OBCP: ONLINE BACKWARD CONFORMAL PREDICTION UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Backward conformal prediction (BCP) controls prediction-set size through a data-dependent level. In online prediction, true labels are revealed after predictions are made. As the data distribution changes, the true-label BCP ratio may no longer satisfy the e-value expectation condition, and past data may no longer reflect current risk. We propose Online Backward Conformal Prediction (OBCP), extending BCP to dynamic prediction and risk reporting under a fixed set-size budget. We introduce a conditional inflation factor and construct a clipped product surrogate to bound miscoverage risk without relying on the e-value condition. We derive a three-term risk bound combining a recency-weighted average of past clipped products, a target-drift term, and a finite-sample concentration term. When a valid drift allowance is available, this bound yields fixed-time high-probability certificates for current conditional miscoverage. We also establish exact quantitative relations between the clipped-product statistic and empirical error counting. Simulations compare oracle-drift certificates, while a chronological real-data audit evaluates the observable product statistic.

1 INTRODUCTION

Conformal prediction provides model-agnostic uncertainty quantification through prediction sets with finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2023). In classification, however, the standard workflow fixes a target miscoverage level and lets the prediction-set size be determined by the data, which can produce overly large and uninformative sets in difficult cases (Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021). A growing literature therefore studies more efficient conformal sets through adaptive scores, model optimization, expected-size analysis, and explicit length or robustness objectives (Stutz et al., 2022; Dhillon et al., 2024; Kiyani et al., 2024; Yan et al., 2024). These methods improve set efficiency but generally retain the coverage-first perspective. Our starting point is the complementary setting in which the operational constraint is the maximum number of labels that may be reported.

Backward conformal prediction (BCP) reverses this workflow by fixing a set-size constraint and selecting a data-dependent miscoverage level (Gauthier et al., 2025a). Its construction builds on e-values, conformal e-prediction, and post-hoc validity (Vovk & Wang, 2021; Vovk, 2025; Gauthier et al., 2025b; Balinsky & Balinsky, 2024), with recent work studying adaptive scores (Gazin et al., 2024), data-dependent coverage policies (Gauthier et al., 2026), post-selection inference (Wang et al., 2024), and tighter BCP constructions (Liu et al., 2026). These results are primarily developed for exchangeable batch settings. By contrast, online conformal methods address distribution shift through sequential adaptation of coverage levels, thresholds, or step sizes (Gibbs & Candes, 2021; Gibbs & Candès, 2024; Zaffran et al., 2022; Bhatnagar et al., 2023; Angelopoulos et al., 2023; 2024a), while related approaches handle temporal dependence, covariate or label shift, and non-exchangeability through conditional modeling or reweighting (Xu & Xie, 2021; 2023; Tibshirani et al., 2019; Podkopaev & Ramdas, 2021; Barber et al., 2023; Farinhas et al., 2024). These methods do not directly address the fixed-budget BCP problem. When the current point and the reference window are no longer exchangeable, the true-label BCP ratio may fail to satisfy the e-value expectation condition, and past observations may no longer represent current risk.

Risk-control methods provide complementary tools for post-deployment risk certification and monitoring (Bates et al., 2021; Angelopoulos et al., 2024b). Related work

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

develops sequential approaches for tracking predictive risk under distribution shift (Podkopaev & Ramdas, 2022; Hultberg et al., 2026). Recent online conformal work also shows that binary miscoverage feedback can discard useful distance-to-threshold information (Wu et al., 2025). We propose Online Backward Conformal Prediction (OBCP), an online extension of BCP for fixed-budget prediction and dynamic risk reporting. Our contributions are threefold.

- **Online BCP with a clipped risk surrogate.** We characterize the inflation of the backward ratio and construct a clipped product that pointwise upper-bounds prediction errors, without assuming the e-value expectation condition.
- **Recency-weighted risk bounds.** We derive a fixed-time bound separating historical products, target drift, and finite-sample uncertainty. A valid drift allowance yields a prediction-time certificate and deterministic weighting exposes a recency-uncertainty tradeoff.
- **Connection with past-error counting.** We show that the clipped-product statistic equals weighted error counting plus a nonnegative soft component from covered predictions. This decomposition identifies when reduced target drift can offset the additional historical conservatism.

The remainder of the paper develops the online fixed-budget setting and OBCP construction in Sections 2-3, establishes the corresponding risk certificates in Section 4, and then evaluates the method in controlled simulations and real data in Sections 5 and 6, respectively. Section 7 concludes with a discussion.

2 Background and Problem Setup

2.1 Conformal prediction

Conformal prediction constructs prediction sets with finite-sample marginal coverage guarantees (Vovk et al., 2005). Consider a calibration sample $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ and a test observation $(X_{\text{test}}, Y_{\text{test}})$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y} := \{1, \ldots, K\}$. Assume that the calibration and test observations are exchangeable. Let $S : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be a fixed nonconformity score, typically obtained from a model trained on separate data, with smaller scores indicating better agreement between an input and a candidate label. For a prespecified miscoverage level $\alpha \in (0, 1)$, split conformal prediction uses the calibration scores $\{S(X_i, Y_i)\}_{i=1}^n$ to calibrate a threshold and includes candidate labels whose test scores satisfy this threshold. The resulting prediction set $\widehat{C}_n^{\alpha}(X_{\text{test}})$ satisfies

$$\mathbb{P}\Big\{Y_{\text{test}} \in \widehat{C}_n^{\alpha}(X_{\text{test}})\Big\} \geq 1 - \alpha.$$

The probability is taken jointly over the calibration sample and the test observation and the guarantee is therefore marginal rather than conditional on a particular test input. While α is chosen in advance, the prediction-set size is data-dependent and need not satisfy a prescribed size limit. Backward conformal prediction reverses this perspective: it starts from a set-size constraint and selects a data-dependent level accordingly.

2.2 Backward conformal prediction

Backward conformal prediction (BCP) (Gauthier et al., 2025a) starts from a set-size constraint and selects a data-dependent level to satisfy it. It builds on *e-variables*: nonnegative random variables E satisfying $\mathbb{E}[E] \leq 1$, whose realizations are called *e-values*. By Markov's inequality, $\mathbb{P}(E \geq 1/a) \leq a$ for every fixed $a \in (0, 1)$. Using the calibration sample and score function defined above, the ratio-based construction assigns each candidate label

$$E_y^{\text{test}} := \frac{(n+1)S(X_{\text{test}}, y)}{\sum_{i=1}^n S(X_i, Y_i) + S(X_{\text{test}}, y)}, \qquad y \in \mathcal{Y}.$$

Under exchangeability and positive denominators almost surely, $\mathbb{E}[E_{Y_{\text{test}}}^{\text{test}}] = 1$, so the true-label ratio is an e-variable. For a fixed budget $B \in \{1, \ldots, K-1\}$, define $\widehat{C}_n^a(X_{\text{test}}) = \{y \in \mathcal{Y} : E_y^{\text{test}} < 1/a\}$ and select $\widetilde{\alpha} = \inf\{a \in (0, 1) : |\widehat{C}_n^a(X_{\text{test}})| \leq B\}$, using only the calibration data and test input. Assuming $\widetilde{\alpha} \in (0, 1)$ almost surely, the resulting set contains at most B labels and satisfies the exact post-hoc guarantee $\mathbb{E}[\mathbf{1}\{Y_{\text{test}} \notin \widehat{C}_n^{\widetilde{\alpha}}(X_{\text{test}})\}/\widetilde{\alpha}] \leq 1$. Without exchangeability, the analogous online

ratio need not satisfy the e-variable expectation bound, motivating a separate analysis of risk reporting under distribution shift.

2.3 Online prediction under a fixed set-size budget

We consider an online sequence $Z_t = (X_t, Y_t) \in \mathcal{X} \times \mathcal{Y}$, $t = 1, \ldots, T$. At each time t, the learner observes X_t and reports a prediction set or declares infeasibility before observing Y_t. The label is then revealed and becomes available for subsequent rounds. We write $\mathcal{F}_{t-1} := \sigma(Z_1, \ldots, Z_{t-1})$ and $\mathcal{G}_t := \sigma(Z_1, \ldots, Z_{t-1}, X_t)$ for the historical and prediction-time information, respectively. All prediction-time decisions must be $\mathcal{G}_t$-measurable. The data distribution may change over time, and we assume neither stationarity nor exchangeability.

The score function S and budget B remain fixed throughout deployment. At time t, the backward construction uses an $\mathcal{F}_{t-1}$-measurable reference set $\mathcal{R}_t \subseteq \{1, \ldots, t-1\}$, such as the most recent L completed observations. Each reported set $\widehat{C}_t(X_t) \subseteq \mathcal{Y}$ must satisfy $1 \le |\widehat{C}_t(X_t)| \le B$. If the backward construction admits no nonempty set within this budget, the learner declares infeasibility. Let $F_t \in \{0, 1\}$ denote the feasibility indicator, which is $\mathcal{G}_t$-measurable. Thus, the budget is enforced on each feasible report rather than on the average set size.

At feasible reporting times, define the realized error $\mathrm{err}_t := \mathbf{1}\{Y_t \notin \widehat{C}_t(X_t)\}$. Our risk-reporting target is the current conditional miscoverage probability

$$r_t^\star := \mathbb{P}\Big\{ Y_t \notin \widehat{C}_t(X_t) \,\Big|\, \mathcal{G}_t \Big\} = \mathbb{E}[\mathrm{err}_t \mid \mathcal{G}_t].$$

Statements involving err_t and $r_t^\star$ are understood on $\{F_t = 1\}$. Our objective is to construct prediction-time upper bounds on this risk from past feedback, with explicit allowances for temporal drift, while retaining the fixed set-size budget.

3 Online Backward Conformal Prediction

3.1 Online backward construction

At time t, we apply the BCP ratio construction to the reference set $\mathcal{R}_t$. For each candidate label $y \in \mathcal{Y}$, define the local ratio

$$R_{t,y} := \frac{(|\mathcal{R}_t| + 1) S(X_t, y)}{\sum_{i \in \mathcal{R}_t} S(X_i, Y_i) + S(X_t, y)}. \tag{1}$$

We assume that the denominator is positive for every candidate label almost surely. These ratios are $\mathcal{G}_t$-measurable: they use only the completed reference observations and the current candidate-label scores. For a nominal level $a \in (0, 1)$, define $\widehat{C}_t^a(X_t) := \{y \in \mathcal{Y} : R_{t,y} < 1/a\}$. Let $\mathcal{A}_t := \{a \in (0, 1) : 1 \le |\widehat{C}_t^a(X_t)| \le B\}$ be the set of feasible levels. If $\mathcal{A}_t \neq \varnothing$, set $F_t = 1$ and define $\widetilde{\alpha}_t := \inf \mathcal{A}_t$, $\widehat{C}_t(X_t) := \widehat{C}_t^{\widetilde{\alpha}_t}(X_t)$. Otherwise, set $F_t = 0$ and declare infeasibility.

After the label is revealed, the true-label ratio R_{t,Y_t} becomes observable. Unlike the BCP ratio $E^{\mathrm{test}}_{Y_{\mathrm{test}}}$, which is an e-variable under exchangeability, R_{t,Y_t} is not assumed to satisfy the e-value expectation bound. Thus, the construction guarantees the set-size budget on feasible reports. It does not, however, make $\widetilde{\alpha}_t$ a valid upper bound on the current risk $r_t^\star$. The next subsection develops a product surrogate for analyzing this risk.

3.2 Clipped product surrogate

The local ratio $R_{t,y}$ need not have conditional mean at most one, so Markov's inequality does not by itself justify $\widetilde{\alpha}_t$ as an upper bound on the current conditional risk. We therefore introduce the conditional mean as an inflation factor in the product bound. Define the conditional inflation factor $c_t := \mathbb{E}[R_{t,Y_t} \mid \mathcal{G}_t]$ and the product surrogate $Q_t := \widetilde{\alpha}_t R_{t,Y_t}$. The threshold construction gives the pointwise relation $\mathrm{err}_t = \mathbf{1}\{Q_t \ge 1\} \le Q_t$. Since $\widetilde{\alpha}_t$ is $\mathcal{G}_t$-measurable, taking conditional expectations yields

$$r_t^\star \le \mathbb{E}[Q_t \mid \mathcal{G}_t] = \widetilde{\alpha}_t c_t.$$

This bound follows directly from the threshold definition and does not require stationarity or exchangeability. If $c_t \leq 1$, it reduces to the familiar bound $r_t^\star \leq \widetilde{\alpha}_t$. Under distribution shift, however, the conditional mean c_t need not be at most one, so $\widetilde{\alpha}_t$ alone may understate the current risk. We therefore retain this possible inflation through the product $Q_t = \widetilde{\alpha}_t R_{t,Y_t}$ and use Q_t as the basic surrogate for risk reporting in the dynamic setting.

Although Q_t upper-bounds the realized loss err_t, it can exceed one. Such excess is unnecessary for a binary loss and only makes subsequent averages more conservative. We therefore define the clipped surrogate $Q_t^{(\tau)} := \min\{Q_t, \tau\}$ for $\tau \geq 1$. Since $\mathrm{err}_t = \mathbf{1}\{Q_t \geq 1\}$, we have $\mathrm{err}_t \leq Q_t^{(\tau)} \leq Q_t$, and hence

$$r_t^\star \leq \mathbb{E}[Q_t^{(\tau)} \mid \mathcal{G}_t] \leq \widetilde{\alpha}_t c_t. \tag{2}$$

Unless stated otherwise, we set $\tau = 1$, the smallest value preserving the pointwise loss bound. Then $Q_t^{(1)} = \mathrm{err}_t + Q_t \mathbf{1}\{Q_t < 1\}$, equaling one on miscoverage events and retaining the original product otherwise. At prediction time, the current clipped product is unavailable because Y_t has not yet been revealed, whereas past clipped products are observed after feedback. We aggregate these historical values next and use them to bound the current conditional risk after accounting for target drift and sampling uncertainty.

3.3 Recency-weighted product statistics

The bound in equation 2 reduces current risk reporting to the conditional mean of the clipped product. We therefore summarize recent history through a recency-weighted calendar-window average. Fix a window length $m \geq 1$ and deterministic weights $w = (w_1, \ldots, w_m)$ satisfying $w_\ell \geq 0$ and $\sum_{\ell=1}^m w_\ell = 1$. At time $t > m$, if $F_i = 1$ for every $i = t-m, \ldots, t$, define

$$\widehat{\beta}_{t,w}^{(\tau),\mathrm{roll}} := \sum_{\ell=1}^{m} w_\ell Q_{t-\ell}^{(\tau)}.$$

We refer to $\widehat{\beta}_{t,w}^{(\tau),\mathrm{roll}}$ as the Product statistic. If this window-feasibility requirement fails, we do not report $\widehat{\beta}_{t,w}^{(\tau),\mathrm{roll}}$ at time t. The weight w_ℓ is assigned to the product from time $t-\ell$, so w_1 corresponds to the most recent past observation. Uniform weights $w_\ell = 1/m$ recover the ordinary rolling average. Weights with larger mass on smaller lags place more emphasis on recent observations and reduce reliance on older data.

Because the past clipped products have already been observed, $\widehat{\beta}_{t,w}^{(\tau),\mathrm{roll}}$ is available before Y_t is revealed. However, it summarizes only historical risk surrogates and is not itself an upper bound on the current risk $r_t^\star$. Relating it to the current conditional target requires two corrections: one for temporal drift in the product mean and one for sampling deviation. Section 4 introduces the corresponding drift allowance and concentration term.

3.4 Relation to empirical counting

Because past labels are revealed after prediction, realized miscoverage indicators provide a natural benchmark for historical risk reporting. We therefore compare the product statistic with empirical counting using the same window and weights. Define $\widehat{r}_{t,w}^{\mathrm{count}} := \sum_{\ell=1}^m w_\ell \, \mathrm{err}_{t-\ell}$. For $\tau = 1$,

$$\widehat{s}_{t,w}^{\mathrm{soft}} := \sum_{\ell=1}^{m} w_\ell Q_{t-\ell} \mathbf{1}\{Q_{t-\ell} < 1\} \geq 0, \qquad \widehat{\beta}_{t,w}^{(1),\mathrm{roll}} = \widehat{r}_{t,w}^{\mathrm{count}} + \widehat{s}_{t,w}^{\mathrm{soft}}. \tag{3}$$

Counting and the clipped product agree on miscoverage events, while the latter retains Q_i on covered predictions. The soft term $\widehat{s}_{t,w}^{\mathrm{soft}}$ preserves proximity-to-boundary information but makes Product more conservative on the same window. Its conditional target may vary less than hard risk. Section 4.3 characterizes when this drift reduction offsets the soft contribution.

4 Risk Certificates and Theoretical Analysis

We derive a drift-explicit bound for the current conditional risk. Throughout this section, we set $\tau = 1$ and write $\beta_t^\circ := \mathbb{E}[Q_t^{(1)} \mid \mathcal{G}_t]$. By equation 2, $r_t^\star \leq \beta_t^\circ$. Since $\widehat{\beta}_{t,w}^{(1),\mathrm{roll}}$ averages past observed

products rather than the current target, the bound includes temporal drift and sampling uncertainty. Observable certification additionally requires a valid drift allowance.

4.1 Target drift and risk certificates

To relate the current target β_t° to its weighted history, define $d_j := |\beta_j^\circ - \beta_{j-1}^\circ|$ and

$$A_q(w) := \sum_{\ell=q}^{m} w_\ell, \qquad \Gamma_{\beta,t}(w) := \sum_{q=1}^{m} A_q(w) d_{t-q+1}. \tag{4}$$

Here $A_q(w)$ is the cumulative weight assigned to targets preceding the qth most recent change, so $\Gamma_{\beta,t}(w)$ measures the weighted drift between the current target and its history. We also define the effective sample size $m_{\text{eff}}(w) := (\sum_{\ell=1}^{m} w_\ell^2)^{-1}$.

Theorem 1. *Fix a time $t > m$, $\eta \in (0,1)$, and deterministic weights w satisfying $w_\ell \geq 0$ and $\sum_{\ell=1}^{m} w_\ell = 1$. With probability at least $1-\eta$, the following implication holds: if the calendar window $t-m,\dots,t$ is feasible, then*

$$r_t^\star \leq \widehat{\beta}_{t,w}^{(1),\text{roll}} + \Gamma_{\beta,t}(w) + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}. \tag{5}$$

The proof uses conditional Hoeffding concentration (Appendix A.1). The three terms in equation 5 are historical products, target drift, and sampling uncertainty. The guarantee is fixed-time, not simultaneous over the deployment trajectory. We denote the right-hand side, truncated at one, by the oracle certificate $\widehat{\rho}_{t,w}^{\text{cert}}$. The drift term $\Gamma_{\beta,t}(w)$ depends on unknown conditional targets. Section 4.2 replaces it with a valid prediction-time allowance when available.

4.2 Drift allowances and the recency-uncertainty tradeoff

To obtain an observable certificate, suppose a prediction-time quantity $\overline{\Gamma}_{t,w}$ satisfies $\mathbb{P}\{\Gamma_{\beta,t}(w) \leq \overline{\Gamma}_{t,w}\} \geq 1-\xi$. Combining this event with Theorem 1 gives the following result.

Corollary 1. *Under the conditions of Theorem 1, suppose $\mathbb{P}\{\Gamma_{\beta,t}(w) \leq \overline{\Gamma}_{t,w}\} \geq 1-\xi$. Then, with probability at least $1-\eta-\xi$, the following implication holds: if the calendar window $t-m,\dots,t$ is feasible, then*

$$r_t^\star \leq \widehat{\beta}_{t,w}^{(1),\text{roll}} + \overline{\Gamma}_{t,w} + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}.$$

Because $\Gamma_{\beta,t}(w)$ depends on unknown conditional targets, OBCP separates historical estimation from control of temporal drift. We treat $\overline{\Gamma}_{t,w}$ as an assumption-dependent input, not a drift estimate supplied by OBCP. Appendix A.3 gives sufficient allowances under bounds on local target variation or distributional drift.

The weighting scheme determines a recency-uncertainty tradeoff. Define the weighted mean lag $\mu_{\text{lag}}(w) := \sum_{\ell=1}^{m} \ell w_\ell$. Under a uniform bound on adjacent target changes, the following corollary makes this tradeoff explicit.

Corollary 2. *Under the conditions of Theorem 1, suppose $d_j \leq \Delta_\beta$ for all $j = t-m+1,\dots,t$. Then $\Gamma_{\beta,t}(w) \leq \mu_{\text{lag}}(w)\Delta_\beta$. Consequently, with probability at least $1-\eta$, if the calendar window $t-m,\dots,t$ is feasible, then*

$$r_t^\star \leq \widehat{\beta}_{t,w}^{(1),\text{roll}} + \mu_{\text{lag}}(w)\Delta_\beta + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}.$$

The two weight summaries expose the tradeoff: a smaller $\mu_{\text{lag}}(w)$ reduces the drift allowance, whereas a smaller $m_{\text{eff}}(w)$ increases the concentration radius. Uniform weights maximize $m_{\text{eff}}(w)$ on a fixed window, while recency-weighted choices seek shorter memory, often at the cost of reduced effective sample size.

4.3 BOUNDARY INFORMATION AND MATCHED CERTIFICATES

By equation 3, the Product statistic exceeds Count by the nonnegative soft contribution $\widehat{s}^{\mathrm{soft}}_{t,w}$. Thus, when the historical term is considered alone, Product is necessarily more conservative. A complete risk certificate, however, also includes a correction for temporal target drift. The OBCP certificate tracks the clipped-product target β°_t, whereas a counting-based certificate tracks the hard-risk target $r^{\star}_t$. These targets need not vary at the same rate. In particular, clipping retains graded information near the error boundary and can therefore be less sensitive to small boundary crossings. The corresponding stability result is given in Appendix A.4.

For a matched comparison, define $\Gamma_{r,t}(w)$ as in equation 4, with β°_i replaced by $r^{\star}_i$, and let $\mathrm{Rad}_{\eta}(w) := \sqrt{\log(1/\eta)/(2m_{\mathrm{eff}}(w))}$. Write $\widetilde{\rho}^{Q}_{t,w} := \widehat{\beta}^{(1),\mathrm{roll}}_{t,w} + \Gamma_{\beta,t}(w) + \mathrm{Rad}_{\eta}(w)$ and $\widetilde{\rho}^{\mathrm{count}}_{t,w} := \widehat{r}^{\mathrm{count}}_{t,w} + \Gamma_{r,t}(w) + \mathrm{Rad}_{\eta}(w)$ for the corresponding untruncated expressions. Their reported certificates are $\widehat{\rho}^{\mathrm{cert}}_{t,w} := \min\{1, \widetilde{\rho}^{Q}_{t,w}\}$ and $\widehat{\rho}^{\mathrm{count}}_{t,w} := \min\{1, \widetilde{\rho}^{\mathrm{count}}_{t,w}\}$.

Corollary 3 (Matched certificate comparison)**.** *On any feasible calendar window, the matched oracle certificate expressions satisfy*

$$\widetilde{\rho}^{Q}_{t,w} - \widetilde{\rho}^{\mathrm{count}}_{t,w} = \widehat{s}^{\mathrm{soft}}_{t,w} + \Gamma_{\beta,t}(w) - \Gamma_{r,t}(w).$$

Consequently, if $\Gamma_{r,t}(w) - \Gamma_{\beta,t}(w) > \widehat{s}^{\mathrm{soft}}_{t,w}$ *and* $\widehat{\rho}^{\mathrm{cert}}_{t,w} < 1$*, then* $\widehat{\rho}^{\mathrm{cert}}_{t,w} < \widehat{\rho}^{\mathrm{count}}_{t,w}$.

The proof is given in Appendix A.5. Corollary 3 gives a conditional comparison: OBCP is strictly tighter when its target-drift reduction exceeds the historical soft surplus and its certificate is below one. Section 5.2 illustrates this tradeoff in constructed near-boundary regimes.

5 SIMULATIONS

We use controlled categorical streams with known conditional laws to study the recency-uncertainty tradeoff, historical risk summaries, and matched OBCP-versus-Count certificates. All complete certificates use oracle target drift to isolate weighting and target-stability effects from drift estimation. These simulations assess the bound under known drift, not an end-to-end drift-estimation procedure. Specifications and additional results are provided in Appendix B.

5.1 RECENCY-UNCERTAINTY TRADEOFF

We vary only the deterministic weighting kernel while holding the underlying prediction process fixed. For each seed, all kernels share the same trajectory, prediction sets, clipped products, oracle targets, and true risks. We compare U200, T200, and exponential kernels E100, E50, and E25, whose half-lives are 100, 50, and 25, under stationarity, one abrupt shift, and recurring shifts with $H \in \{100, 500\}$. Each environment uses 200 paired trajectories with $B = 3$, $L = m = 200$, $\tau = 1$, and $\eta = 0.05$. Complete specifications are given in Appendices B.1.1-B.1.4.

For each kernel, Table 1 reports the signed certificate gap $\widehat{\rho}^{\mathrm{cert}}_{t,w} - r^{\star}_t$, averaged over the full common reportable trajectory and then equally across trajectories. This metric summarizes average certificate tightness across fixed kernels rather than trajectory-wise validity.

U200 has the smallest mean gap under stationarity and the abrupt shift. Under recurring shifts, E25 reduces the gap by 28.2% relative to U200 at $H = 100$, whereas T200 is smallest at $H = 500$, improving on U200 by 3.4%; E25 then has a larger mean gap than U200. Thus, consistent with Corollary 2, the useful memory scale depends on the temporal frequency of distribution shift.

5.2 RISK REPORTING AND CERTIFICATE COMPARISON

This subsection addresses two questions. We first compare historical reporting signals under known distribution shifts, and then compare the complete OBCP certificate with a matched counting certificate to isolate the effect of target stability. Both experiments use five-class categorical streams with $B = 3$, $L = m = 200$, $\tau = 1$, and $\eta = 0.05$, and each scenario averages over 200 paired trajectories. U200 denotes uniform weighting over the previous 200 observations. Full models and protocols are given in Appendix B.2.

Table 1: Kernel geometry and mean signed oracle-certificate gap. The last four columns average the gap over the full common reportable trajectory and 200 paired trajectories; bold marks the smallest value in each environment.

Kernel	m_{eff}	μ_{lag}	Stationary	Abrupt	Switch $H = 100$	Switch $H = 500$
U200	200.0	100.5	**0.242**	**0.288**	0.677	0.353
E100	173.1	78.1	0.249	0.294	0.610	0.343
T200	150.4	67.3	0.255	0.301	0.572	**0.341**
E50	127.3	59.3	0.264	0.310	0.549	0.344
E25	71.6	35.8	0.300	0.346	**0.486**	0.362

Historical reporting under temporal shift. We first compare four observable historical signals-Static BCP, Roll-α, Count, and Product-under three known risk paths: shock-recovery, gradual increase, and recurring shifts with $H = 500$. The fixed score model yields feasible sets $\{1, 2, 3\}$ and conditional risks between 0.08 and 0.16. Full specifications are given in Appendix B.2.2.

Static BCP is fixed after initialization, whereas the other three signals average past levels, errors, and clipped products, respectively. None of these four signals includes a drift or concentration correction. Figure 1 also overlays Oracle OBCP, which adds the exact product-target drift and concentration radius to Product.

For any historical risk summary A_t among Static BCP, Roll-α, Count, and Product, we quantify downward and upward discrepancies by the time-averaged shortfall $(r_t^\star - A_t)_+$ and overestimation $(A_t - r_t^\star)_+$, averaged across trajectories. During the upward-shift phase, Product reduces Count's mean shortfall from 0.0420 to 0.0180. After the risk decreases, this additional margin makes Product more conservative, increasing mean overestimation from 0.0389 to 0.0685. A similar pattern appears under recurring shifts: Product lowers mean shortfall from 0.0145 to 0.0042, while mean overestimation increases from 0.0135 to 0.0341. These results illustrate a margin tradeoff: the soft contribution reduces shortfall during rising risk but increases overestimation after risk decreases. Detailed uncertainty summaries are reported in Appendix B.2.2.

Matched complete certificates. We next compare Oracle OBCP with a matched Oracle Count certificate. Both use the same predictions, U200 history, confidence level, concentration radius, and final truncation, while each uses the drift of its own conditional target.

To isolate the effect of target stability, we transfer probability mass across two labels lying close to the prediction boundary, so that the hard risk alternates between 0.08 and 0.16 while the clipped-product target changes much less. The two regimes alternate in blocks of length $H \in \{25, 50, 200\}$; full construction details are given in Appendix B.2.3.

In this construction, Oracle OBCP is tighter for fast switching: mean certificates are 0.269 versus 0.528 for Count at $H = 25$, and 0.268 versus 0.368 at $H = 50$. At $H = 200$, Count is tighter (0.247 versus 0.268 for OBCP), consistent with the surplus–drift tradeoff in Corollary 3 as switching slows. Monte Carlo uncertainties and component decompositions are reported in Appendix B.2.3.

6 Real-Data Analysis

We evaluate the observable historical summaries on a chronological CLEAR10 image stream (Lin et al., 2021). Because the current conditional risk and target drift are unknown, we compare these summaries with subsequently realized errors through a retrospective local audit. This examines the historical surrogate, not the full high-probability certificate.

CLEAR10 setup. CLEAR10 has $K = 11$ classes. We fit a multinomial logistic-regression classifier on fixed ImageNet-pretrained ResNet-18 features (He et al., 2016) using only training bucket 1, and keep the representation and score model fixed throughout deployment. The evaluation stream consists of $4{,}950$ chronologically ordered observations from test buckets 2-10. We use $B = 3$, $L = m = 200$, $\tau = 1$, and uniform weighting over the previous 200 observations (U200). All $4{,}950$ predictions are feasible and have size three, with 116 realized miscoverage events (2.343%). We compare four

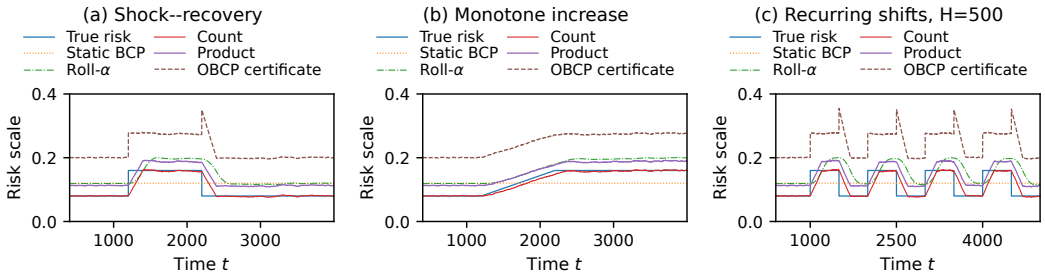


Figure 1: Comparison of Static BCP, Roll-α, Count, Product, and Oracle OBCP with the true conditional risk under (a) shock-recovery, (b) gradual increase, and (c) recurring shifts with $H = 500$. Curves are means over 200 trajectories. Static BCP, Roll-α, Count, and Product are uncorrected historical risk summaries, whereas Oracle OBCP augments Product with the oracle target-drift correction and concentration radius.

historical risk summaries: Static BCP, Roll-α U200, Count U200, and Product U200. Data partitions, model fitting, online updates, and summary definitions are detailed in Appendices C.1-C.5.

Retrospective risk audit. Because $r_t^\star$ is unavailable, we use subsequently revealed labels to construct a local retrospective benchmark. At each evaluation time t, let

$$\widehat{r}_{t,100}^{\,\mathrm{future}} := \frac{1}{100}\sum_{j=0}^{99} \mathrm{err}_{t+j}, \tag{6}$$

the empirical miscoverage over the 100 predictions starting at t. For a historical summary A_t, we measure shortfall by $[\widehat{r}_{t,100}^{\,\mathrm{future}} - A_t]_+$ and define the underestimation rate as the fraction of times $A_t < \widehat{r}_{t,100}^{\,\mathrm{future}}$. The common evaluation support contains 4,651 starting times with both a complete previous-200 history and a complete future-100 audit window. Table 2 reports the resulting local shortfall diagnostics.

Product U200 remains at or above the future-100 empirical error at all 4,651 evaluated starting times, whereas Count U200 underestimates it at 41.41% of the starting times and reaches a worst shortfall of 0.060. The difference is consistent with equation 3: Product retains the nonnegative soft contribution from covered predictions, providing additional margin beyond binary error counting. Static BCP and Roll-α have substantially smaller shortfalls than Count but still fall below the retrospective benchmark at 2.75% and 3.44% of evaluation times, respectively.

Budget sensitivity is reported in Appendix C.7. Across $B = 2, \ldots, 6$, all predictions remain feasible, while empirical miscoverage decreases from 0.04727 to 0.00465 as the allowed set size increases.

7 Conclusion and Discussion

We introduced OBCP for online fixed-budget prediction under distribution shift. OBCP uses a clipped-product risk surrogate that retains soft boundary information, and derives a recency-weighted bound separating historical products, target drift, and finite-sample uncertainty. Given a valid drift allowance, the bound yields fixed-time risk certificates for current conditional miscoverage. The analysis further exposes a recency-uncertainty tradeoff and characterizes when reduced target drift can offset the additional conservatism of the product surrogate relative to empirical error counting.

Operational certification uses an application-dependent drift allowance, with sufficient constructions discussed in Appendix A.3. This separation allows the core OBCP construction to accommodate different mechanisms of distribution shift. Extending the fixed-time analysis and developing sharper certificates are natural directions for future work.

Table 2: Retrospective future-100 shortfall audit on CLEAR10 with $B = 3$. All four summaries are evaluated at the same 4,651 starting times. Underestimation rate measures the frequency of shortfalls, while average and worst shortfall measure their magnitude. Worst shortfall is the maximum observed positive gap between the future-100 empirical error and the historical summary.

Metric	Static BCP	Roll-α	Count U200	Product U200
Underestimation rate	2.75%	3.44%	41.41%	0.00%
Average shortfall	0.000245	0.000218	0.007405	0.000000
Worst shortfall	0.01405	0.01585	0.06000	0.00000

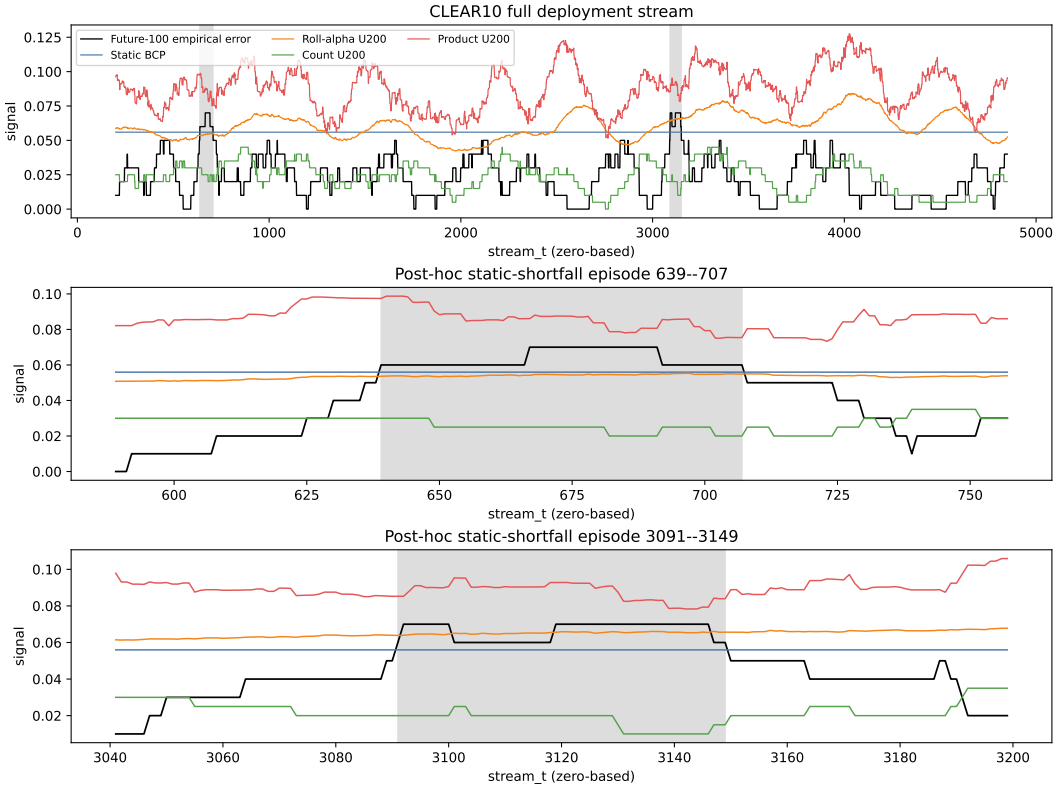


Figure 2: Static BCP, Roll-α, Count U200, Product U200, and the retrospective future-100 empirical error on CLEAR10. The top panel shows all 4,651 evaluation times. The lower panels enlarge the two retrospectively identified intervals where Static BCP falls below the future-block error. Historical summaries use only information available before the current label is revealed, whereas the audit curve uses subsequently observed labels. The gray shading marks the intervals over which Static BCP is below the future-100 empirical error.

Reproducibility Statement

The assumptions underlying our theoretical results are stated in the main text, with complete proofs provided in Appendix A. Detailed simulation specifications and evaluation protocols are given in Appendix B, while the CLEAR10 data processing, model fitting, and evaluation procedures are documented in Appendix C.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript, including improvements to clarity and readability, and for code review. All AI-assisted edits were reviewed and verified by the authors, who take responsibility for the final content and implementation of this work.

REFERENCES

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074, 2023.

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

Anastasios Nikolas Angelopoulos, Stephen Bates, Jitendra Malik, and Michael Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1616–1630. PMLR, 21–27 Jul 2024a.

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

Alexander A. Balinsky and Alexander David Balinsky. Enhancing conformal prediction using e-test statistics. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pp. 65–72. PMLR, 09–11 Sep 2024.

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

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

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 2337–2363. PMLR, 2023.

Guneet S Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1549–1557. PMLR, 2024.

António Farinhas, Chrysoula Zerva, Dennis Thomas Ulmer, and Andre Martins. Non-exchangeable conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. Backward conformal prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Etienne Gauthier, Francis Bach, and Michael I Jordan. E-values expand the scope of conformal prediction. *arXiv preprint arXiv:2503.13050*, 2025b.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. Adaptive coverage policies in conformal prediction. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3196–3204. PMLR, 02–05 May 2026.

Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics*, pp. 1504–1512. PMLR, 2024.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Bror Hultberg, Dave Zachariah, and Antônio H Ribeiro. Anytime-valid conformal risk control. *arXiv preprint arXiv:2602.04364*, 2026.

Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563, 2024.

Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The clear benchmark: Continual learning on real-world imagery. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

Junxian Liu, Hao Zeng, and Hongxin Wei. Improving backward conformal prediction via non-conformity score transformation. In *Forty-third International Conference on Machine Learning*, 2026.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pp. 844–853. PMLR, 2021.

Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. In *International Conference on Learning Representations*, 2022.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk. Conformal e-prediction. *Pattern Recognition*, 166:111674, 2025.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Xiaoning Wang, Yuyang Huo, Liuhua Peng, and Changliang Zou. Conformalized multiple testing after data-dependent selection. *Advances in Neural Information Processing Systems*, 37:58574–58609, 2024.

Junxi Wu, Dongjian Hu, Yajie Bao, Shu-Tao Xia, and Changliang Zou. Error-quantified conformal inference for time series. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pp. 38707–38727. PMLR, 2023.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved efficiency. In *The Twelfth International Conference on Learning Representations*, 2024.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International conference on machine learning*, pp. 25834–25866. PMLR, 2022.

A PROOFS

A.1 PROOF OF THEOREM 1

Fix a time $t > m$, a confidence level $\eta \in (0,1)$, and deterministic weights $w = (w_1, \dots, w_m)$ satisfying $w_\ell \geq 0$ and $\sum_{\ell=1}^m w_\ell = 1$. Let $\mathcal{E}_{t,m} := \cap_{i=t-m}^t \{F_i = 1\}$ denote the event that the calendar window is feasible. We prove that there exists an event $\mathcal{H}_{t,w,\eta}$ with probability at least $1-\eta$ such that, on $\mathcal{H}_{t,w,\eta} \cap \mathcal{E}_{t,m}$, the bound in equation 5 holds. Equivalently, $\mathcal{H}_{t,w,\eta} \subseteq \{\mathcal{E}_{t,m} \Rightarrow equation\ 5\}$.

For this fixed-time proof, set $Q_i^{(1)} = 0$ on $\{F_i = 0\}$. This zero-extension is only an analytical convention. On $\mathcal{E}_{t,m}$, the zero-extended values coincide with the clipped products used in $\widehat{\beta}_{t,w}^{(1),\mathrm{roll}}$. Since F_i is $\mathcal{G}_i$-measurable, the conditional target $\beta_i^\circ := \mathbb{E}[Q_i^{(1)} \mid \mathcal{G}_i]$ is also zero on $\{F_i = 0\}$ under this convention.

For any lag $\ell \in \{1, \dots, m\}$, the definition $d_j = |\beta_j^\circ - \beta_{j-1}^\circ|$ gives $\beta_t^\circ \leq \beta_{t-\ell}^\circ + \sum_{q=1}^\ell d_{t-q+1}$. Multiplying by w_ℓ and summing over ℓ yields

$$\beta_t^\circ \leq \sum_{\ell=1}^m w_\ell \beta_{t-\ell}^\circ + \sum_{\ell=1}^m w_\ell \sum_{q=1}^\ell d_{t-q+1}.$$

For a fixed q, the increment d_{t-q+1} appears for lags $\ell \geq q$. Hence

$$\sum_{\ell=1}^m w_\ell \sum_{q=1}^\ell d_{t-q+1} = \sum_{q=1}^m \left(\sum_{\ell=q}^m w_\ell \right) d_{t-q+1} = \Gamma_{\beta,t}(w).$$

Therefore,

$$\beta_t^\circ \leq \sum_{\ell=1}^m w_\ell \beta_{t-\ell}^\circ + \Gamma_{\beta,t}(w). \tag{7}$$

We next control the sampling deviation between the weighted past targets $\sum_{\ell=1}^m w_\ell \beta_{t-\ell}^\circ$ and the observable weighted products $\sum_{\ell=1}^m w_\ell Q_{t-\ell}^{(1)}$. Define $Z_{t,w} := \sum_{\ell=1}^m w_\ell (\beta_{t-\ell}^\circ - Q_{t-\ell}^{(1)})$. For each time i, let $D_i := \beta_i^\circ - Q_i^{(1)}$. By the definition $\beta_i^\circ = \mathbb{E}[Q_i^{(1)} \mid \mathcal{G}_i]$, we have $\mathbb{E}[D_i \mid \mathcal{G}_i] = 0$. Conditional on $\mathcal{G}_i$, the value β_i° is fixed and $Q_i^{(1)} \in [0,1]$, so $D_i \in [\beta_i^\circ - 1, \beta_i^\circ]$. Thus the conditional range width of D_i is at most one.

For any $\lambda > 0$, conditional Hoeffding's lemma gives

$$\mathbb{E}\left[\exp\left\{ \lambda w_\ell (\beta_{t-\ell}^\circ - Q_{t-\ell}^{(1)}) \right\} \,\middle|\, \mathcal{G}_{t-\ell} \right] \leq \exp\left\{ \frac{\lambda^2 w_\ell^2}{8} \right\}.$$

Iterating this conditional moment bound over $t-m, \dots, t-1$ gives $\mathbb{E}[\exp\{\lambda Z_{t,w}\}] \leq \exp\{(\lambda^2/8) \sum_{\ell=1}^m w_\ell^2\}$. By Chernoff's method,

$$\mathbb{P}(Z_{t,w} > s) \leq \exp\left\{ -\frac{2s^2}{\sum_{\ell=1}^m w_\ell^2} \right\}.$$

Taking $s = \sqrt{\log(1/\eta)/(2m_{\text{eff}}(w))}$ gives an event $\mathcal{H}_{t,w,\eta}$ with probability at least $1-\eta$ on which

$$\sum_{\ell=1}^{m} w_\ell \beta^\circ_{t-\ell} \le \sum_{\ell=1}^{m} w_\ell Q^{(1)}_{t-\ell} + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}. \tag{8}$$

On $\mathcal{H}_{t,w,\eta}$, combining equation 7 and equation 8 gives

$$\beta^\circ_t \le \sum_{\ell=1}^{m} w_\ell Q^{(1)}_{t-\ell} + \Gamma_{\beta,t}(w) + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}.$$

Now assume $\mathcal{E}_{t,m}$ holds. Then $\sum_{\ell=1}^{m} w_\ell Q^{(1)}_{t-\ell} = \widehat{\beta}^{(1),\text{roll}}_{t,w}$. Feasibility at time t also allows us to use equation 2, giving $r^\star_t \le \beta^\circ_t$. Therefore, on $\mathcal{H}_{t,w,\eta} \cap \mathcal{E}_{t,m}$,

$$r^\star_t \le \widehat{\beta}^{(1),\text{roll}}_{t,w} + \Gamma_{\beta,t}(w) + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}.$$

Since $\mathbb{P}(\mathcal{H}_{t,w,\eta}) \ge 1-\eta$, with probability at least $1-\eta$ the implication $\mathcal{E}_{t,m} \Rightarrow equation\ 5$ holds. This proves the theorem.

A.2 Proofs for Section 4.2

Proof of Corollary 1. Let $\mathcal{H}_{t,w,\eta}$ denote the high-probability event from Theorem 1. On this event, whenever the calendar window $t-m,\dots,t$ is feasible,

$$r^\star_t \le \widehat{\beta}^{(1),\text{roll}}_{t,w} + \Gamma_{\beta,t}(w) + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}.$$

By assumption, the drift allowance satisfies $\mathbb{P}\{\Gamma_{\beta,t}(w) \le \overline{\Gamma}_{t,w}\} \ge 1-\xi$. Let $\mathcal{D}_{t,w} := \{\Gamma_{\beta,t}(w) \le \overline{\Gamma}_{t,w}\}$. Since $\mathbb{P}(\mathcal{H}_{t,w,\eta}) \ge 1-\eta$ and $\mathbb{P}(\mathcal{D}_{t,w}) \ge 1-\xi$, the union bound gives $\mathbb{P}(\mathcal{H}_{t,w,\eta} \cap \mathcal{D}_{t,w}) \ge 1-\eta-\xi$. On $\mathcal{H}_{t,w,\eta} \cap \mathcal{D}_{t,w}$, substituting $\Gamma_{\beta,t}(w) \le \overline{\Gamma}_{t,w}$ into the oracle bound yields

$$r^\star_t \le \widehat{\beta}^{(1),\text{roll}}_{t,w} + \overline{\Gamma}_{t,w} + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}.$$

This proves the corollary.

Proof of Corollary 2. By definition, we have

$$\Gamma_{\beta,t}(w) = \sum_{q=1}^{m} A_q(w) d_{t-q+1}, \qquad A_q(w) = \sum_{\ell=q}^{m} w_\ell.$$

If $d_j \le \Delta_\beta$ for all $j = t-m+1,\dots,t$, then

$$\Gamma_{\beta,t}(w) \le \Delta_\beta \sum_{q=1}^{m} A_q(w) = \Delta_\beta \sum_{q=1}^{m}\sum_{\ell=q}^{m} w_\ell.$$

Interchanging the order of summation gives

$$\sum_{q=1}^{m}\sum_{\ell=q}^{m} w_\ell = \sum_{\ell=1}^{m}\sum_{q=1}^{\ell} w_\ell = \sum_{\ell=1}^{m} \ell w_\ell = \mu_{\text{lag}}(w).$$

Therefore $\Gamma_{\beta,t}(w) \le \mu_{\text{lag}}(w)\Delta_\beta$. Substituting this upper bound into Theorem 1 gives

$$r^\star_t \le \widehat{\beta}^{(1),\text{roll}}_{t,w} + \mu_{\text{lag}}(w)\Delta_\beta + \sqrt{\frac{\log(1/\eta)}{2m_{\text{eff}}(w)}}$$

on feasible calendar windows with probability at least $1-\eta$. This proves the corollary.

A.3 Examples of drift allowances

Corollary 1 leaves the construction of $\overline{\Gamma}_{t,w}$ unspecified so that the main theory does not depend on a particular model of distribution shift. Here we give two examples of sufficient drift allowances under additional assumptions on temporal variation.

Local variation budgets. Suppose a nonnegative, prediction-time measurable quantity $V_{t,m}$ satisfies

$$\mathbb{P}\left\{\sum_{q=1}^{m} d_{t-q+1} \le V_{t,m}\right\} \ge 1-\xi.$$

Since $0 \le A_q(w) \le 1$,

$$\Gamma_{\beta,t}(w) = \sum_{q=1}^{m} A_q(w) d_{t-q+1} \le \sum_{q=1}^{m} d_{t-q+1},$$

and therefore one may take $\overline{\Gamma}_{t,w} := V_{t,m}$. This construction is useful when one can bound the cumulative target variation over the reporting window without bounding each individual increment separately.

Bounds based on distributional drift. Appendix A.4 gives, for adjacent feasible times,

$$d_j \le \|q_j - q_{j-1}\|_\infty + \mathrm{TV}(\pi_j, \pi_{j-1}),$$

where $\pi_j(y) := \mathbb{P}(Y_j = y \mid \mathcal{G}_j)$ and $q_j(y) := \widetilde{\alpha}_j R_{j,y}$. Suppose nonnegative, prediction-time measurable quantities δ_j satisfy

$$\mathbb{P}\{\mathrm{TV}(\pi_j, \pi_{j-1}) \le \delta_j \text{ for all } j = t-m+1, \dots, t\} \ge 1-\xi.$$

Then one may take

$$\overline{\Gamma}_{t,w} := \sum_{q=1}^{m} A_q(w) \left[\|q_{t-q+1} - q_{t-q}\|_\infty + \delta_{t-q+1}\right].$$

The product-vector differences are observable from the prediction-time quantities, while δ_j encodes any additional information or assumptions about changes in the conditional label law.

Thus, the drift allowance is application dependent rather than fixed by OBCP itself. Any construction satisfying $\mathbb{P}\{\Gamma_{\beta,t}(w) \le \overline{\Gamma}_{t,w}\} \ge 1-\xi$ can be inserted into Corollary 1.

A.4 Target stability for Section 4.3

At a feasible time i, define $\pi_i(y) := \mathbb{P}(Y_i = y \mid \mathcal{G}_i)$, $q_i(y) := \widetilde{\alpha}_i R_{i,y}$, $u_i(y) := \min\{q_i(y), 1\}$, and $v_i(y) := \mathbf{1}\{q_i(y) \ge 1\}$. Then $\beta_i^\circ = \langle \pi_i, u_i \rangle$ and $r_i^\star = \langle \pi_i, v_i \rangle$. Since $q \mapsto \min\{q, 1\}$ is 1-Lipschitz, any adjacent feasible pair satisfies

$$\begin{aligned} \|u_j - u_{j-1}\|_\infty &\le \|q_j - q_{j-1}\|_\infty, \\ |\beta_j^\circ - \beta_{j-1}^\circ| &\le \|q_j - q_{j-1}\|_\infty + \mathrm{TV}(\pi_j, \pi_{j-1}), \end{aligned}$$

where $\mathrm{TV}(p, p') := \frac{1}{2}\sum_y |p(y) - p'(y)|$. For the second inequality, add and subtract $\langle \pi_j, u_{j-1} \rangle$ and use

$$|\beta_j^\circ - \beta_{j-1}^\circ| \le \|u_j - u_{j-1}\|_\infty + |\langle \pi_j - \pi_{j-1}, u_{j-1} \rangle|.$$

The last term is at most $\mathrm{TV}(\pi_j, \pi_{j-1})$ because $u_{j-1}(y) \in [0,1]$. In contrast to clipping, the scalar hard threshold is discontinuous at one. These properties explain the different target sensitivities but do not by themselves order $\Gamma_{\beta,t}(w)$ and $\Gamma_{r,t}(w)$.

A.5 Proof of Corollary 3

Proof. The two matched untruncated expressions share the same concentration radius, so by equation 3,

$$\widetilde{\rho}_{t,w}^{Q} - \widetilde{\rho}_{t,w}^{\mathrm{count}} = \widehat{s}_{t,w}^{\mathrm{soft}} + \Gamma_{\beta,t}(w) - \Gamma_{r,t}(w).$$

Hence the stated drift-saving condition implies $\widetilde{\rho}_{t,w}^{Q} < \widetilde{\rho}_{t,w}^{\mathrm{count}}$. If $\widehat{\rho}_{t,w}^{\mathrm{cert}} < 1$, then $\widetilde{\rho}_{t,w}^{Q} = \widehat{\rho}_{t,w}^{\mathrm{cert}} < 1$. Capping the larger counting expression at one therefore preserves the strict inequality. □

B Details and Additional Results for Section 5

This section provides the complete experimental specifications and supplementary results for Section 5. Each appendix subsection corresponds to one subsection in the main simulation section. We begin with the deterministic-kernel experiment from Section 5.1.

B.1 Details for Section 5.1: Recency-uncertainty tradeoff

This subsection gives the complete specification of the categorical stream, deterministic weighting kernels, temporal environments, and paired Monte Carlo protocol used in Section 5.1. All five kernels are fixed before the trajectories are generated and are evaluated on the same realized paths without data-adaptive selection.

B.1.1 Categorical stream and online construction

We consider a categorical stream without covariates, or equivalently with a degenerate covariate, on the label space $\mathcal{Y} = \{1, \dots, 5\}$. At each time t, the label is generated independently according to a time-varying categorical law,

$$Y_t \sim \mathrm{Categorical}(p_t).$$

The low-risk and shifted label laws for the four main-text environments are

$$p_0 = (0.80, 0.10, 0.07, 0.02, 0.01), \qquad p_1 = (0.25, 0.20, 0.15, 0.25, 0.15).$$

The candidate-label score is fixed throughout the stream:

$$S(y) := -\log p_{0,y}.$$

The set-size budget is $B = 3$, and the reference window contains the previous $L = 200$ completed observations. At paper time $t > 200$, the reference set is

$$\mathcal{R}_t = \{t - 200, \dots, t - 1\},$$

and the candidate-label ratio from equation 1 becomes

$$R_{t,y} = \frac{201 S(y)}{\sum_{i=t-200}^{t-1} S(Y_i) + S(y)}.$$

Because the score values are strictly ordered and the map $s \mapsto 201s/(A + s)$ is strictly increasing for every positive reference-score sum A, the candidate-ratio ordering is fixed across time. Let $R_{t,(4)}$ denote the fourth-smallest candidate ratio. When $R_{t,(4)} > 1$, the backward level is $\widetilde{\alpha}_t = 1/R_{t,(4)}$ and the strict-threshold prediction set is $\{1, 2, 3\}$. When $R_{t,(4)} \le 1$, no level in $(0, 1)$ satisfies the set-size budget and the time is declared infeasible. Thus, at feasible times the conditional miscoverage risk is

$$r_t^\star = p_t(4) + p_t(5),$$

which equals 0.03 under p_0 and 0.40 under p_1.

After Y_t is revealed at a feasible time, the clipped product is

$$Q_t^{(1)} = \min\{\widetilde{\alpha}_t R_{t,Y_t}, 1\}.$$

Since p_t is known in the simulation, the oracle conditional clipped-product target is available exactly:

$$\beta_t^\circ = \mathbb{E}[Q_t^{(1)} \mid \mathcal{G}_t] = \sum_{y=1}^{5} p_t(y) \min\{\widetilde{\alpha}_t R_{t,y}, 1\}.$$

The reporting support is $m = 200$, the clipping level is $\tau = 1$, and the confidence parameter is $\eta = 0.05$. At a reportable time, the oracle certificate for a deterministic kernel $w = (w_1, \dots, w_{200})$ is

$$\widehat{\rho}_{t,w}^{\mathrm{cert}} = \min\left\{1, \sum_{\ell=1}^{200} w_\ell Q_{t-\ell}^{(1)} + \Gamma_{\beta,t}(w) + \sqrt{\frac{\log(1/\eta)}{2 m_{\mathrm{eff}}(w)}}\right\}.$$

The corresponding oracle drift term is

$$\Gamma_{\beta,t}(w) = \sum_{q=1}^{200} A_q(w) \left|\beta^{\circ}_{t-q+1} - \beta^{\circ}_{t-q}\right|, \qquad A_q(w) := \sum_{\ell=q}^{200} w_\ell.$$

The certificate at time t uses the completed products from times $t-200, \ldots, t-1$ and never includes the current product $Q_t^{(1)}$. Reporting requires the consecutive calendar window $t-200, \ldots, t$ to be feasible and all required quantities to be available. An unavailable time is not replaced by an older feasible observation.

Under the one-based paper-time convention, the first backward ratio is available at $t = 201$, after the first reference window has been observed. The first strict certificate can be available at $t = 401$, after 200 completed products have also become available. In the four environments used in Section 5.1, all times $t = 401, \ldots, T$ are feasible and reportable in every realized trajectory.

B.1.2 Deterministic weighting kernels

Weights are assigned in reverse chronological order: w_1 is assigned to the most recent product $Q_{t-1}^{(1)}$, whereas w_{200} is assigned to the oldest product $Q_{t-200}^{(1)}$. Every kernel has nonnegative weights summing to one.

The uniform kernel U200 is

$$w_\ell^{\mathrm{U}} = \frac{1}{200}, \qquad \ell = 1, \ldots, 200.$$

The triangular kernel T200 assigns linearly decreasing weight with lag:

$$w_\ell^{\mathrm{T}} = \frac{2(201-\ell)}{200 \cdot 201}, \qquad \ell = 1, \ldots, 200.$$

For an exponential kernel with half-life h, define

$$\kappa_h := 2^{-1/h}.$$

The normalized exponential weights are

$$w_\ell^{\mathrm{E}(h)} = \frac{(1-\kappa_h)\kappa_h^{\ell-1}}{1-\kappa_h^{200}}, \qquad \ell = 1, \ldots, 200.$$

We use $h \in \{100, 50, 25\}$ and denote the resulting kernels by E100, E50, and E25.

For every kernel, the effective sample size and weighted mean lag are

$$m_{\mathrm{eff}}(w) = \frac{1}{\sum_{\ell=1}^{200} w_\ell^2}, \qquad \mu_{\mathrm{lag}}(w) = \sum_{\ell=1}^{200} \ell w_\ell.$$

Table 3 gives their numerical values and the corresponding concentration radius.

Moving from U200 toward E25 in the order shown in Table 3 decreases $\mu_{\mathrm{lag}}(w)$, but also decreases $m_{\mathrm{eff}}(w)$ and increases the concentration radius.

B.1.3 Distribution-shift environments

We evaluate four pre-specified environments. The stationary and moderate-abrupt streams have horizon $T = 3000$, while the recurring-shift streams have horizon $T = 5000$.

In the stationary environment,

$$p_t = p_0, \qquad t = 1, \ldots, 3000.$$

In the moderate-abrupt environment,

$$p_t = \begin{cases} p_0, & 1 \le t \le 1500, \\ p_1, & 1501 \le t \le 3000. \end{cases}$$

Table 3: Geometry of the deterministic weighting kernels used in Section 5.1. The radius is $\sqrt{\log(1/\eta)/(2m_{\mathrm{eff}}(w))}$ with $\eta = 0.05$ and $\tau = 1$.

Kernel	$m_{\mathrm{eff}}(w)$	$\mu_{\mathrm{lag}}(w)$	Radius
U200	200.000	100.500	0.086541
E100	173.124	78.103	0.093016
T200	150.374	67.333	0.099804
E50	127.299	59.303	0.108474
E25	71.578	35.785	0.144659

Thus, the conditional risk remains at 0.03 in the stationary stream and changes from 0.03 to 0.40 at $t = 1501$ in the moderate-abrupt stream.

For the recurring-shift environments, let $H \in \{100, 500\}$. The generator begins with an initial low-risk period of length 500 and then an additional low-risk block of length H, so the first low-to-high switch occurs at time $501 + H$. For $t > 500 + H$, define

$$b_t := \left\lfloor \frac{t - (501 + H)}{H} \right\rfloor.$$

The distribution path is

$$p_t = \begin{cases} p_0, & 1 \le t \le 500 + H, \\ p_1, & t > 500 + H \text{ and } b_t \text{ is even}, \\ p_0, & t > 500 + H \text{ and } b_t \text{ is odd}. \end{cases}$$

Hence, the first low-to-high switch occurs at $t = 601$ for $H = 100$ and at $t = 1001$ for $H = 500$, after which the risk alternates between 0.03 and 0.40 in blocks of length H.

Since $L = m = 200$, complete certificates are reportable from $t = 401$ onward. We use the full common reportable trajectory in every environment rather than selecting a separate post-shift evaluation interval for each temporal path.

B.1.4 Paired Monte Carlo protocol

Each environment uses 200 trajectories with seeds $r = 0, \ldots, 199$. For a given seed, all five kernels are evaluated on the same realized trajectory and therefore share the same prediction sets, clipped products, oracle targets, and true risks; only the weighted certificate terms differ. All reported comparisons are computed over the common reportable times and are first averaged within each trajectory and then equally across the 200 trajectories.

B.2 Details for Section 5.2: Risk Reporting and Certificate Comparison

The experiments in Section 5.2 have two purposes. The temporal experiment compares four historical risk summaries with the complete Oracle OBCP certificate under several risk paths. The near-boundary experiment compares the complete Oracle OBCP and Oracle Count certificates under controlled boundary crossings. The two experiments use different fixed score models but share the prediction and evaluation protocol described below.

B.2.1 Common prediction and evaluation protocol

Prediction and feedback. The covariate is constant and labels are independent over time, with $Y_t \sim \mathrm{Categorical}(p_t)$ on $\mathcal{Y} = \{1, \ldots, 5\}$. Scores $s_y := S(y)$ are deterministic and fixed within each experiment. We use $B = 3$, $n_{\mathrm{cal}} = L = m = 200$, $\tau = 1$, and $\eta = 0.05$. The first 200 labels initialize the calibration sample and reference window. Prediction begins at $t = 201$ using $\mathcal{R}_t = \{t - L, \ldots, t - 1\}$; after prediction, Y_t is revealed and the reference window advances. Complete reporting uses the strict previous-200 calendar window, so the first possible complete report is at $t = 401$.

Both score vectors used below are strictly increasing and satisfy $s_4 = 1$. Writing $V_t := \sum_{i=t-L}^{t-1} s_{Y_i}$, the online construction gives

$$R_{t,y} = \frac{(L+1)s_y}{V_t + s_y}, \quad \widetilde{\alpha}_t = \frac{V_t + 1}{L+1},$$
$$q_t(y) := \widetilde{\alpha}_t R_{t,y} = \frac{s_y(V_t+1)}{V_t + s_y}, \quad \widehat{C}_t = \{1,2,3\}.$$

Hence the exact conditional risk is $r_t^\star = \sum_{y \notin \widehat{C}_t} p_t(y)$.

Historical risk summaries. For Static BCP, each calibration point $j \le n_{\mathrm{cal}}$ is treated as a pseudo-test observation, with candidate ratios

$$R_y^{(-j)} := \frac{n_{\mathrm{cal}} s_y}{\sum_{i \le n_{\mathrm{cal}},\, i \ne j} s_{Y_i} + s_y}.$$

The backward rule gives the corresponding pseudo-level $\widetilde{\alpha}^{(-j)}$, and we report the leave-one-out estimate $\widehat{\alpha}^{\mathrm{LOO}} := n_{\mathrm{cal}}^{-1} \sum_j \widetilde{\alpha}^{(-j)}$ of Gauthier et al. (2025a), held fixed after initialization. Roll-α U200 is $m^{-1} \sum_{\ell=1}^m \widetilde{\alpha}_{t-\ell}$, Count U200 is $\widehat{r}_{t,w}^{\mathrm{count}} = m^{-1} \sum_{\ell=1}^m \mathrm{err}_{t-\ell}$, and Product U200 is $\widehat{\beta}_{t,w}^{(1),\mathrm{roll}} = m^{-1} \sum_{\ell=1}^m Q_{t-\ell}^{(1)}$. These four quantities are uncorrected historical risk summaries. Product is the observable historical component of OBCP, not the complete certificate.

Oracle OBCP and evaluation. The known label law and realized reference window give $\beta_t^\circ = \sum_y p_t(y) \min\{q_t(y), 1\}$. For U200,

$$\Gamma_{\beta,t}(w) = \sum_{q=1}^m \frac{m-q+1}{m} |\beta_{t-q+1}^\circ - \beta_{t-q}^\circ|,$$
$$\widehat{\rho}_{t,w}^{\mathrm{cert}} = \min\left\{1,\, \widehat{\beta}_{t,w}^{(1),\mathrm{roll}} + \Gamma_{\beta,t}(w) + \sqrt{\frac{\log(1/\eta)}{2m}}\right\}. \tag{9}$$

The concentration radius is approximately 0.08654. The oracle drift uses the actual sequence of conditional product targets, including changes in both p_t and the realized reference window.

For any historical summary A_t, and reportable times $\mathcal{T}_b$ within trajectory b, define

$$\mathrm{SF}_{\mathcal{T}_b}(A) := \frac{1}{|\mathcal{T}_b|} \sum_{t \in \mathcal{T}_b} (r_{t,b}^\star - A_{t,b})_+, \qquad \mathrm{OV}_{\mathcal{T}_b}(A) := \frac{1}{|\mathcal{T}_b|} \sum_{t \in \mathcal{T}_b} (A_{t,b} - r_{t,b}^\star)_+.$$

All reported means first average over reportable times within each trajectory and then equally over 200 paired trajectories. MCSEs are computed across the resulting trajectory-level quantities.

B.2.2 Historical risk summaries under temporal shift

Score model and temporal paths. We use

$$(s_1, s_2, s_3, s_4, s_5) = (0.02, 0.04, 0.06, 1.00, 1.10),$$
$$p(\theta) = \big(0.50(1-\theta),\, 0.30(1-\theta),\, 0.20(1-\theta),\, 0.60\theta,\, 0.40\theta\big), \qquad 0.08 \le \theta \le 0.16. \tag{10}$$

The scores are used directly rather than as class probabilities. Since $\widehat{C}_t = \{1,2,3\}$, the exact conditional risk is $r_t^\star = p_t(4) + p_t(5) = \theta_t$.

The shock-recovery path has $T = 4000$, with $\theta_t = 0.08$ through $t = 1200$, 0.16 over $t = 1201{:}2200$, and 0.08 over $t = 2201{:}4000$. The monotone path also has $T = 4000$, with

$$\theta_t = \begin{cases} 0.08, & 1 \le t \le 1200, \\ 0.08 + 0.08(t - 1200)/1000, & 1201 \le t \le 2200, \\ 0.16, & 2201 \le t \le 4000. \end{cases}$$

The recurring path has $T = 5000$: it remains at 0.08 through $t = 1000$ and then alternates between 0.16 and 0.08 in blocks of length $H = 500$, beginning with a high-risk block at $t = 1001$. Figure 1 shows the full reportable trajectories.

For the numerical summaries below, the shock-recovery experiment uses $t = 1201{:}1400$ for the upward-shift phase and $t = 2201{:}2400$ for recovery, while the recurring experiment uses $t = 1001{:}5000$.

Results. Table 4 reports the full numerical comparison. Product reduces the shortfall of Count during increasing-risk periods through the soft contribution in equation 3, while the same additional mass increases overestimation after downward changes.

B.2.3 Matched oracle-certificate comparison

The second experiment compares the matched Oracle OBCP and Oracle Count certificates from Section 4.3. Both routes use the same predictions, U200 weights, confidence level, concentration radius, and final truncation, while using the oracle drift of their respective conditional targets. Their difference follows directly from the surplus-drift decomposition in Corollary 3.

Near-boundary probability transfer. We use

$$\begin{aligned} s &= (0.001, 0.002, 0.995, 1.000, 1.005), \\ p_{\text{low}} &= (0.50, 0.32, 0.10, 0.06, 0.02), \\ p_{\text{high}} &= (0.50, 0.32, 0.02, 0.14, 0.02). \end{aligned}$$

Thus probability 0.08 moves from included label 3 to excluded label 4, while the score function and feasible prediction set remain unchanged. The hard risk therefore increases from 0.08 to 0.16. At a fixed reference sum V, let $\beta_{\text{low}}(V)$ and $\beta_{\text{high}}(V)$ denote the corresponding conditional product targets. Since the two boundary scores differ by only 0.005,

$$\beta_{\text{high}}(V) - \beta_{\text{low}}(V) = 0.08 \left(1 - \frac{0.995(V+1)}{V + 0.995}\right) = \frac{0.0004V}{V + 0.995} \leq 0.0004.$$

Hence this construction creates a substantial hard-risk change while keeping the direct product-target change small. The implemented oracle drift additionally reflects the realized evolution of the reference sum V_t.

Each path has $T = 5000$ and remains in the low-risk regime through $t = 600$. Beginning at $t = 601$, it alternates between high- and low-risk blocks of length $H \in \{25, 50, 200\}$, starting with high risk. Each value of H uses 200 paired trajectories. We evaluate $t = 1001{:}5000$, which contains complete high-low cycles for every H and has mean true risk 0.12.

Certificate comparison. Table 5 reports the two complete certificate values.

Oracle OBCP is smaller on average for $H = 25$ and $H = 50$, whereas Oracle Count is smaller for $H = 200$. Table 6 shows the mechanism behind this change in ordering.

The mean soft surplus remains near 0.06. For $H = 25$ and $H = 50$, the reduction from $\Gamma_{r,t}(w)$ to $\Gamma_{\beta,t}(w)$ exceeds this cost, yielding a smaller OBCP certificate. At $H = 200$, the hard-risk drift is smaller and the drift saving no longer offsets the soft surplus. Thus the experiment isolates the mechanism in Corollary 3: Product is more conservative as a historical statistic, but a smoother conditional target can make the complete OBCP certificate tighter.

C Details and Additional Results for Section 6

C.1 Dataset and chronological partitions

We use the CLEAR-10 benchmark of Lin et al. (2021), a temporally organized image-classification dataset constructed from YFCC100M imagery. CLEAR is designed to capture natural temporal evolution in real-world visual concepts through time-ordered data buckets. We use the official CLEAR-10 training and test splits. The data used here contain 39,050 images: 33,000 training

Table 4: Temporal risk-reporting results, with across-trajectory MCSE in parentheses. The upward-shift and recovery intervals are $t = 1201{:}1400$ and $t = 2201{:}2400$, respectively; recurring results use $t = 1001{:}5000$. The first four rows are historical risk summaries, while the final row is the complete Oracle OBCP certificate.

Method	Up-shift shortfall	Recovery overestimation	Recurring shortfall	Recurring overestimation
Static BCP	0.03994 (0.00135)	0.04033 (0.00139)	0.01998 (0.00068)	0.02029 (0.00073)
Roll-α U200	0.02896 (0.00098)	0.10261 (0.00135)	0.00632 (0.00011)	0.04430 (0.00028)
Count U200	0.04202 (0.00122)	0.03889 (0.00113)	0.01450 (0.00019)	0.01354 (0.00019)
Product U200	0.01803 (0.00081)	0.06846 (0.00116)	0.00416 (0.00009)	0.03414 (0.00027)
Oracle OBCP U200	0.00000 (0.00000)	0.19444 (0.00116)	0.00000 (0.00000)	0.13281 (0.00032)

Table 5: Matched complete oracle certificates under near-boundary probability transfer. Entries are means over 200 trajectories and $t = 1001{:}5000$; parentheses give across-trajectory MCSE. Both routes use U200, their respective oracle target-drift terms, and $\eta = 0.05$. Difference is Oracle OBCP minus Oracle Count.

H	Oracle OBCP	Oracle Count	Difference
25	0.26914 (0.00042)	0.52818 (0.00036)	−0.25904
50	0.26836 (0.00042)	0.36820 (0.00035)	−0.09984
200	0.26777 (0.00042)	0.24726 (0.00034)	+0.02051

Table 6: Mean components of the matched certificate difference over $t = 1001{:}5000$. The shared concentration radius cancels and neither certificate is truncated on these intervals. The final column is Oracle OBCP minus Oracle Count.

H	Soft surplus	$\Gamma_{\beta,t}$	$\Gamma_{r,t}$	OBCP minus Count
25	0.060961	0.001596	0.321600	−0.259042
50	0.060148	0.000815	0.160800	−0.099837
200	0.060488	0.000229	0.040200	+0.020517

images and 6,050 test images. The label space has eleven classes: BACKGROUND, baseball, bus, camera, cosplay, dress, hockey, laptop, racing, soccer, and sweater. Training buckets 1-10 each contain 3,300 images, and test buckets 0-10 each contain 550 images, with equal class counts within each bucket.

All classifier fitting, cross-validation, and temperature calibration use training bucket 1. Within each class, its 300 observations are assigned to 250 classifier-fitting/CV observations, 25 temperature-calibration observations, and 25 reference-reserve observations. This yields three disjoint subsets of sizes 2,750, 275, and 275. Within each class, observations are ordered by upload time, Flickr identifier interpreted as a string, and row identifier. Within training bucket 1, connected components under the union of exact-image and exact-feature duplicate relations are kept together, and two contiguous component-level cut points are chosen to match the class-specific quotas. The saved split attains these quotas exactly, and no duplicate component crosses the classifier-fitting/CV, temperature-calibration, and reference-reserve roles. Detected duplicate observations are retained rather than deleted.

All fitting, temperature calibration, and initial-reference construction are performed within training bucket 1. Test buckets 0 and 1 provide backward and same-period diagnostics, respectively. The primary forward stream consists of all 4,950 test observations in buckets 2-10. These observations are globally ordered by upload timestamp, stable string Flickr identifier, and row identifier. The ordering is fixed before online evaluation. The initial reference contains the latest 200 of the 275 reference-reserve observations under the same ordering rule and is disjoint from the deployment stream.

The duplicate audit records one exact-image match between a training-bucket 6 image and a test-bucket 6 image. The training member belongs to a training bucket outside the fitting, calibration, and initial-reference subsets used by the deployed procedure.

C.2 FROZEN REPRESENTATION AND SCORE MODEL

Images are converted to RGB and passed through an ImageNet-pretrained ResNet-18 (He et al., 2016) using the `IMAGENET1K_V1` weights. The final classification layer is replaced by the identity, giving 512-dimensional features. Preprocessing uses resizing to 256, a 224×224 center crop, bilinear interpolation with antialiasing, and channel means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. Feature extraction uses the fixed pretrained weights and deterministic preprocessing, and the resulting features are stored in float32.

The classifier is an ℓ_2-regularized multinomial logistic regression preceded by feature standardization. Hyperparameter selection uses five-fold stratified group cross-validation within the 2,750 fitting observations, keeping duplicate groups in a single fold. The scaler is fitted separately on each training fold. The grid is $C \in \{0.001, 0.01, 0.1, 1, 10, 100\}$, and the selection criterion is mean held-out negative log-likelihood. Numerical ties within 10^{-12} prefer smaller C. Cross-validation uses seed 20260808. The selected value is $C = 0.1$; the final scaler and classifier are refitted on all fitting observations. Optimization uses L-BFGS, a maximum of 2,000 iterations, and tolerance 10^{-6}.

A single positive temperature $T_{\text{temp}} = 1.10155$ is fitted on the separate 275-observation temperature subset using temperature scaling (Guo et al., 2017). The objective is negative log-likelihood of the logits divided by T_{temp}; optimization is over $\log T_{\text{temp}} \in [-10, 10]$ with tolerance 10^{-12}. The final score is $S(x, y) = -\log \max\{p_{\text{cal}}(y \mid x), 10^{-12}\}$, producing a frozen float64 matrix with 39,050 rows and 11 columns. The representation, classifier, standardization, and temperature remain fixed throughout deployment.

C.3 ONLINE STATE AND REPORTING WINDOWS

The canonical budget is $B = 3$, with reference length $L = 200$, reporting length $m = 200$, and clipping level $\tau = 1$. The externally initialized reference enables prediction from the first deployment time. At each time, the current candidate-label scores and completed reference determine the ratios, selected level, and prediction set. The historical risk summaries are formed before the current label is revealed. After revelation, the current true-label score updates the reference, and the realized loss and product enter the reporting history. A revealed observation updates the reference even if its prediction is infeasible. Both the reference and reporting histories continue across bucket boundaries.

With one-based paper time t, U200 uses exactly $t - 200, \ldots, t - 1$. Reporting requires the current prediction and all 200 preceding calendar predictions to be feasible. Calendar positions are preserved in the reporting history rather than replaced by older observations. The initial reference reserve initializes the ratios, while the reporting history is built from completed deployment predictions. Consequently, on the all-feasible stream observed here, U200 is first available at $t = 201$ and has 4,750 available times through $t = 4950$.

The stored index `stream_t` is zero-based, so it equals $t - 1$. All 4,950 canonical predictions are feasible and have set size three, with 116 realized errors.

C.4 NUMERICAL CONVENTIONS AND REPRODUCIBILITY

Prediction-set membership defines the error indicator. We retain the raw floating-point product $Q_t = \widetilde{\alpha}_t R_{t,Y_t}$. If membership and the numerical comparison $Q_t \geq 1$ disagree, their disagreement is accepted only within 10^{-12} of the threshold; a larger discrepancy raises an error. At $\tau = 1$, error events store the clipped value 1 and soft contribution 0, while covered events store $\min\{Q_t, 1\}$ and the same value as their soft contribution. This convention preserves nonnegative soft contributions and the clipped-product/error/soft decomposition while retaining the original prediction sets and errors.

The score-level replay begins from the frozen scores, labels, class mapping, initial-reference IDs, and stream ordering. It performs CPU-only online evaluation using these fixed inputs. Upstream data-processing, feature-extraction, training, and calibration scripts are retained separately; reproducing those upstream stages additionally requires the external images, feature cache, and pretrained weights. The package therefore distinguishes score-level reproduction from regeneration starting with raw images.

C.5 STATIC AND DYNAMIC SIGNALS

Static BCP uses the same $n = 200$ early labeled observations that initialize the online reference. For each observation j in this sample $\mathcal{I}_0$, the leave-one-out candidate ratio is

$$E_{j,y}^{\mathrm{LOO}} := \frac{nS(X_j, y)}{\sum_{i \in \mathcal{I}_0 \setminus \{j\}} S(X_i, Y_i) + S(X_j, y)}.$$

Applying the size-constrained level-selection rule to these ratios gives a pseudo-level $\widetilde{\alpha}_j^{\mathrm{LOO}}$. The Static BCP summary is $\widehat{\alpha}_{\mathrm{LOO}} = n^{-1} \sum_{j \in \mathcal{I}_0} \widetilde{\alpha}_j^{\mathrm{LOO}}$. At $B = 3$, all 200 pseudo-predictions are feasible and the estimate is 0.05595198817572568. We compute this LOO estimate once from the initial sample and hold it fixed as the early-period reporting baseline.

At an available reporting time, the three rolling summaries are

$$\widehat{\alpha}_t^{\mathrm{roll}} = \frac{1}{200} \sum_{\ell=1}^{200} \widetilde{\alpha}_{t-\ell}, \qquad \widehat{r}_t^{\mathrm{count}} = \frac{1}{200} \sum_{\ell=1}^{200} \mathrm{err}_{t-\ell},$$

$$\widehat{\beta}_t^{(1),\mathrm{roll}} = \frac{1}{200} \sum_{\ell=1}^{200} Q_{t-\ell}^{(1)}.$$

Roll-α tracks a moving average of the levels selected for actual past predictions. Product additionally incorporates the corresponding true-label ratios before clipping and averaging. Counting and Product use the same completed labels and satisfy

$$\widehat{\beta}_t^{(1),\mathrm{roll}} = \widehat{r}_t^{\mathrm{count}} + \widehat{s}_t^{\mathrm{soft}},$$

where $\widehat{s}_t^{\mathrm{soft}}$ averages the covered-event soft contributions over that history using the numerical convention in Appendix C.4. We evaluate all four quantities as historical risk summaries on the same CLEAR10 deployment stream.

C.6 RETROSPECTIVE AUDIT AND FIGURE CONSTRUCTION

The target in equation 6 includes the error at its starting time and the next 99 errors. Its complete future block exists through $t = 4851$. Intersecting this condition with U200 availability gives

$$\mathcal{T} = \{201, \ldots, 4851\},$$

or `stream_t`$= 200, \ldots, 4850$. Every method is evaluated on these same $4{,}651$ starting times, and all three rolling summaries use the same previous-200 history.

Let $z_t = \widehat{r}_{t,100}^{\mathrm{future}}$ and let A_t denote any of the four historical risk summaries. The underestimation rate is

$$\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbf{1}\{A_t < z_t\},$$

using a strict comparison. Average shortfall is

$$\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} [z_t - A_t]_+,$$

and worst shortfall is

$$\max_{t \in \mathcal{T}} [z_t - A_t]_+.$$

The average shortfall uses every starting time in $\mathcal{T}$, with zero contribution whenever $A_t \geq z_t$.

Static BCP, Roll-α, Count, and Product fall below the future-block empirical error at 128, 160, 1,926, and 0 starting times, respectively. The 128 static-BCP underestimations form two maximal contiguous episodes: paper times 640-708 and 3092-3150, or zero-based indices 639-707 and 3091-3149. Their lengths are 69 and 59. The figure displays all common-support starting times and enlarges these two retrospectively identified episodes with 50 observations of context on each side. It plots the original signal values at every evaluation starting time.

Adjacent future blocks share 99 of their 100 observations. We summarize the resulting sequence as overlapping local comparisons along the realized deployment trajectory. The enlarged intervals illustrate the observed static-BCP shortfall episodes, while Table 2 summarizes the entire common support. Product U200 stays at or above the retrospective audit target at every evaluated starting time, illustrating the contribution of its covered-event soft component to the observed reporting margin.

C.7 SET-BUDGET SENSITIVITY

We replay the same frozen score stream for $B \in \{2, 3, 4, 5, 6\}$, holding the model, initial-reference observations, ordering, and window lengths fixed. The canonical setting is $B = 3$; $B = 2, 4$ are planned sensitivity settings and $B = 5, 6$ are exploratory extensions. All 4,950 predictions are feasible at every tested budget, and each set has exactly the prescribed size.

Table 7: CLEAR10 set-budget sensitivity over all 4,950 deployment observations. Errors and empirical error summarize realized miscoverage, while Mean $Q_t^{(1)}$ is the average instantaneous clipped product at each budget.

B	Errors	Empirical error	Mean $Q_t^{(1)}$
2	234	0.04727	0.10853
3	116	0.02343	0.08836
4	65	0.01313	0.07605
5	38	0.00768	0.06807
6	23	0.00465	0.06172

Increasing the budget lowers the observed error frequency and mean clipped product while requiring larger reported sets. The clipped product remains above the empirical error because it retains positive mass on covered predictions. These sensitivity results compare set-size choices using the same fitted model and realized temporal stream.