcal{C}_\lambda(X), Y)]$ is continuous at $\hat{\lambda}(\alpha) = \inf\{\lambda : R(\lambda) \le \alpha\}$, which lies in the interior of Λ. Consider any procedure of the form (6) with a deterministic correction sequence $\gamma_n \ge 0$ satisfying $\gamma_n = o\big(\sqrt{n^{-1}\log\log n}\big)$. Then*

$$\mathbb{P}\Big(\forall n \in \mathbb{N} : \ \mathbb{E}[\ell(\mathcal{C}_{\lambda_n}(X), Y) \mid \mathcal{D}_n] \le \alpha\Big) = 0.$$

In particular, no such correction sequence achieves anytime-valid risk control (9) for any $\delta \in (0,1)$.

The proof, given in Appendix B, relies on the law of the iterated logarithm. In particular, the standard correction (7) and the fixed-time correction (8), which decay as $O(n^{-1})$ and $O(n^{-1/2})$, respectively, cannot achieve anytime-valid risk control, regardless of δ. Since the correction term (10) is of order $\sqrt{n^{-1}\log\log n}$, Theorem 4.1 and Propositions 4.6 and 4.12 together show that our correction has the optimal rate.

5 Numerical Experiments

We illustrate the performance of the anytime-valid risk controlled sets in five examples below.

Synthetic Regression. The data are generated according to the linear model

$$Y_i = 2X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,1), \quad X_i \sim \text{Uniform}(-3,3),$$

where $i \in \mathbb{N}$. We set the target miscoverage level to $\alpha = 5\%$, $\delta = 10\%$, and let $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$. Assume we have access to a (correctly specified) predictive model f. We use a simple score function $s(X,y) = |y - f(X)|$. Following Corollary 4.2, we construct a sequence of prediction sets $\{\mathcal{C}_{\lambda_n}(X)\}$. For each $n \in \mathbb{N}$, we evaluate the corresponding miscoverage probability $\mathbb{P}(Y \notin \mathcal{C}_{\lambda_n}(X) | \mathcal{D}_n)$.

Since the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ depends on the realized calibration sample $\mathcal{D}_n$ the resulting miscoverage probabilities are random. To characterize this variability, we repeat the entire data collection and prediction-set construction procedure over multiple independent runs, and for each run plot the miscoverage probability as a function of n in Figure 1c.

For comparison, we also construct prediction sets using the standard correction term in (7) and the fixed-time valid correction term in (8). The latter shows a clearer structure in which the miscoverage generally approaches α from below. While the probability of achieving miscoverage below α is at least $1 - \delta$ for any fixed n, the procedure is not uniformly valid over time, and consequently many runs still exceed the miscoverage threshold at some sample size n.

Indeed, by the law of the iterated logarithm, using either correction term in (7) or (8) yields a sequence of risks that almost surely exceeds the tolerance level α infinitely often (see, e.g., Theorem 8.5.2 in Durrett (2019)). In finite simulations, this behavior may not be observed, since only a finite time horizon is considered. In contrast, the anytime-valid procedure maintains miscoverage below α across runs, with most sequences remaining green, and the realized miscoverage approaching α from below as n increases. Figure 3 shows the corresponding average prediction set sizes.

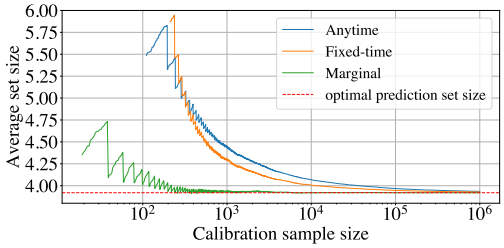


Figure 3: Average prediction set sizes across methods.

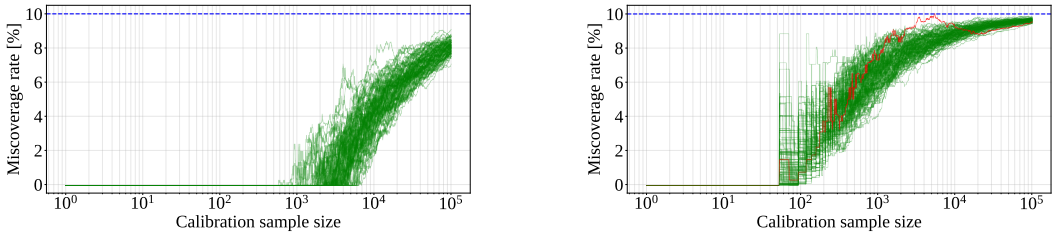
Synthetic Regression under Distribution Shift. We illustrate Theorem 4.8 using the example in Section 2 of Shimodaira (2000). We assume that the data are generated according to

$$Y_i = -X_i + X_i^3 + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 0.3^2),$$

where $X_i \sim \mathcal{N}(0.5, 0.5^2)$. A linear model f is fitted using $n = 1000$ held-out data points. We assume that test points come from a different distribution P^*, with which $X \sim \mathcal{N}(0, 0.3^2)$. Thus $\omega(x,y) = p^*(x)/p(x)$ is a ratio of Gaussian densities. The aim is to control the miscoverage rate

$$\mathbb{E}_{(X,Y)\sim P^*}\left[\mathbf{1}\{Y \notin \mathcal{C}_{\lambda_n}(X)\} | \mathcal{D}_n\right].$$

We set $\alpha = 10\%, \delta = 10\%$, and perform a Monte Carlo simulation as in the synthetic experiment above. The resulting miscoverage behavior is illustrated in Figure 4a. For comparison, we also plot the resulting miscoverage in the absence of distribution shift, $\omega(x, y) \equiv 1$, using Corollary 4.2. Figure 4b shows the convergence towards α is substantially faster in this case.



(a) With distribution shift.

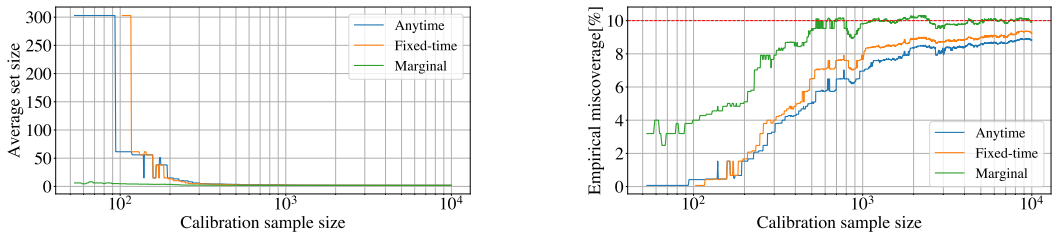
(b) Without distribution shift ($\omega \equiv 1$).

Figure 4: Miscoverage rates of prediction sets $\{\mathcal{C}_n(X)\}$ versus calibration sample size n, with $\alpha = 10\%, \delta = 10\%$.

Image Classification. We apply Corollary 4.2 to image classification using the data set ImageNet-1K, which consists of $K = 1000$ different classes. We use the base implementation from Angelopoulos & Bates (2022). Our goal is, given an image, to output a set of classes containing the true class with probability of at least $1 - \alpha$. Let $\mathcal{X}$ denote the image space and $\mathcal{Y} = \{1, \dots, K\}$ the set of classes. For an image x, the model gives estimated class probabilities

$$f(x) \in [0, 1]^K, \quad \sum_{i=1}^{K} f_i(x) = 1.$$

For a sample (X, Y), where $X \in \mathcal{X}$ is the image and $Y \in \mathcal{Y}$ its true class, we define the score function as $s(X, Y) = 1 - f_Y(X)$. We target miscoverage at level $\alpha = 10\%$ with $\delta = 10\%$, and compare the prediction sets obtained using Corollary 4.2, the fixed-time correction (8), and the standard correction (7). Using a held-out evaluation set, we estimate the expected sizes of these prediction sets and plot the results in Figure 5a. They illustrate that anytime-validity can be gained at equivalent prediction set sizes even with a small increase of sample size n. We compare the empirical miscoverage in Figure 5b.



(a)

(b)

Figure 5: Comparison of anytime-valid conformal prediction using Corollary 4.2, the fixed-time method (8), and marginal conformal prediction (7) for ImageNet-1K classification with $\alpha = 10\%$ and $\delta = 10\%$.

Synthetic False Negative Rate Control. We construct a synthetic multilabel classification problem with $K = 3$ labels and $D = 3$-dimensional input features. The covariates are generated as $X_i \sim \mathcal{N}(0, I_D)$. For each instance i, the label vector $Y_i \in \{0, 1\}^K$ has conditionally independent components given X_i, with

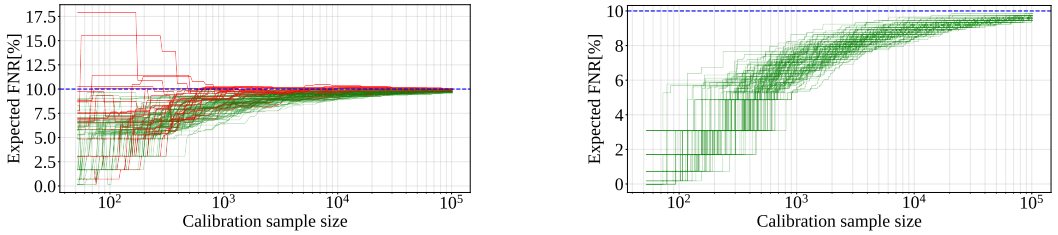
$$Y_{ik} \sim \text{Bernoulli}\big(\sigma\big(e_k^\top X_i - 1\big)\big), \quad k = 1, \dots, K,$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function and e_k is the k-th standard basis vector in $\mathbb{R}^D$.

We assume access to a perfectly specified model f that outputs the true class probabilities, i.e., $f_k(X) = \mathbb{P}(Y_k = 1 \mid X)$. Prediction sets are constructed via thresholding:

$$\mathcal{C}_\lambda(X) = \{\, k \in \{1, \dots, K\} : f_k(X) \geq 1 - \lambda \,\}.$$

We target FNR control from Example 2.2 at level $\alpha = 10\%$ with $\delta = 10\%$, and repeat the same experiment as in the synthetic regression experiment above. Figure 6b shows the realized FNR of the sequence of prediction sets constructed using Theorem 4.1, while Figure 6a shows the corresponding results using the method of Bates et al. (2021) (Proposition 5 therein), which applies the fixed-time Waudby-Smith–Ramdas upper confidence bound (Waudby-Smith & Ramdas, 2024) to achieve sample-conditional risk control at a single calibration size n.



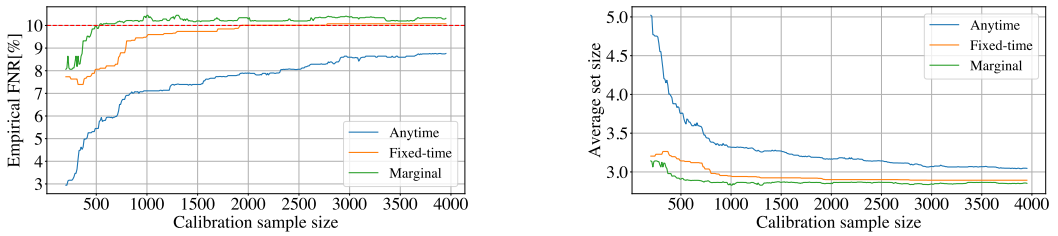
(a) Fixed-time valid

(b) Anytime-valid

Figure 6: FNR of prediction sets versus calibration sample size n, for the synthetic multilabel classification experiment with $\alpha = 10\%$, $\delta = 10\%$.

False Negative Rate Control in Multilabel Classification. We use the multilabel classification setup of Angelopoulos & Bates (2022) (Section 5.1). The data consist of images with multiple binary labels, and the prediction set is constructed as $\mathcal{C}_\lambda(X) = \{k : f_k(X) \geq 1 - \lambda\}$, where $f_k(X) \in [0, 1]$ is the estimated probability of label k. We target FNR control at level $\alpha = 10\%$ with $\delta = 10\%$.

Figure 7a shows the realized FNR of the sequence of prediction sets constructed using Theorem 4.1, compared to the fixed-time WSR bound of Bates et al. (2021) and marginal risk control (Angelopoulos et al., 2024). Figure 7b shows the corresponding average prediction set sizes.



(a)

(b)

Figure 7: Comparison of anytime-valid risk control using Theorem 4.1, fixed-time valid risk control using the method of Bates et al. (2021), and marginal risk control for multilabel image classification with $\alpha = 10\%$, $\delta = 10\%$.

6 Conclusion

We proposed an anytime-valid conformal risk control framework that guarantees, with high probability, that the risk of a sequence of prediction sets remains below a target level at all times while converging to it asymptotically. This is particularly useful in problems with cumulative data collection processes. The approach yields interpretable correction terms, tightness guarantees together with a proof that the rate of the correction is optimal, and extensions to test-time distribution shift. The framework applies to general bounded loss functions, including miscoverage, false negative rate, and others, as validated by our theoretical analysis and experiments.

A natural limitation is that the method is more conservative at small sample sizes, requiring additional data before producing informative prediction sets. Another limitation is that the distribution-shift setting requires knowledge of the importance weights $\omega(x, y)$; handling misspecified weights remains an open direction. Furthermore, future work includes extensions to online model updates as well as time-series models with uncertainty updates. This is an interesting direction as anytime-valid guarantees are especially relevant for models that need to run over extended periods in online settings.

Acknowledgments and Disclosure of Funding

This work was partially supported by the Swedish Research Council (contract number 2024-03903). AHR is supported by eSSENCE, SciLifeLab, and the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. https://arxiv.org/abs/2107.07511.

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

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *International Conference on Learning Representations (ICLR)*, 2024. https://openreview.net/forum?id=33XGfHLtZg.

Angelopoulos, A. N., Barber, R. F., and Bates, S. Theoretical foundations of conformal prediction, 2025. https://arxiv.org/abs/2411.11824.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021. https://doi.org/10.1145/3478535.

Darling, D. A. and Robbins, H. Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences of the United States of America*, 57(5):1188–1192, 1967. http://www.jstor.org/stable/57876.

Duchi, J. Sample-conditional coverage in split-conformal prediction. In *Advances in Neural Information Processing Systems*, 2025. https://openreview.net/forum?id=aTBM5j3oyA.

Duchi, J. and Haque, S. An information-theoretic lower bound in time-uniform estimation. *Conference on Learning Theory*, 2024. https://proceedings.mlr.press/v247/duchi24a.html.

Durrett, R. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.

Gauthier, E., Bach, F., and Jordan, M. I. E-Values Expand the Scope of Conformal Prediction. *arXiv preprint arXiv:2503.13050*, 2025. https://arxiv.org/abs/2503.13050.

Ge, H., Bastani, H., and Bastani, O. Stochastic online conformal prediction with semi-bandit feedback. *Forty-second International Conference on Machine Learning (ICML)*, 2025. https://openreview.net/forum?id=IdRrKDsTZ8.

Gibbs, I. and Candès, E. Adaptive Conformal Inference Under Distribution Shift. *Advances in Neural Information Processing Systems*, 2021. https://openreview.net/forum?id=6vaActvpcp3.

Howard, S. R. and Ramdas, A. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704–1728, 2022. https://doi.org/10.3150/21-BEJ1388.

Howard, S. R., Ramdas, A., McAuliffe, J. D., and Sekhon, J. S. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. https://doi.org/10.1214/18-PS321.

Hulsman, R. Distribution-free finite-sample guarantees and split conformal prediction. *arXiv:2210.14735*, 2022.

Marques F., P. C. Universal distribution of the empirical coverage in split conformal prediction. *Statistics & Probability Letters*, 219:110350, 2025. https://doi.org/10.1016/j.spl.2024.110350.

Pournaderi, M. and Xiang, Y. Training-conditional coverage bounds under covariate shift. *Transactions on Machine Learning Research*, 2026. https://openreview.net/forum?id=F6hHT3qWxT.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4), 576–601, 2023. https://doi.org/10.1214/23-STS894.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. https://doi.org/10.1016/S0378-3758(00)00115-4.

Vovk, V. Conditional validity of inductive conformal predictors. *Asian Conference on Machine Learning*, 2012. https://proceedings.mlr.press/v25/vovk12.html.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer, 2005.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. https://doi.org/10.1093/jrsssb/qkad009.

Waudby-Smith, I., Wu, L., Ramdas, A., Karampatziakis, N., and Mineiro, P. Anytime-valid off-policy inference for contextual bandits. *ACM / IMS J. Data Sci.*, 1(3), 2024. https://doi.org/10.1145/3643693.

Xu, Z., Karampatziakis, N., and Mineiro, P. Active, anytime-valid risk controlling prediction sets. *Advances in Neural Information Processing Systems*, 2024. https://openreview.net/forum?id=4ZH48aGD60.

Zwart, P. H. Probabilistic conformal coverage guarantees in small-data settings, 2025. https://arxiv.org/abs/2509.15349.

A Background: Anytime-Valid Concentration Inequalities

A short introduction. We briefly recall the main ideas behind time-uniform concentration; see Howard et al. (2020, 2021) for a detailed treatment and Ramdas et al. (2023) for an accessible overview. A process $(L_t)_{t\geq 0}$ adapted to a filtration $(\mathcal{F}_t)_{t\geq 0}$ is a *supermartingale* if $\mathbb{E}[L_t \mid \mathcal{F}_{t-1}] \leq L_{t-1}$ for all $t \geq 1$, i.e., it does not increase in conditional expectation. It is commonly viewed as the wealth of a gambler in a game that is fair or unfavorable to them. The key tool is Ville's inequality. If (L_t) is a nonnegative supermartingale with $\mathbb{E}[L_0] \leq 1$, then

$$\mathbb{P}\big(\exists t \geq 0 : \ L_t \geq 1/\delta\big) \leq \delta.$$

This is a time-uniform analogue of Markov's inequality, which only yields $\mathbb{P}(L_t \geq 1/\delta) \leq \delta$ for a single, fixed t.

For a sum $M_t = \sum_{i=1}^t \xi_i$ of independent mean-zero increments, a typical nonnegative supermartingale is of the form $L_t(\tau) = \exp(\tau M_t - \psi(\tau)V_t)$, where $\tau > 0$ is a free parameter, V_t is a variance process, and $\psi(\tau)V_t$ upper bounds the cumulant generating function of τM_t. Definition A.1 below formalizes this condition. Ville's inequality then yields that, with probability at least $1-\delta$, $M_t < \log(1/\delta)/\tau + \psi(\tau)V_t/\tau$ for all t simultaneously. This boundary is linear in V_t, and a fixed τ makes it tight only for V_t near a single value. *Stitching* (Howard et al., 2021) overcomes this by splitting the range of V_t into geometrically growing epochs $[\eta^k m, \eta^{k+1} m)$, $k = 0, 1, \ldots$, using a separate parameter τ_k tuned to epoch k with error budget $\delta/h(k)$, and taking a union bound over the epochs. Since $\sum_k 1/h(k) \leq 1$, the total error probability is at most δ, and the resulting boundary grows as $\sqrt{V_t \log\log V_t}$, in line with the law of the iterated logarithm. Theorem A.2 below states the result.

Definition A.1 (Sub-ψ condition, Definition 1 Howard et al. (2020))**.** Let $(M_t)_{t=0}^\infty, (V_t)_{t=0}^\infty$ be real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t=0}^\infty$ with $M_0 = V_0 = 0$ and $V_t \geq 0$ for all t. For $c \in \mathbb{R}$, define

$$\psi_{G,c}(\tau) = \frac{\tau^2}{2(1-c\tau)}, \quad \text{for} \quad 0 \leq \tau < \tau_{\max},$$

where $\tau_{\max} = 1/\max\{0, c\}$ with the convention that $1/0 = \infty$. We say that (M_t) is a *sub-gamma process* with scale parameter $c \in \mathbb{R}$ and variance process (V_t) if, for each $\tau \in [0, \tau_{\max})$, there exists a supermartingale $(L_t(\tau))_{t=0}^\infty$ w.r.t. $(\mathcal{F}_t)$ such that $\mathbb{E}[L_0(\tau)] \leq 1$ and

$$\exp\left(\tau M_t - \psi_{G,c}(\tau)V_t\right) \leq L_t(\tau) \text{ a.s. for all } t.$$

Howard et al. (2021) prove the following Theorem.

Theorem A.2 (Stitched boundary, Theorem 1 Howard et al. (2021))**.** *Let (M_t) be a sub-gamma process with scale parameter $c \geq 0$ and variance process (V_t). For any $\delta \in (0,1), \eta > 1, m > 0$, and $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ increasing such that $\sum_{k=0}^\infty 1/h(k) \leq 1$, define*

$$\mathcal{S}_\delta(v) = \sqrt{k_1^2 v \cdot d(v) + k_2^2 c^2 \cdot d^2(v)} + k_2 c \cdot d(v), \text{ where } \begin{cases} d(v) = \log h(\log_\eta(\frac{v}{m})) + \log(\frac{l_0}{\delta}), \\ k_1 = (\eta^{1/4} + \eta^{-1/4})/\sqrt{2}, \\ k_2 = (\sqrt{\eta}+1)/2. \end{cases} \tag{11}$$

Then,

$$\mathbb{P}(\forall t \geq 1 : M_t < \mathcal{S}_\delta(\max\{V_t, m\})) \geq 1 - \delta.$$

As a direct corollary of the above Theorem, taking $\eta = 2$, $l_0 = 1$, and $h(k) = (k+1)^2\pi^2/6$. With $\eta = 2$, Theorem A.2 gives $k_1 = 2^{-1/4} + 2^{-3/4} \approx 1.4355$ and $k_2 = (\sqrt{2}+1)/2 \approx 1.2071$, and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ yields $\mathcal{S}_\delta(v) \leq k_1\sqrt{v\, d(v)} + 2k_2\, c\, d(v)$. The constants 1.44 and 2.42 are k_1 and $2k_2$ rounded up for readability, we obtain the following explicit result.

Corollary A.3. *Define*

$$d(v) = 2\log\left(\log_2(v/m) + 1\right) + \log\frac{\pi^2}{6\delta}.$$

Let (M_t) be a sub-gamma process with scale parameter $c \geq 0$ and variance process (V_t). Then, for any $\delta \in (0,1)$, and $m > 0$,

$$\mathbb{P}(\forall t \geq 1 : M_t < \tilde{\mathcal{S}}_\delta(\max\{V_t, m\})) \geq 1 - \delta.$$

with

$$\tilde{\mathcal{S}}_\delta(v) = 1.44\sqrt{v\, d(v)} \;+\; 2.42\, c\, d(v).$$

Remark A.4 (Alternative constructions of correction terms). Our construction of the correction terms is based on Corollary A.3, which in turn relies on Theorem A.2 instantiated with specific parameter choices. Theorem A.2 depends on the parameters η and h, which govern a trade-off between finite-sample constants and asymptotic tightness. Smaller values of η and more slowly growing functions h yield tighter asymptotic behavior, at the cost of larger additive constants. Conversely, larger values lead to looser asymptotic constants but improved behavior at smaller calibration sets. We refer the interested reader to Howard et al. (2021) for further discussion.

While Corollary A.3 provides a time-uniform bound on (M_t), we will also need a fixed-time tail bound for a sub-gamma process. The following lemma provides this via a Markov inequality argument.

Lemma A.5. *Let $c \geq 0$. Let $(M_t)_{t\in\mathbb{N}}$ be sub-gamma with scale parameter c and deterministic variance process $(V_t)_{t\in\mathbb{N}}$. Then, for all $n \in \mathbb{N}$, such that $V_n > 0$, and $\delta \in (0,1)$,*

$$\mathbb{P}\left(M_n \geq c\,\log\frac{1}{\delta} + \sqrt{2V_n\,\log\frac{1}{\delta} + \left(c\,\log\frac{1}{\delta}\right)^2}\right) \leq \delta.$$

Proof. For any $0 \leq \tau < 1/c$ and $x \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{P}(M_n \geq x) &= \mathbb{P}\big(\exp(\tau M_n - \psi_{G,c}(\tau)V_n) \geq e^{\tau x - V_n\psi_{G,c}(\tau)}\big)\\ &\leq \exp\Big(-\tau x + V_n\psi_{G,c}(\tau)\Big),\end{aligned}$$

by Markov's inequality. Set $\tau = \frac{x}{V_n + cx} \in [0, 1/c)$ and set the RHS equal to δ, that is

$$\exp\Big(-\frac{x^2}{2(V_n + cx)}\Big) = \delta.$$

Then,

$$x = c\,\log\frac{1}{\delta} + \sqrt{2V_n\,\log\frac{1}{\delta} + \left(c\,\log\frac{1}{\delta}\right)^2}.$$

□

Lemma A.6 (Proposition S2 Howard et al. (2021))**.** *Let $(M_t)_{t\in\mathbb{N}}$ be sub-gamma with scale parameter $c = 0$ and possibly random variance process $(V_t)_{t\in\mathbb{N}}$. Then, for all $\rho > 0$ and $\delta \in (0,1)$,*

$$\mathbb{P}\big(\exists t \geq 1 : M_t \geq u_\rho(V_t)\big) \leq \delta, \qquad u_\rho(v) = \sqrt{2(v+\rho)\log\Big(\frac{1}{2\delta}\sqrt{\frac{v+\rho}{\rho}} + 1\Big)}.$$

In particular, $\mathbb{P}\big(M_n \geq u_\rho(V_n)\big) \leq \delta$ for every fixed $n \in \mathbb{N}$.

B Derivations

For notational convenience, we denote the loss associated with the prediction set $\mathcal{C}_\lambda(x)$ by

$$\ell_\lambda(x, y) := \ell\left(\mathcal{C}_\lambda(x), y\right).$$

The population risk is defined as $R(\lambda) = \mathbb{E}[\ell_\lambda(X, Y)]$. To construct suitable sub-gamma processes, we will rely on certain properties of R, and we first address the additional considerations that arise when R is discontinuous. By assumption, $\ell_\lambda(X, Y)$ is right-continuous in λ and bounded. Hence, by the Dominated Convergence Theorem, R is right-continuous in λ. Furthermore, R is monotonically decreasing. For $\alpha \in [0, \sup_{\lambda\in\Lambda} R(\lambda)]$, define

$$\hat{\lambda}(\alpha) := \inf\{\lambda : R(\lambda) \leq \alpha\}.$$

If $\alpha \geq \sup_{\lambda\in\Lambda} R(\lambda)$, every threshold has risk at most α, and the upper bounds hold trivially. We also define the left-limit of R at $\hat{\lambda}(\alpha)$ by

$$R(\hat{\lambda}(\alpha)^-) := \lim_{\lambda\uparrow\hat{\lambda}(\alpha)} R(\lambda),$$

and, similarly, the corresponding left-limit of the loss function by

$$\ell_{\hat{\lambda}(\alpha)^-}(x,y) := \lim_{\lambda\uparrow\hat{\lambda}(\alpha)} \ell_\lambda(x,y).$$

Since $\ell_\lambda(x,y)$ is monotonically decreasing in λ and bounded, it follows from the Dominated Convergence Theorem that

$$\begin{aligned} R(\hat{\lambda}(\alpha)^-) &= \lim_{\lambda\uparrow\hat{\lambda}(\alpha)} \mathbb{E}[\ell_\lambda(X,Y)] = \mathbb{E}\Big[\lim_{\lambda\uparrow\hat{\lambda}(\alpha)} \ell_\lambda(X,Y)\Big] \\ &= \mathbb{E}[\ell_{\hat{\lambda}(\alpha)^-}(X,Y)]. \end{aligned}$$

By monotonicity and right-continuity of R, we have

$$R(\hat{\lambda}(\alpha)) \leq \alpha \leq R(\hat{\lambda}(\alpha)^-).$$

The proofs require increments with mean exactly zero at the oracle threshold, i.e., $\mathbb{E}[\ell_{\hat{\lambda}(\alpha)}(X,Y)] = \alpha$. When R is continuous at $\hat{\lambda}(\alpha)$, this holds automatically. However, R may jump across the level α at $\hat{\lambda}(\alpha)$, so that $R(\hat{\lambda}(\alpha)) < \alpha \leq R(\hat{\lambda}(\alpha)^-)$ and no threshold attains the level α exactly. The weight $\pi(\alpha)$ defined below is the interpolation coefficient that splits this jump. It defines an effective loss, a convex combination of the loss at $\hat{\lambda}(\alpha)$ and its left-limit, whose mean is exactly α.

Define

$$\pi(\alpha) := \begin{cases} 1, & \text{if } R(\hat{\lambda}(\alpha)) = \alpha, \\ \dfrac{R(\hat{\lambda}(\alpha)^-) - \alpha}{R(\hat{\lambda}(\alpha)^-) - R(\hat{\lambda}(\alpha))}, & \text{if } R(\hat{\lambda}(\alpha)) < \alpha. \end{cases}$$

Then $\pi(\alpha) \in [0,1]$, and

$$\alpha = \pi(\alpha)\, R(\hat{\lambda}(\alpha)) + (1-\pi(\alpha))\, R(\hat{\lambda}(\alpha)^-).$$

Equivalently,

$$\mathbb{E}\Big[\pi(\alpha)\, \ell_{\hat{\lambda}(\alpha)}(X,Y) + (1-\pi(\alpha))\, \ell_{\hat{\lambda}(\alpha)^-}(X,Y)\Big] = \alpha.$$

To establish our main results, we construct suitable processes $(M_t)_{t=0}^\infty$ and $(V_t)_{t=0}^\infty$ that satisfy a sub-gamma condition, allowing us to apply Corollary A.3.

Theorem 4.1. *Let ℓ be an arbitrary loss function bounded in $[0,B]$, monotone in λ, and right-continuous in λ. Let $c_\alpha = \max\{(2\alpha - B)/3, 0\}$. Construct the prediction sets according to (6) and set the correction term as*

$$\gamma_n = f_{c_\alpha, \alpha(B-\alpha)m^*, \delta}\big(\alpha(B-\alpha)n\big)/n, \tag{10}$$

where

$$m^* \;=\; \min\Big\{ m' \in \mathbb{N} : f_{c_\alpha, \alpha(B-\alpha)m', \delta}\big(\alpha(B-\alpha)m'\big)/m' \leq \alpha \Big\}.$$

Then the resulting sequence of prediction sets $\{\mathcal{C}_{\lambda_n}(X)\}$ achieves anytime-valid risk control (9).

Proof of Theorem 4.1. For the growing calibration sequence $\{(X_i,Y_i)\}_{i\geq 1}$, define

$$\xi_i(\alpha) = \pi(\alpha)\ell_{\hat{\lambda}(\alpha)}(X_i,Y_i) + (1-\pi(\alpha))\ell_{\hat{\lambda}(\alpha)^-}(X_i,Y_i),$$

and let

$$M_t(\alpha) = \sum_{i=1}^{t} (\alpha - \xi_i(\alpha)).$$

By construction, $\mathbb{E}[\alpha - \xi_i(\alpha)] = 0$ and $(\alpha - \xi_i(\alpha)) \in [\alpha - B, \alpha]$. First, note that these are sufficient conditions for a sub-Bernoulli process Howard et al. (2020), (Table 3, row "Bernoulli 1"). Applying Proposition 2 of Howard et al. (2020) (Table 5, rows 3 and 5) then yields that $(M_t(\alpha))_{t\in\mathbb{N}}$ is sub-gamma with scale parameter $c_\alpha = \max\left\{\frac{2\alpha - B}{3}, 0\right\}$, where we used monotonicity of the scale parameter, and variance process $(V_t)_{t\in\mathbb{N}}$ given by

$$V_t = \alpha(B - \alpha)t. \tag{12}$$

By Corollary A.3, with $c = c_\alpha$ and $\tilde{m} = \alpha(B - \alpha)m^*$, we obtain

$$\mathbb{P}\left(\forall t \in \mathbb{N} : M_t(\alpha) < f_{c_\alpha, \tilde{m}, \delta}(\max\{\tilde{m}, V_t\})\right) \geq 1 - \delta.$$

Rewriting, yields, with probability at least $1 - \delta$, for all t,

$$\alpha - \frac{f_{c_\alpha, \tilde{m}, \delta}(\max\{\tilde{m}, V_t\})}{t} < \frac{1}{t}\sum_{i=1}^{t} \ell_{\hat{\lambda}(\alpha)^-}(X_i, Y_i).$$

Thus, assuming the above is true, for $t \geq m^*$ we have $V_t \geq \tilde{m}$, so that $f_{c_\alpha, \tilde{m}, \delta}(\max\{\tilde{m}, V_t\}) = t\gamma_t$ with γ_t as in (10), and

$$\begin{aligned}
\lambda_t &= \inf\left\{\lambda : R_t(\lambda) \leq \alpha - \frac{f_{c_\alpha, \tilde{m}, \delta}(\max\{\tilde{m}, V_t\})}{t}\right\} \\
&\geq \inf\left\{\lambda : R_t(\lambda) < \frac{1}{t}\sum_{i=1}^{t} \ell_{\hat{\lambda}(\alpha)^-}(X_i, Y_i)\right\} \\
&\geq \hat{\lambda}(\alpha),
\end{aligned}$$

by monotonicity of the loss. For $t < m^*$, we have $\gamma_t > \alpha$ by definition of m^*, so $\lambda_t = \lambda_{\max} \geq \hat{\lambda}(\alpha)$. Hence,

$$\mathbb{P}\left(\forall t \in \mathbb{N} : \lambda_t \geq \hat{\lambda}(\alpha)\right) \geq 1 - \delta.$$

□

Proposition 4.6. *In the setting of Theorem 4.1, let $\{\lambda_n\}$ be non-increasing (using a running minimum if necessary) and assume that $\alpha \leq B/2$ and $\alpha \leq \sup_{\lambda\in\Lambda} R(\lambda)$, where $R(\lambda) = \mathbb{E}[\ell(\mathcal{C}_\lambda(X), Y)]$. Furthermore, let $d_n = \sup_{\lambda\in\Lambda} J(R_n, \lambda)$. If d_n converges to 0, then the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ is asymptotically tight with the risk approaching α at the rate $O\left(\max\left\{d_n, B\sqrt{n^{-1}\log\log n}\right\}\right)$. In particular, the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ satisfies*

$$\mathbb{P}\big(\forall n : \alpha - k_n \leq \mathbb{E}[\ell\left(\mathcal{C}_{\lambda_n}(X), Y\right) | \mathcal{D}_n] \leq \alpha\big) \geq 1 - 2\delta,$$

where $k_n = d_{g(n)} + \gamma_{g(n)} + \gamma'_{g(n)}(\delta_{\lfloor \log_2 n\rfloor + 1})$, $g(n) = 2^{\lfloor \log_2 n\rfloor}$, $\delta_i = \delta/(i(i+1))$, and

$$\gamma'_n(\delta) = \frac{B}{n}\log\frac{1}{\delta} + \sqrt{2\frac{\alpha(B-\alpha)}{n}\log\frac{1}{\delta} + \left(\frac{B}{n}\log\frac{1}{\delta}\right)^2}.$$

To prove Proposition 4.6, the key idea is to apply Lemma A.5 to a sub-gamma process constructed at a fixed threshold λ, combined with a union bound over a dyadic grid.

Proof of Proposition 4.6. By Theorem 4.1, the upper bound holds for all n with probability at least $1 - \delta$. Taking a running minimum does not affect this, since it preserves $\lambda_n \geq \hat{\lambda}(\alpha)$. As $\mathbb{E}[\ell(\mathcal{C}_{\lambda_n}(X), Y) \mid \mathcal{D}_n] = R(\lambda_n)$, it remains to show that, with probability at least $1 - \delta$, $R(\lambda_n) \geq \alpha - k_n$ for all n.

Fix $i \geq 0$, let $N = 2^i$, and set $\beta_i = \alpha - \gamma_N - \gamma'_N(\delta_{i+1})$. If $\beta_i \leq 0$, or if $R(\lambda) > \beta_i$ for all λ, then $R(\lambda_N) \geq \beta_i$ trivially. Otherwise, let $\check{\lambda}_i = \inf\{\lambda : R(\lambda) \leq \beta_i\}$, a deterministic threshold. By

right-continuity, $R(\check{\lambda}_i) \le \beta_i$, and since $\sup_\lambda R(\lambda) \ge \alpha > \beta_i$, every $\lambda < \check{\lambda}_i$ satisfies $R(\lambda) > \beta_i$, so that $R(\check{\lambda}_i^-) \ge \beta_i$.

Let $\xi_j = \ell_{\check{\lambda}_i}(X_j, Y_j) - R(\check{\lambda}_i)$. These are i.i.d., have mean zero, and take values in $[-R(\check{\lambda}_i), B - R(\check{\lambda}_i)]$. Hence, as in the proof of Theorem 4.1, $\sum_{j \le t} \xi_j$ is sub-gamma with scale at most B and variance process $R(\check{\lambda}_i)(B - R(\check{\lambda}_i))\, t \le \alpha(B-\alpha)t$. Lemma A.5 at time N thus shows that the event

$$E_i = \left\{ R_N(\check{\lambda}_i) < R(\check{\lambda}_i) + \gamma'_N(\delta_{i+1}) \right\}$$

has probability at least $1 - \delta_{i+1}$. On E_i, $R_N(\check{\lambda}_i) < \beta_i + \gamma'_N(\delta_{i+1}) = \alpha - \gamma_N$, so $\check{\lambda}_i$ is feasible in (6) at sample size N, and hence $\lambda_N \le \check{\lambda}_i$, as the running minimum can only decrease λ_N. We distinguish two cases.

If $\lambda_N < \check{\lambda}_i$, then $R(\lambda_N) \ge R(\check{\lambda}_i^-) \ge \beta_i$.

If $\lambda_N = \check{\lambda}_i$, then every $\lambda < \lambda_N$ is infeasible at sample size N (it lies below the infimum in (6)), so $R_N(\lambda) > \alpha - \gamma_N$ for all $\lambda < \lambda_N$ and $R_N(\lambda_N^-) \ge \alpha - \gamma_N$. Hence, $R_N(\check{\lambda}_i) = R_N(\lambda_N^-) - J(R_N, \lambda_N) \ge \alpha - \gamma_N - d_N$, and, on E_i,

$$R(\lambda_N) = R(\check{\lambda}_i) > R_N(\check{\lambda}_i) - \gamma'_N(\delta_{i+1}) \ge \alpha - \gamma_N - \gamma'_N(\delta_{i+1}) - d_N.$$

In both cases, $R(\lambda_N) \ge \alpha - k_N$.

Since $\sum_{i \ge 0} \delta_{i+1} = \delta \sum_{i \ge 1} \frac{1}{i(i+1)} = \delta$, a union bound shows that, with probability at least $1 - \delta$, $R(\lambda_{2^i}) \ge \alpha - k_{2^i}$ for all $i \ge 0$. For $2^i \le n < 2^{i+1}$, we have $g(n) = 2^i$, $k_n = k_{2^i}$, and $\lambda_n \le \lambda_{2^i}$, hence $R(\lambda_n) \ge R(\lambda_{2^i}) \ge \alpha - k_n$. The rate follows because $\gamma_{g(n)}$ and $\gamma'_{g(n)}(\delta_{\lfloor \log_2 n \rfloor + 1})$ are of order $B\sqrt{n^{-1} \log\log n}$. Since only $J(R_N, \lambda_N) \le d_N$ was used, d_N may be replaced by $J(R_N, \lambda_N)$.

□

Before proving Theorem 4.8, we note the following fact, which follows from a Taylor expansion argument.

Lemma B.1. *Let $x \ge 0$, then*

$$\exp\left(x - \frac{x^2}{2}\right) \le 1 + x.$$

Theorem 4.8. *Let ℓ be an arbitrary loss function bounded in $[0, B]$, monotone in λ, and right-continuous in λ. Define $b_{m,\delta}(v) = 1.44\sqrt{v h_{m,\delta}(v)}$, i.e., $b_{m,\delta} = f_{0,m,\delta}$. Define the importance weights $\omega_i := \omega(X_i, Y_i)$ using (5) and $W_n = \sum_{i=1}^n \omega_i^2$, and fix $m > 0$ before observing the calibration data. Construct the prediction sets according to (6), using the importance-weighted loss $\omega(x,y)\,\ell(\mathcal{C}_\lambda(x), y)$, i.e., with empirical risk $R_n^\omega(\lambda) = \frac{1}{n}\sum_{i=1}^n \omega_i\, \ell(\mathcal{C}_\lambda(X_i), Y_i)$, with the correction term as*

$$\gamma_n = B\left(1 - \frac{1}{n}\sum_{i=1}^n \omega_i\right) + \frac{b_{m,\delta}(\max\{B^2 W_n, m\})}{n}.$$

Then the resulting sequence of prediction sets $\{\mathcal{C}_{\lambda_n}(X)\}$ achieves anytime-valid risk control (9) under distribution shift (4). That is, with probability at least $1 - \delta$, we obtain a sequence of prediction sets such that

$$\mathbb{E}_{P^*}[\ell(\mathcal{C}_{\lambda_n}(X), Y) \mid \mathcal{D}_n] \;\le\; \alpha$$

holds for all n.

Proof of Theorem 4.8. We proceed similarly as the proof of Theorem 4.1, where R, $\hat{\lambda}(\alpha)$, and $\pi(\alpha)$ now refer to the test-time risk $R(\lambda) = \mathbb{E}_{P^*}[\ell_\lambda(X, Y)]$. For simplicity, we assume that R is continuous. If this is not the case, then as in the proof of Theorem 4.1, we may replace each occurrence of $\ell_{\hat{\lambda}(\alpha)}$ with

$$\left(\pi(\alpha)\ell_{\hat{\lambda}(\alpha)}(X_i, Y_i) + (1 - \pi(\alpha))\ell_{\hat{\lambda}(\alpha)^-}(X_i, Y_i)\right).$$

Define

$$\xi_i(\alpha) = \omega(X_i, Y_i)\left(B - \ell_{\hat{\lambda}(\alpha)}(X_i, Y_i)\right)$$

and

$$\mu(\alpha) = \mathbb{E}\left[\xi_i(\alpha)\right] = B - \alpha.$$

Set

$$M_t(\alpha) = \sum_{i=1}^{t} \left(\xi_i(\alpha) - \mu(\alpha)\right).$$

For any $\tau \geq 0$, note that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\tau(\xi_i(\alpha) - \mu(\alpha)) - \frac{\tau^2 \xi_i(\alpha)^2}{2}\right)\right] \\
&= \exp(-\tau\mu(\alpha))\mathbb{E}\left[\exp\left(\tau\xi_i(\alpha) - \frac{\tau^2 \xi_i(\alpha)^2}{2}\right)\right] \\
&\leq \exp(-\tau\mu(\alpha))\mathbb{E}\left[1 + \tau\xi_i(\alpha)\right] \\
&= \exp(-\tau\mu(\alpha))(1 + \tau\mu(\alpha)) \\
&\leq \exp(-\tau\mu(\alpha))\exp(\tau\mu(\alpha)) = 1
\end{aligned}$$

due to Lemma B.1, since $\xi_i(\alpha)\tau \geq 0$, and the fact that $1 + x \leq \exp(x)$, for all $x \in \mathbb{R}$.

Hence, $(M_t(\alpha))_{t\in\mathbb{N}}$ is sub-gamma with scale parameter 0 and variance process $(V_t)_{t\in\mathbb{N}}$ given by

$$V_t = \sum_{i=1}^{t} \omega_i^2 \left(B - \ell_{\hat{\lambda}(\alpha)}(X_i, Y_i)\right)^2 \leq B^2 W_t.$$

By Corollary A.3, with $c = 0$ and the fixed m, we obtain,

$$\mathbb{P}\left(\forall t \in \mathbb{N} : M_t(\alpha) < b_{m,\delta}(\max\{m, B^2 W_t\})\right) \geq 1 - \delta.$$

Rewriting, yields

$$\mathbb{P}\left(\forall t \in \mathbb{N} : \alpha - \gamma_t < R_t^{\omega}(\hat{\lambda}(\alpha))\right) \geq 1 - \delta,$$

for

$$\gamma_t = \frac{b_{m,\delta}(\max\{m, B^2 W_t\})}{t} + B\left(1 - \frac{1}{t}\sum_{i=1}^{t} \omega_i\right).$$

The result follows by monotonicity and right-continuity of R, exactly as in the proof of Theorem 4.1. □

Remark B.2. In the proofs of Theorems 4.1 and 4.8, we apply Corollary A.3 to processes of different nature. In the i.i.d. setting, the increments $\alpha - \xi_i(\alpha)$ have mean zero and lie in $[\alpha - B, \alpha]$. Only their upper tail matters, and it is bounded by α. For $\alpha \leq B/2$, this asymmetry makes the process sub-Gaussian with variance process $\alpha(B - \alpha)t$, so no range term is needed ($c_\alpha = 0$). Under distribution shift, the increments $\xi_i(\alpha) = \omega_i(B - \ell_{\hat{\lambda}(\alpha)}(X_i, Y_i))$ are nonnegative but unbounded, since they involve the importance weights. Without assumptions on ω_i, we control them through their observed squares (self-normalization), which yields the variance process $\sum_{i\leq t} \xi_i(\alpha)^2 \leq B^2 W_t$.

Proposition 4.11. *In the distribution-shift setting of Theorem 4.8, let $\{\lambda_n\}$ be non-increasing (using a running minimum if necessary) and assume that $\alpha \leq \sup_{\lambda\in\Lambda} \mathbb{E}_{P^*}[\ell(\mathcal{C}_\lambda(X), Y)]$. Let $d_n = \sup_{\lambda\in\Lambda} J(R_n^{\omega}, \lambda)$. If d_n converges to 0 and $\sqrt{\frac{W_n \log\log W_n}{n^2}} \to 0$, where the latter holds almost surely whenever $\mathbb{E}_P[\omega^2] < \infty$, then the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ is asymptotically tight. In particular, the sequence $\{\mathcal{C}_{\lambda_n}(X)\}$ satisfies*

$$\mathbb{P}\big(\forall n : \alpha - k_n \leq \mathbb{E}_{P^*}[\ell\left(\mathcal{C}_{\lambda_n}(X), Y\right) | \mathcal{D}_n] \leq \alpha\big) \geq 1 - 2\delta,$$

where $k_n = d_{g(n)} + \gamma_{g(n)} + \gamma'_{g(n)}$, $g(n) = 2^{\lfloor \log_2 n \rfloor}$, and $\gamma'_n = 2^{j_n}/n - \gamma_n$, where

$$j_n = \min\left\{j \geq 0 : \frac{2^j}{n} \geq \gamma_n + \frac{b_{m,\delta_{n,j}}(\max\{B^2 W_n, m\})}{n}\right\}, \qquad \delta_{n,j} = \frac{\delta_{\lfloor \log_2 n \rfloor + 1}}{(j+1)(j+2)},$$

and $\delta_i = \frac{\delta}{i(i+1)}$.

Proof of Proposition 4.11. Write $R(\lambda) = \mathbb{E}_{P^*}[\ell_\lambda(X, Y)]$ for the test-time risk. By Theorem 4.8, the upper bound holds for all n with probability at least $1-\delta$. Since $\mathbb{E}_{P^*}[\ell(\mathcal{C}_{\lambda_n}(X), Y) \mid \mathcal{D}_n] = R(\lambda_n)$, it remains to show that, with probability at least $1-\delta$, $R(\lambda_n) \geq \alpha - k_n$ for all n. We follow the proof of Proposition 4.6, with two differences. The weighted losses may be unbounded, so their deviations are controlled by Corollary A.3, which allows for a random variance process and γ_N is random, so we use a grid of deterministic thresholds and a union bound over the grid.

Fix $i \geq 0$ and let $N = 2^i$, so that $\lfloor \log_2 N \rfloor + 1 = i + 1$ and $\delta_{N,j} = \delta_{i+1}/((j+1)(j+2))$. For $j \geq 0$, let $\beta_j = \alpha - 2^j/N$. Call j active if $\beta_j > 0$ and $R(\lambda) \leq \beta_j$ for some λ. For active j, let $\check{\lambda}_j = \inf\{\lambda : R(\lambda) \leq \beta_j\}$, a deterministic threshold. By right-continuity, $R(\check{\lambda}_j) \leq \beta_j$, and every $\lambda < \check{\lambda}_j$ satisfies $R(\lambda) > \beta_j$.

For active j, let $Z_l = \omega_l \ell_{\check{\lambda}_j}(X_l, Y_l) \geq 0$, $l \geq 1$. These are i.i.d. with $\mathbb{E}_P[Z_l] = R(\check{\lambda}_j)$. Applying Lemma B.1 with $x = \tau Z_l \geq 0$ shows, exactly as in the proof of Theorem 4.8, that $\sum_{l \leq t}(Z_l - R(\check{\lambda}_j))$ is sub-gamma with scale 0 and variance process $\sum_{l \leq t} Z_l^2 \leq B^2 W_t$. By Corollary A.3 with parameter m and error probability $\delta_{N,j}$, evaluated at time N, and since $b_{m,\delta}(\max\{\cdot, m\})$ is non-decreasing, the event

$$E_{i,j} = \left\{R_N^\omega(\check{\lambda}_j) < R(\check{\lambda}_j) + u_{N,j}\right\},$$

where $u_{N,j} = b_{m,\delta_{N,j}}(\max\{B^2 W_N, m\})/N$, has probability at least $1 - \delta_{N,j}$. Since $\sum_{j \geq 0} \frac{1}{(j+1)(j+2)} = 1$ and $\sum_{i \geq 0} \delta_{i+1} = \delta$, all events $E_{i,j}$, $i \geq 0$, j active, hold simultaneously with probability at least $1 - \delta$. We work on this event.

Let $j^* = j_N$. By the definition of γ'_N, we have $\gamma_N + \gamma'_N = 2^{j^*}/N$, so that

$$\beta_{j^*} = \alpha - 2^{j^*}/N = \alpha - \gamma_N - \gamma'_N = \alpha - k_N + d_N,$$

and, by the definition of j_N, $\gamma'_N = 2^{j^*}/N - \gamma_N \geq u_{N,j^*}$. If j^* is not active, then $R(\lambda_N) \geq \beta_{j^*} \geq \alpha - k_N$ trivially. Otherwise, on E_{i,j^*},

$$R_N^\omega(\check{\lambda}_{j^*}) < R(\check{\lambda}_{j^*}) + u_{N,j^*} \leq \beta_{j^*} + u_{N,j^*} \leq \beta_{j^*} + \gamma'_N = \alpha - \gamma_N,$$

so $\check{\lambda}_{j^*}$ is feasible in (6) at sample size N, and $\lambda_N \leq \check{\lambda}_{j^*}$, as the running minimum can only decrease λ_N. We distinguish two cases.

If $\lambda_N < \check{\lambda}_{j^*}$, then $R(\lambda_N) > \beta_{j^*} \geq \alpha - k_N$ by definition of $\check{\lambda}_{j^*}$.

If $\lambda_N = \check{\lambda}_{j^*}$, then every $\lambda < \lambda_N$ is infeasible at sample size N, so $R_N^\omega(\lambda_N^-) \geq \alpha - \gamma_N$ and $R_N^\omega(\check{\lambda}_{j^*}) = R_N^\omega(\lambda_N^-) - J(R_N^\omega, \lambda_N) \geq \alpha - \gamma_N - d_N$. On E_{i,j^*}, therefore,

$$R(\lambda_N) = R(\check{\lambda}_{j^*}) > R_N^\omega(\check{\lambda}_{j^*}) - u_{N,j^*} \geq \alpha - \gamma_N - u_{N,j^*} - d_N \geq \alpha - \gamma_N - \gamma'_N - d_N = \alpha - k_N.$$

In both cases, $R(\lambda_N) \geq \alpha - k_N$.

For $2^i \leq n < 2^{i+1}$, we have $g(n) = 2^i$, $k_n = k_{2^i}$, and $\lambda_n \leq \lambda_{2^i}$, hence $R(\lambda_n) \geq R(\lambda_{2^i}) \geq \alpha - k_n$. A union bound with the event of Theorem 4.8 gives probability at least $1 - 2\delta$.

Asymptotic tightness. Since $d_n \to 0$ by assumption, it suffices to show $\gamma_n \to 0$ and $\gamma'_n \to 0$. Write $v_n = \max\{B^2 W_n, m\}$. First, $\gamma_n = B(1 - \frac{1}{n}\sum_{l \leq n} \omega_l) + b_{m,\delta}(v_n)/n \to 0$, where the first term vanishes because $\frac{1}{n}\sum_{l \leq n} \omega_l \to 1$, and the second decays as $O(\sqrt{v_n \log\log v_n}/n) \to 0$.

Second, the grid is geometric, so rounding up to the next grid point at most doubles the distance. If $j_n \geq 1$, minimality of j_n gives $2^{j_n - 1}/n < \gamma_n + u_{n,j_n - 1} \leq \gamma_n + u_{n,j_n}$, since $u_{n,j}$ is non-decreasing in j ($\delta_{n,j}$ decreases in j), and hence $\gamma'_n = 2^{j_n}/n - \gamma_n < \gamma_n + 2u_{n,j_n}$. If $j_n = 0$, then $\gamma'_n = 1/n - \gamma_n$. Together with the definition of j_n, this gives $u_{n,j_n} \leq \gamma'_n \leq |\gamma_n| + 2u_{n,j_n} + 1/n$.

It remains to show $u_{n,j_n} \to 0$. For $j \leq \log_2 n + 2$ and $i = \lfloor \log_2 n \rfloor + 1$, we have $\log(1/\delta_{n,j}) = \log(1/\delta) + \log(i(i+1)) + \log((j+1)(j+2)) = O(\log\log n)$, so that $u_{n,j} = O\big(\sqrt{v_n(\log\log v_n + \log\log n)}/n\big)$ uniformly in such j. This tends to zero as $v_n \log\log v_n/n^2 \to 0$ by assumption, and $v_n \log\log n/n^2 \to 0$ because it is at most of order $v_n \log\log v_n/n^2$ if $v_n \geq n/2$, and at most $\log\log n/(2n)$ otherwise. In particular, $\bar{j} = \lceil \log_2 n \rceil + 1$ satisfies $2^{\bar{j}}/n \geq 2 \geq \gamma_n + u_{n,\bar{j}}$ for large n, so that $j_n \leq \bar{j}$, and hence $u_{n,j_n} \to 0$. Therefore, $\gamma'_n \to 0$ and $k_n \to 0$. □

Proof of Proposition 4.12. Since the miscoverage loss is $\{0,1\}$-valued, nonincreasing, and right-continuous in λ, we can write $\ell_\lambda(X_i, Y_i) = \mathbf{1}\{\lambda < T_i\}$ with $T_i = \inf\{\lambda \in \Lambda : Y_i \in \mathcal{C}_\lambda(X_i)\}$. Note that for $\mathcal{C}_\lambda(x) = \{y : s(x,y) \le \lambda\}$, $T_i = s(X_i, Y_i)$ is simply the score. Then $R(\lambda) = \mathbb{P}(T_1 > \lambda)$ and $R_n(\lambda) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{T_i > \lambda\}$. Write $\hat\lambda = \hat\lambda(\alpha)$.

By definition of $\hat\lambda$, $R(\lambda) > \alpha$ for all $\lambda < \hat\lambda$, and $R(\hat\lambda) \le \alpha$ by right-continuity. Since $\hat\lambda$ is an interior point of Λ and R is continuous at $\hat\lambda$, it follows that $R(\hat\lambda) = \alpha$ and $\mathbb{P}(T_1 = \hat\lambda) = R(\hat\lambda^-) - R(\hat\lambda) = 0$. Hence, almost surely, no T_i equals $\hat\lambda$.

Let $\xi_i = \ell_{\hat\lambda}(X_i, Y_i) - \alpha$. The ξ_i are i.i.d. with mean zero and variance $\sigma^2 = \alpha(1-\alpha) > 0$. By the law of the iterated logarithm (see, e.g., Theorem 8.5.2 in Durrett, 2019),

$$\liminf_{n\to\infty} \frac{\sum_{i=1}^n \xi_i}{\sqrt{2\sigma^2 n \log\log n}} = -1 \qquad \text{a.s.}$$

Fix $\epsilon \in (0,1)$. Almost surely, $\sum_{i=1}^n \xi_i < -(1-\epsilon)\sqrt{2\sigma^2 n \log\log n}$ for infinitely many n. Since $\gamma_n = o(\sqrt{n^{-1}\log\log n})$, we have $n\gamma_n < (1-\epsilon)\sqrt{2\sigma^2 n\log\log n}$ for all sufficiently large n. Consequently, almost surely,

$$R_n(\hat\lambda) = \alpha + \frac{1}{n}\sum_{i=1}^n \xi_i < \alpha - \gamma_n \qquad \text{for infinitely many } n.$$

Fix such an n and assume that no T_i equals $\hat\lambda$. As there are only finitely many T_i and $\hat\lambda$ is an interior point of the interval Λ, there exists $\lambda^\circ \in \Lambda$ with $\lambda^\circ < \hat\lambda$ such that no T_i lies in $[\lambda^\circ, \hat\lambda]$. Then $R_n(\lambda^\circ) = R_n(\hat\lambda) < \alpha - \gamma_n$, so λ° is feasible in (6), and therefore $\lambda_n \le \lambda^\circ < \hat\lambda$. By the first step, $R(\lambda_n) > \alpha$. Since λ_n is a function of $\mathcal{D}_n$ and the test point is independent of $\mathcal{D}_n$, we obtain $\mathbb{E}[\ell(\mathcal{C}_{\lambda_n}(X), Y) \mid \mathcal{D}_n] = R(\lambda_n) > \alpha$.

Hence, almost surely, the risk exceeds α for infinitely many n. □

C Fixed-Time Sample-Conditional Bounds under Covariate Shift

We derive a fixed-time analogue of Theorem 4.8, obtaining a sharper correction term at the cost of uniformity in n. Recall that $W_n = \sum_{i=1}^n \omega_i^2$ and that the importance-weighted empirical risk is

$$R_n^\omega(\lambda) = \frac{1}{n}\sum_{i=1}^n \omega_i\, \ell_\lambda(X_i, Y_i).$$

Theorem C.1 (Fixed-time sample-conditional risk control under covariate shift)**.** *Let ℓ be an arbitrary loss function bounded in $[0, B]$, monotone in λ, and right-continuous in λ. For fixed $n \in \mathbb{N}$ and $\delta \in (0,1)$, let $\rho > 0$ be chosen independently of the calibration data (it may depend on n), and construct λ_n via (6), with R_n replaced by R_n^ω, with correction term*

$$\gamma_n = B\left(1 - \frac{1}{n}\sum_{i=1}^n \omega_i\right) + \frac{u_\rho(B^2 W_n)}{n}, \qquad u_\rho(v) = \sqrt{2(v+\rho)\log\left(\frac{1}{2\delta}\sqrt{\frac{v+\rho}{\rho}} + 1\right)}. \tag{13}$$

Then

$$\mathbb{P}\big(\mathbb{E}_{P^*}\big[\ell(\mathcal{C}_{\lambda_n}(X), Y) \mid \mathcal{D}_n\big] \le \alpha\big) \ge 1 - \delta.$$

Remark C.2. If ρ does not depend on n, the guarantee holds simultaneously for all n (Lemma A.6). Choosing $\rho = B^2 n/(-W_{-1}(-4\delta^2/e) - 1)$, where W_{-1} is the lower branch of the Lambert W function, approximately minimizes the boundary at $B^2 W_n \approx B^2 n$ (Howard et al., 2021, Proposition 3). For $\delta = 10\%$, this gives $\rho \approx B^2 n/5$.

Proof of Theorem C.1. The proof is identical to that of Theorem 4.8, with Corollary A.3 replaced by Lemma A.6. With $M_t(\alpha)$ and $V_t \le B^2 W_t$ as defined there, Lemma A.6 and the fact that u_ρ is non-decreasing give

$$\mathbb{P}\big(\forall t \in \mathbb{N} : M_t(\alpha) < u_\rho(B^2 W_t)\big) \ge 1 - \delta.$$

The same rearrangement yields $\mathbb{P}\big(\forall t \in \mathbb{N} : \alpha - \gamma_t < R_t^\omega(\hat\lambda(\alpha))\big) \ge 1 - \delta$ with $\gamma_t = u_\rho(B^2 W_t)/t + B\big(1 - \frac{1}{t}\sum_{i=1}^t \omega_i\big)$, in particular at $t = n$. The result follows as in the proof of Theorem 4.8. □

Proposition C.3 (Lower bound, fixed-time under covariate shift)**.** *In the setting of Theorem C.1, assume that* $\alpha \le \sup_{\lambda\in\Lambda} R(\lambda)$*, where* $R(\lambda) = \mathbb{E}_{P^*}[\ell_\lambda(X,Y)]$*, and let* $d_n = \sup_{\lambda\in\Lambda} J(R_n^\omega, \lambda)$*. Let* $u_\rho(v;\delta')$ *denote* $u_\rho(v)$ *from* (13) *with* δ *replaced by* δ'*, and, for* $j \ge 0$*, let*

$$\delta_j = \frac{\delta}{(j+1)(j+2)}, \qquad u_j = \frac{u_\rho(B^2 W_n; \delta_j)}{n}, \qquad j_n = \min\left\{j \ge 0 : \frac{2^j}{\sqrt{n}} \ge \gamma_n + u_j\right\}.$$

Then, with probability at least $1 - 2\delta$*,*

$$\alpha - k_n \le \mathbb{E}_{P^*}\left[\ell(\mathcal{C}_{\lambda_n}(X), Y) \mid \mathcal{D}_n\right] \le \alpha, \qquad k_n = d_n + \gamma_n + \gamma_n', \qquad \gamma_n' = \frac{2^{j_n}}{\sqrt{n}} - \gamma_n,$$

where γ_n' *is computable from the calibration data.*

Proof. The upper bound holds with probability at least $1-\delta$ by Theorem C.1. For the lower bound, we follow the proof of Proposition 4.11 at the fixed sample size n, with the grid $\beta_j = \alpha - 2^j/\sqrt{n}$, $j \ge 0$, and with Lemma A.6 in place of Corollary A.3. Call j active if $\beta_j > 0$ and $R(\lambda) \le \beta_j$ for some λ, and, for active j, let $\check{\lambda}_j = \inf\{\lambda : R(\lambda) \le \beta_j\}$. Then $R(\check{\lambda}_j) \le \beta_j$, and every $\lambda < \check{\lambda}_j$ satisfies $R(\lambda) > \beta_j$.

For active j, the variables $Z_l = \omega_l \ell_{\check{\lambda}_j}(X_l, Y_l) \ge 0$ are i.i.d. with mean $R(\check{\lambda}_j)$, and, as in the proof of Theorem C.1, $\left(\exp(\tau \sum_{l\le t}(Z_l - R(\check{\lambda}_j)) - \tau^2 \sum_{l\le t} Z_l^2/2)\right)_{t\ge 0}$ is a supermartingale for every $\tau \ge 0$. Since $\sum_{l\le n} Z_l^2 \le B^2 W_n$ and $u_\rho(\cdot\,;\delta_j)$ is non-decreasing, Lemma A.6 with error probability δ_j shows that the event

$$E_j = \left\{R_n^\omega(\check{\lambda}_j) < R(\check{\lambda}_j) + u_j\right\}$$

has probability at least $1-\delta_j$. As $\sum_{j\ge 0}\delta_j = \delta$, all events E_j, j active, hold simultaneously with probability at least $1-\delta$. We work on this event.

Let $j^* = j_n$. By the definitions of γ_n' and j_n, we have $\beta_{j^*} = \alpha - \gamma_n - \gamma_n' = \alpha - k_n + d_n$ and $\gamma_n' \ge u_{j^*}$. If j^* is not active, then $R(\lambda_n) \ge \beta_{j^*} \ge \alpha - k_n$ trivially. Otherwise, on E_{j^*}, $R_n^\omega(\check{\lambda}_{j^*}) < \beta_{j^*} + u_{j^*} \le \beta_{j^*} + \gamma_n' = \alpha - \gamma_n$, so $\check{\lambda}_{j^*}$ is feasible in (6) and $\lambda_n \le \check{\lambda}_{j^*}$. If $\lambda_n < \check{\lambda}_{j^*}$, then $R(\lambda_n) > \beta_{j^*} \ge \alpha - k_n$ by definition of $\check{\lambda}_{j^*}$. If $\lambda_n = \check{\lambda}_{j^*}$, then every $\lambda < \lambda_n$ is infeasible, so $R_n^\omega(\lambda_n^-) \ge \alpha - \gamma_n$ and $R_n^\omega(\check{\lambda}_{j^*}) \ge \alpha - \gamma_n - d_n$; on E_{j^*}, therefore,

$$R(\lambda_n) = R(\check{\lambda}_{j^*}) > R_n^\omega(\check{\lambda}_{j^*}) - u_{j^*} \ge \alpha - \gamma_n - \gamma_n' - d_n = \alpha - k_n.$$

A union bound with the event of Theorem C.1 completes the proof. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the three main contributions: anytime-valid risk control, a matching lower bound proving asymptotic optimality, and extension to distribution shift. These align well with Theorems 4.1, 4.8, and Propositions 4.6, 4.11 and 4.12.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section 6, where we also outline directions for improvement.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems state their assumptions explicitly, and the proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All experiments are fully specified in the paper. The synthetic experiments include complete data generating processes, model choices, score functions, and parameter settings (α, δ, sample sizes). The real-data experiments reference established baseline implementations. Code to reproduce all experimental results is provided in the Supplementary Material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code to reproduce the experimental results is provided in the Supplementary Material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and test details are fully specified in the paper and supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The figures show distributions of miscoverage rates over many independent runs, directly illustrating variability across random calibration datasets.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments are lightweight and do not require significant compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: This is a methods paper, and no immediate societal impacts are apparent.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No models, datasets, or high-risk assets are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and code used are properly cited with references to their original sources.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects research was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a core methodological component of this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.