SCHEDULED INFORMATION VALUE FOR DELAYED SELECTIVE LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective-label policy learning is challenging because current decisions affect both immediate payoffs and the outcomes available for future learning. This problem becomes harder under delayed feedback, where previously selected but unresolved outcomes may already provide substantial information in the near future. Because queue size alone does not reveal when such feedback will arrive, we then develop *Scheduled-aware Information Value* (SAIV) for sequential decision making under delayed selective feedback. SAIV combines a survival-based uncertainty baseline, which accounts for information expected from the pending queue, with a scheduled-information term that uses elapsed waiting times to capture timing information beyond a count-aware reference. This allows the policy to adapt exploration to both how much feedback is pending and when it is likely to arrive. Under a Beta–Bernoulli model, we derive the associated Bayes-risk reduction, establish a memoryless-delay null result, and bound sensitivity to delay-kernel estimation. Simulations and a sequential lending replay show consistent gains over exploration and delayed-feedback baselines, with mechanism experiments confirming that the improvements arise from exploiting the timing structure of unresolved feedback.

1 INTRODUCTION

Selective labels arise when a decision determines whether the corresponding outcome can be observed and used for learning (Lakkaraju et al., 2017; De-Arteaga et al., 2018). Lending is an example where repayment is observed only for approved applicants, but not for rejected ones. Similar selective feedback arises in medical testing and online recommendation. This creates a fundamental challenge for policy learning: each action affects both its immediate payoff and the information available for future decisions. Excessive exploration incurs current costs, whereas overly conservative selection may leave the learner with insufficient feedback to correct its beliefs.

The challenge is amplified when selected outcomes arrive with delay. For example, a repayment outcome may remain unresolved for months after a lending decision, forcing the learner to act while previous outcomes are still pending. Such pending feedback contains two distinct sources of information: *queue-size information*, which describes how many labels are unresolved, and *schedule information*, which describes their elapsed waiting times and hence their likely time to maturation. These sources can lead to different decisions even when the current posterior and queue size are identical. If pending labels are close to maturation, informative feedback may arrive soon without further exploration; if they are newly selected, additional exploration may be more valuable.

We develop *Scheduled-aware Information Value* (SAIV) to capture this timing structure. It uses elapsed waiting times to estimate residual maturation probabilities and the feedback expected within a planning window. SAIV combines the resulting timing-specific information with a survival-based uncertainty baseline, allowing the learner to exploit scheduled feedback without treating unresolved outcomes as observed.

Under the Beta–Bernoulli model, we characterize the resulting information quantities and their decision value. We show that scheduled-aware and count-aware information coincide under memoryless delays, so the timing correction vanishes, and we bound the score's sensitivity to delay-kernel estimation error. We evaluate SAIV in simulations and a sequential replay of longitudinal lending data.

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.

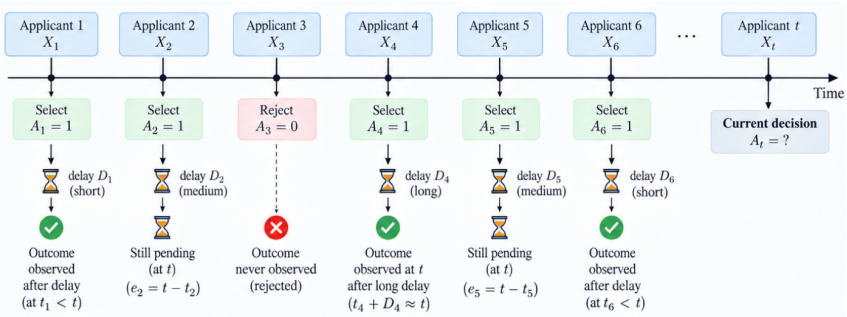


Figure 1: Illustration of sequential decisions under delayed selective feedback.

Mechanism tests show when timing changes decisions and how those changes affect continuation utility, while the memoryless-delay experiment isolates the role of elapsed-time information.

Our contributions are fourfold:

- We identify the delivery schedule of unresolved labels as a decision-relevant source of prospective information, distinct from the number of pending labels.
- We develop SAIV, which combines a survival-based uncertainty term with a timing-specific information increment relative to a count-aware reference.
- We characterize the construction under the Beta–Bernoulli model, including its local information value, memoryless-delay null result, and sensitivity to delay-kernel estimation error.
- We evaluate SAIV in controlled simulations and a sequential replay of real lending decisions, including mechanism tests that isolate the role of feedback timing.

2 Methodology

2.1 Problem Setup

We consider a finite-horizon sequential decision problem with delayed and selective feedback. At each time $t = 1, \dots, T$, the learner observes covariates $X_t \in \mathbb{R}^d$, assigns the instance to a profile $Z_t = g_\phi(X_t) \in [K] = \{1, \dots, K\}$, and chooses $A_t \in \{0, 1\}$, where $A_t = 1$ denotes selection. Conditional on $Z_t = j$, the outcome satisfies $Y_t \mid (Z_t = j) \sim \text{Bernoulli}(p^j)$, and A_t is based on the information $\mathcal{F}_t$ available immediately before the decision.

Feedback is selective and delayed: if $A_t = 1$, Y_t becomes observable after a random delay D_t; if $A_t = 0$, it is never observed. We assume D_t is conditionally independent of Y_t given the selection-time information and covariates. Thus, selected instances may remain unresolved, and their future feedback can affect subsequent decisions. Figure 1 illustrates this structure: arrived outcomes (e.g., Applicants 1 and 6) are observed, pending instances (e.g., Applicants 2 and 5) form the pending queue with elapsed times $e_\tau = t - \tau$ that determine future delivery probabilities, and never-selected instances (e.g., Applicant 3) generate no observable outcome. Further, let $\pi = \{\pi_t\}_{t=1}^T$ be a policy which specifies $A_t = \pi_t(\mathcal{F}_t)$. The learner solves

$$\max_\pi \mathbb{E}\left[\sum_{t=1}^{T} \pi_t(\mathcal{F}_t)(Y_t - c_t)\right] := \max_\pi \mathcal{V}^\pi, \tag{1}$$

where c_t is the selection cost. Each selection yields an immediate payoff and may generate future information when its label arrives; under delay, the value of this information depends on its expected arrival relative to the remaining horizon.

Since p^j is unknown, we place an independent Beta prior on it. Immediately before decision t, the posterior is $\text{Beta}(\alpha_t^j, \beta_t^j)$, with $n_t^j = \alpha_t^j + \beta_t^j$, $\mu_t^j = \alpha_t^j / n_t^j$, and $V_t^j = \mu_t^j(1 - \mu_t^j)/(n_t^j + 1)$. Here μ_t^j and V_t^j are the posterior mean and variance of p^j. The posterior is updated only when a selected label is delivered. The following sections quantify the value of a new selection when delayed feedback from earlier selections is already scheduled to arrive.

2.2 INFORMATION ON THE SCHEDULE

For profile j, let L_t be the number of pending labels at time t, and e_i the elapsed delay of the i-th pending instance. Under delay distribution G_j, the survival function is $\mathrm{Sur}_{G_j}(e) = \mathbb{P}_{G_j}(D > e)$. A pending instance i matures within the next s time units with probability $a^{j,i}_{t,G_j}(s) = \mathbb{P}_{G_j}(D \le e_i + s \mid D > e_i) = 1 - \mathrm{Sur}_{G_j}(e_i + s)/\,\mathrm{Sur}_{G_j}(e_i)$, whereas a newly selected instance has zero elapsed time and matures with probability $a^{j,0}_{t,G_j}(s) = \mathbb{P}_{G_j}(D \le s) = 1 - \mathrm{Sur}_{G_j}(s)$.

A new selection may benefit from pending feedback. We distinguish two types of future information: count-aware information, which uses only the number of pending instances, and scheduled information, which additionally accounts for their elapsed times. When m pending labels are resolved within the planning window, the effective posterior sample size increases from n^j_t to $n^j_t + m$. Under the Beta–Bernoulli model, the expected marginal reduction in posterior variance from one new selection is

$$h^j_t(m; s) = a^{j,0}_{t,G_j}(s) V^j_t \frac{n^j_t}{(n^j_t + m)(n^j_t + m + 1)}. \tag{2}$$

The count-aware and scheduled measures differ only in how this conditional quantity is averaged over the number of earlier pending labels that mature.

Count-aware future information $I^{\mathrm{cnt}}_{G_j}(s)$. The count-aware baseline discards elapsed times and assigns every pending label the same maturation probability $a^{j,i}_{t,G_j}(s) = a^{j,0}_{t,G_j}(s)$ for $1 \le i \le L_t$. Let $M^{j,\mathrm{cnt}}_t(s) \sim \mathrm{Binomial}(L_t, a^{j,0}_{t,G_j}(s))$ count the pending labels that mature within s units. Averaging equation 2 over this distribution gives

$$I^{\mathrm{cnt}}_{G_j}(s) = \sum_{m=0}^{L_t} \binom{L_t}{m} a^{j,0}_{t,G_j}(s)^m \left(1 - a^{j,0}_{t,G_j}(s)\right)^{L_t - m} h^j_t(m; s). \tag{3}$$

Scheduled-aware future information $I^{\mathrm{sch}}_{G_j}(s)$. The scheduled measure retains each pending label's elapsed time. Let $\mathbf{a}^j_{t,G_j}(s) = \left(a^{j,1}_{t,G_j}(s), \ldots, a^{j,L_t}_{t,G_j}(s)\right)$ be the heterogeneous maturation probabilities and $w^j_{m,t}(\mathbf{a}^j_{t,G_j}(s); s)$ the probability that exactly m of the L_t pending labels are observed within the next s time units. Averaging equation 2 over this scheduled distribution gives

$$I^{\mathrm{sch}}_{G_j}(s) = \sum_{m=0}^{L_t} w^j_{m,t}\left(\mathbf{a}^j_{t,G_j}(s); s\right) h^j_t(m; s). \tag{4}$$

Figure 2 illustrates why retaining elapsed times can matter even when the current posterior and queue size are fixed: two queues with the same L_t and posterior but different elapsed-time configurations imply different conditional maturation probabilities and hence different future-feedback distributions under a non-memoryless delay (see Section B for details).

Unlike $I^{\mathrm{cnt}}_{G_j}(s)$, $I^{\mathrm{sch}}_{G_j}(s)$ retains the elapsed-time configuration. We therefore define the timing-specific increment as

$$\Delta I_{G_j}(s) = \left[\sqrt{I^{\mathrm{sch}}_{G_j}(s)} - \sqrt{I^{\mathrm{cnt}}_{G_j}(s)}\right]_+. \tag{5}$$

The square root places the information terms on the posterior standard-deviation scale. When elapsed time is uninformative, the two quantities coincide and $\Delta I_{G_j}(s) = 0$. Detailed derivations are provided in Appendix C.1–C.3.

2.3 SCHEDULED-AWARE INFORMATION VALUE

The exact Bellman representation of equation 1 is given by the Bellman value function (see Theorem 2 in Appendix D.1). Let $\mathcal{V}^\pi_t(\mathcal{F}_t) = \mathbb{E}[\sum_{\tau=t}^{T} \pi_\tau(\mathcal{F}_\tau)(Y_\tau - c_\tau) | \mathcal{F}_t]$ and $\mathcal{V}^*_t(\mathcal{F}_t) = \sup_\pi \mathcal{V}^\pi_t(\mathcal{F}_t)$. Then the optimal action of equation 1 is

$$A^{*,j}_t = \mathbf{1}\{\mu^j_t + \mathcal{C}^j_t > c_t\} \qquad \textit{where}$$
$$\mathcal{C}^j_t(\mathcal{F}_t) = \mathbb{E}\left[\mathcal{V}^*_{t+1}(\mathcal{F}_{t+1}) \mid \mathcal{F}_t, Z_t = j, A_t = 1\right] - \mathbb{E}\left[\mathcal{V}^*_{t+1}(\mathcal{F}_{t+1}) \mid \mathcal{F}_t, Z_t = j, A_t = 0\right]. \tag{6}$$

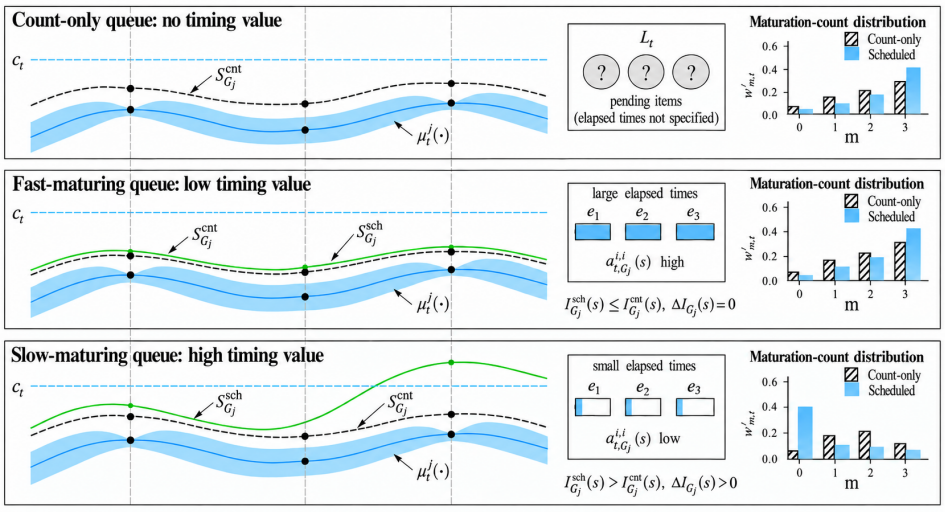


Figure 2: Schematic illustration of the timing mechanism in SAIV. Three states have the same current posterior and the same number of pending labels, but different elapsed-time configurations.

Here, $\mathcal{C}_t^j(\mathcal{F}_t)$ denotes the continuation gain from acquiring the current label through its effect on future beliefs and decisions. Its exact evaluation is generally intractable because it depends on the current posterior and the full pending-feedback state, and requires integrating over future feedback arrivals, posterior updates, and subsequent actions.

A benchmark is the posterior-based myopic policy $\pi_t^{\text{post}}(\mathcal{F}_t) = \mathbf{1}\{\mu_t^j > c_t\}$, which ignores the continuation gain when making the current decision. It provides a natural base rule for rollout, a standard approximate dynamic programming approach (Bertsekas et al., 1997; Bertsekas & Castanon, 1999). Specifically, replacing $\mathcal{V}_{t+1}^*(\cdot;\cdot)$ in equation 6 with $\mathcal{V}_{t+1}^{\pi^{\text{post}}}(\cdot;\cdot)$ defines the base-policy continuation gain $\mathcal{C}_t^{j,\text{post}}(\mathcal{F}_t)$ and yields the rollout rule $\mu_t^j + \mathcal{C}_t^{j,\text{post}}(\mathcal{F}_t) > c_t$. With exact continuation evaluation, repeated rollout weakly improves upon the myopic base policy. This representation suggests that progressively richer descriptions of pending feedback can be used to construct increasingly informative continuation adjustments.

We thus consider three successively richer decision summaries:

$$\xi_t^{\text{post}} = \left(X_t, Z_t, \alpha_t^j, \beta_t^j\right), \quad \xi_t^{\text{cnt}} = \left(\xi_t^{\text{post}}, L_t\right), \quad \xi_t^{\text{sch}} = \left(\xi_t^{\text{cnt}}, (e_i)_{i=1}^{L_t}\right).$$

The posterior-only class consists of the myopic benchmark, $\Pi^{\text{post}} = \{\pi^{\text{post}}\}$. The count-aware and schedule-aware classes allow the continuation adjustment to depend on the corresponding summary:

$$\Pi^r = \left\{\pi : \pi_t(\mathcal{F}_t) = \mathbf{1}\left\{\mu_t^j + b_t^r(\xi_t^r) > c_t\right\},\ b_t^r \text{ measurable},\ t = 1, \ldots, T\right\}, \qquad r \in \{\text{cnt}, \text{sch}\}.$$

Here, b_t^r is a real-valued continuation adjustment. The count-aware class (Π^{cnt}) can use the pending count but not individual elapsed times, whereas the schedule-aware class (Π^{sch}) can use both. All policies are evaluated under the same underlying feedback process; the distinction concerns only which information enters the decision rule.

Recall that $\mathcal{V}^\pi$ given in equation 1, and define $\mathcal{V}^{r,\star} = \sup_{\pi \in \Pi^r} \mathcal{V}^\pi, r \in \{\text{post}, \text{cnt}, \text{sch}\}$. Because a policy with richer information can always ignore the additional information and reproduce a rule based on a coarser summary, the policy classes are nested, following the standard value-of-information principle (Blackwell, 1953; De Oliveira, 2018). Proposition 1 formalizes this nesting and its value implication.

Proposition 1 (Nested continuation-aware policy classes)**.** *Suppose all policies are evaluated under the same initial distribution, outcome and delay laws, exogenous covariate process, and cost sequence. Then* $\Pi^{\text{post}} \subseteq \Pi^{\text{cnt}} \subseteq \Pi^{\text{sch}}$, *and consequently*

$$\mathcal{V}^{\text{post},\star} = \mathcal{V}^{\pi^{\text{post}}} \leq \mathcal{V}^{\text{cnt},\star} \leq \mathcal{V}^{\text{sch},\star}. \tag{7}$$

Equation 7 is a value-of-information ordering over policy classes where richer summaries weakly expand the feasible policy set, although the inequalities need not be strict. It does not imply that an arbitrary schedule-aware adjustment dominates an arbitrary count-aware one. This motivates the tractable constructions below.

We therefore construct two tractable scores using the information quantities developed in Section 2.2. For the current profile j, let $\bar{m}^j_{t,G_j}(s) = L_t a^{j,0}_{t,G_j}(s)$ denote the count-aware expected number of pending labels delivered within the planning window. We summarize the residual posterior uncertainty after these expected arrivals by $B_{G_j}(s) = \left(\frac{\mu^j_t(1-\mu^j_t)}{n^j_t+\bar{m}^j_{t,G_j}(s)+1}\right)^{1/2}$. This yields the first count-aware information value (CAIV) score

$$S^{\mathrm{cnt}}_{G_j}(\rho^j; s) = \mu^j_t + \rho^j B_{G_j}(s), \tag{8}$$

which adjusts the current posterior mean for the residual uncertainty associated with feedback expected to arrive within the planning window.

The count-aware score in equation 8 summarizes the pending schedule through the expected maturation count, but does not capture how the full distribution of future arrivals. We then incorporate the timing-specific increment $\Delta I_{G_j}(s)$ defined in Section 2.2. The resulting second score, termed scheduled-aware information value (SAIV) score, is

$$S^{\mathrm{sch}}_{G_j}(\rho^j, \gamma^j, \kappa; s) = S^{\mathrm{cnt}}_{G_j}(\rho^j; s) + \gamma^j \kappa\, \Delta I_{G_j}(s), \tag{9}$$

where $\rho^j, \gamma^j, \kappa \geq 0$. The policy selects the current instance whenever $S^{\mathrm{sch}}_{G_j}(\rho^j, \gamma^j, \kappa; s) > c_t$.

The timing-specific increment in equation 9 has two implications. First, because $\Delta I_{G_j}(s)$ is defined using the positive part, adding the timing term cannot lower the CAIV score. Second, under memoryless delays, elapsed waiting time does not change residual maturation probabilities, so the timing correction disappears. These properties are formalized in Proposition 2.

Proposition 2 (Consistency of the timing correction)**.** *For a fixed posterior, feedback queue, decision cost, and delay kernel, the following hold (the proof is deferred to Appendix D):*

1. *If $\rho^j, \gamma^j, \kappa \geq 0$, then $S^{\mathrm{sch}}_{G_j}(\rho^j, \gamma^j, \kappa; s) \geq S^{\mathrm{cnt}}_{G_j}(\rho^j; s)$.*
2. *If G_j is an uncapped geometric delay distribution, then $a^{j,i}_{t,G_j}(s) = a^{j,0}_{t,G_j}(s)$ for all $i = 1, \ldots, L_t$, and consequently $I^{\mathrm{sch}}_{G_j}(s) = I^{\mathrm{cnt}}_{G_j}(s)$. Hence, $S^{\mathrm{sch}}_{G_j}(\rho^j, \gamma^j, \kappa; s) = S^{\mathrm{cnt}}_{G_j}(\rho^j; s)$.*

2.4 Delay-Kernel Estimation, Online Computation, and Stability

In practice, G_j is unknown and must be estimated from historical feedback data (Appendix D.4). Once fitted, $\widehat{G}_j$ is kept fixed during online evaluation, and all SAIV quantities are obtained by direct substitution.

At decision time t, the learner first incorporates any arrived feedback and updates the corresponding posterior. Given $Z_t = j$ with elapsed delays $e_1, \ldots, e_{L_t}$, the probabilities $a^{j,i}_{t,\widehat{G}_j}(s)$ are evaluated using the fitted survival function. Let $w^{j,(i)}_{m,t}(\widehat{G}_j; s)$ denote the probability that exactly m of the first i pending labels mature within the planning window under $\widehat{G}_j$. The Poisson–binomial weights $w^j_{m,t}(\mathbf{a}^j_{t,\widehat{G}_j}(s); s)$ from Section 2.2 are computed by the recursion

$$w^{j,(i)}_{m,t}(\widehat{G}_j; s) = (1 - a^{j,i}_{t,\widehat{G}_j}(s)) w^{j,(i-1)}_{m,t}(\widehat{G}_j; s) + a^{j,i}_{t,\widehat{G}_j}(s) w^{j,(i-1)}_{m-1,t}(\widehat{G}_j; s), \quad w^{j,(i-1)}_{-1,t}(\widehat{G}_j; s) = 0, \tag{10}$$

with $w^{j,(0)}_{0,t}(\widehat{G}_j; s) = 1$ and $w^{j,(0)}_{m,t}(\widehat{G}_j; s) = 0$ for $m > 0$. After all L_t pending labels are processed, $w^{j,(L_t)}_{m,t}(\widehat{G}_j; s) = w^j_{m,t}(\mathbf{a}^j_{t,\widehat{G}_j}(s); s)$, $m = 0, \ldots, L_t$, and compute $S^{\mathrm{sch}}_{\widehat{G}_j}(\rho^j, \gamma^j, \kappa; s)$. The current instance is selected whenever $S^{\mathrm{sch}}_{\widehat{G}_j}(\rho^j, \gamma^j, \kappa; s) > c_t$. The recursion requires $O(L_t^2)$ arithmetic operations and $O(L_t)$ memory; the full procedure is summarized in Algorithm 1, Section E.

Since the information terms depend on the delay kernel, we study their stability under kernel estimation error. Let $\epsilon_0^j(s) = |a_{t,G_j}^{j,0}(s) - a_{t,\widehat{G}_j}^{j,0}(s)|$ denote the error for a newly selected label and $E_t^j(s) = \sum_{i=1}^{L_t} |a_{t,G_j}^{j,i}(s) - a_{t,\widehat{G}_j}^{j,i}(s)|$ the aggregate error over pending labels. Theorem 1 gives the corresponding stability result.

Theorem 1 (Stability). *Let* $\Delta_{B,t}^j(s) = \frac{\sqrt{\mu_t^j(1-\mu_t^j)}}{2(n_t^j+1)^{3/2}} L_t \epsilon_0^j(s)$, $\Delta_{A,t}^j(s) = V_t^j \frac{\epsilon_0^j(s)}{n_t^j+1} + \frac{2V_t^j E_t^j(s)}{(n_t^j+1)(n_t^j+2)}$, *and* $\Delta_{0,t}^j(s) = V_t^j \frac{\epsilon_0^j(s)}{n_t^j+1} + \frac{2V_t^j L_t \epsilon_0^j(s)}{(n_t^j+1)(n_t^j+2)}$. *Then*

$$\left|B_{G_j}(s) - B_{\widehat{G}_j}(s)\right| \le \Delta_{B,t}^j(s), \quad \left|I_{G_j}^{\mathrm{sch}}(s) - I_{\widehat{G}_j}^{\mathrm{sch}}(s)\right| \le \Delta_{A,t}^j(s), \quad \left|I_{G_j}^{\mathrm{cnt}}(s) - I_{\widehat{G}_j}^{\mathrm{cnt}}(s)\right| \le \Delta_{0,t}^j(s),$$

and consequently

$$\left|S_{G_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s) - S_{\widehat{G}_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s)\right| \le \rho^j \Delta_{B,t}^j(s) + \gamma^j \kappa \left(\sqrt{\Delta_{A,t}^j(s)} + \sqrt{\Delta_{0,t}^j(s)}\right).$$

Finally, suppose the effective number of observed outcomes for profile j satisfies $n_t^j \to \infty$. Under the Beta–Bernoulli model, $\mu_t^j \to p^j$ and $V_t^j \to 0$ almost surely. For any fitted delay kernel $\widehat{G}_j$ yielding valid maturation probabilities, $B_{\widehat{G}_j}(s) \le \sqrt{V_t^j}$, $I_{\widehat{G}_j}^{\mathrm{sch}}(s) \le \frac{V_t^j}{n_t^j+1}$. Since $\Delta I_{\widehat{G}_j}(s) \le \sqrt{I_{\widehat{G}_j}^{\mathrm{sch}}(s)}$, it follows that $0 \le S_{\widehat{G}_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s) - \mu_t^j \le \rho^j \sqrt{V_t^j} + \gamma^j \kappa \sqrt{\frac{V_t^j}{n_t^j+1}}$. Therefore, $S_{\widehat{G}_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s) \to p^j$ almost surely as feedback accumulates.

3 Simulation Experiment

3.1 Experimental Design

Consider a controlled delayed selective-label environment with K covariate profiles. At each time $t = 1, \ldots, T$, a profile $Z_t \sim \mathrm{DiscUniform}\{[K]\}$ arrives, with $Y_t \mid Z_t = j \sim \mathrm{Bernoulli}(p^j)$. The learner maintains a profile-specific prior $\mathrm{Beta}(\alpha_0^j, \beta_0^j)$, with mean $m_j^0 = \alpha_0^j/(\alpha_0^j + \beta_0^j)$, and chooses $A_t \in \{0, 1\}$. When $A_t = 1$, the outcome is revealed after a delay $D_t \sim G_{\mathrm{mix}}$; otherwise, it remains permanently unobserved. To generate heterogeneous feedback delays, we use a mixture lognormal distribution:

$$G_{\mathrm{mix}} = \begin{cases} \exp(\mu_{\mathrm{fast}} + \sigma\varepsilon_t), & \text{with probability } 0.25, \\ \exp(\mu_{\mathrm{slow}} + \sigma\varepsilon_t), & \text{with probability } 0.75, \end{cases} \qquad \varepsilon_t \sim N(0, 1).$$

In the main experiment, we set $T = 4{,}000$, and $K = 20$. Profiles are divided into six low-, eight middle-, and six high-success groups. The low and high groups are correctly specified, with prior means matching the true success probabilities. The middle group represents the selective-feedback challenge: prior means are below the decision threshold while true success probabilities are above it, so overly conservative policies may fail to acquire the feedback needed for belief correction. Exact profile parameters are reported in Appendix F.1.

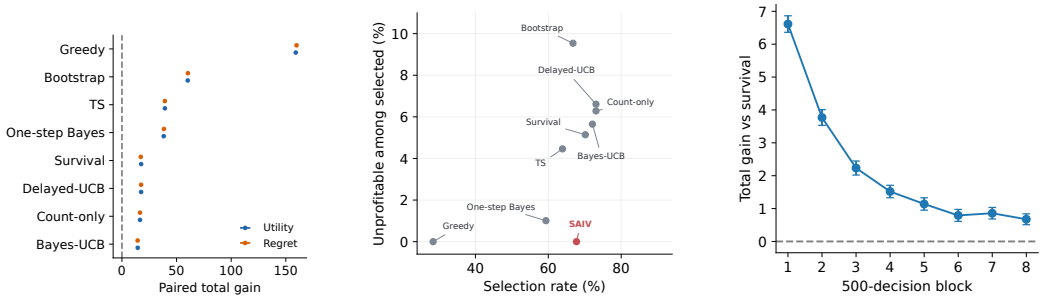
We organize the baselines into two groups. First, to examine the nested information structure characterized in Proposition 1, we compare three representative policies that progressively retain richer information. The posterior-only policy is matured greedy, the count-aware policy is CAIV, and the schedule-aware policy is proposed SAIV. Second, we evaluate SAIV against six competing sequential decision policies. Details of all policies are provided in Appendix F.2. Within each environment, all methods face the same profile sequence, potential outcomes, delays, costs, priors, and calibration data, so that performance differences are evaluated on paired stochastic environments. We assess sequential decision quality using utility (U_T) and pseudo-regret (R_T) over $T = 4{,}000$ periods:

$$U_T = \sum_{t=1}^{T} A_t(Y_t - c_t), \qquad R_T = \sum_{t=1}^{T} \left\{[p^{Z_t} - c_t]_+ - A_t(p^{Z_t} - c_t)\right\}. \tag{11}$$

We additionally report oracle agreement (relative to $A_t^\star = \mathbf{1}_{\{p^{Z_t} > c_t\}}$), the overall selection rate, and the unprofitable-selection rate (the fraction of selected instances with $p^{Z_t} \le c_t$).

Table 1: Held-out simulation results over 100 paired environment seeds; values are mean ± std. Bold marks the best utility, pseudo-regret, and oracle agreement.

Method	Utility ↑	Regret ↓	Oracle agreement (%) ↑	Selection rate (%)	Unprofitable selections (%)
Panel A: Nested information policies					
Matured greedy	194.92 ± 19.63	168.68 ± 14.41	58.45 ± 2.62	28.39 ± 2.63	0.00 ± 0.00
CAIV	337.46 ± 23.82	25.39 ± 11.00	93.89 ± 2.83	73.09 ± 3.01	6.28 ± 2.25
SAIV	$\mathbf{354.03 \pm 25.75}$	$\mathbf{8.86 \pm 12.88}$	$\mathbf{97.80 \pm 3.19}$	67.74 ± 3.33	0.00 ± 0.00
Panel B: Competing sequential decision policies					
Online bootstrap	293.83 ± 22.97	69.25 ± 7.95	84.06 ± 1.88	66.77 ± 1.98	9.54 ± 1.32
Tempered TS	314.72 ± 23.94	48.13 ± 10.87	88.23 ± 2.67	63.90 ± 2.73	4.46 ± 1.31
One-step Bayes	315.79 ± 33.60	47.28 ± 24.76	88.21 ± 6.00	59.35 ± 6.00	1.01 ± 1.41
Tuned survival	336.43 ± 26.23	26.06 ± 16.47	92.92 ± 4.32	70.16 ± 4.18	5.14 ± 2.89
Delayed-UCB	336.48 ± 24.12	26.41 ± 12.98	93.37 ± 3.42	73.09 ± 3.45	6.61 ± 2.92
Bayes-UCB	339.58 ± 24.55	23.21 ± 13.54	93.92 ± 3.56	72.11 ± 3.62	5.65 ± 2.83
SAIV	$\mathbf{354.03 \pm 25.75}$	$\mathbf{8.86 \pm 12.88}$	$\mathbf{97.80 \pm 3.19}$	67.74 ± 3.33	0.00 ± 0.00



(a) Paired gains vs. baselines. **(b)** Selection behavior. **(c)** Paired gain in successive 500-decision blocks.

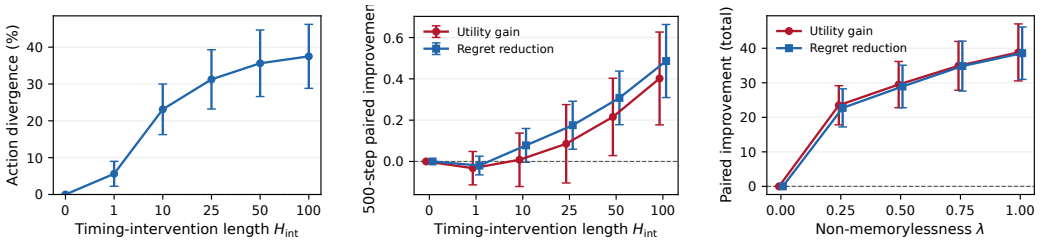
Figure 3: Sequential decision performance on 100 paired seeds.

3.2 Decision Performance

Panel A of Table 1 shows a clear monotone pattern across the three representative information levels. Utility increases from 194.92 under the posterior-only policy to 337.46 under the count-aware policy and further to 354.03 under the schedule-aware SAIV policy, while pseudo-regret decreases from 168.68 to 25.39 and then to 8.86. Oracle agreement exhibits the same pattern, increasing from 58.45% to 93.89% and 97.80%, respectively. Panel B shows that SAIV also performs favorably against the broader set of competing sequential decision policies, achieving the highest utility, the lowest pseudo-regret, and the highest oracle agreement. Figure 3a further reports paired seed-level comparisons, showing significant improvements.

Figure 3b also shows that these gains are not driven by more aggressive selection. SAIV selects 67.74% of arrivals, compared with 70.16% for tuned survival, 72.11% for Bayes-UCB, and 73.09% for delayed-UCB. Moreover, SAIV makes no unprofitable selections in the reported simulations, whereas the corresponding rates are 5.14%, 5.65%, and 6.61%, respectively. More aggressive policies acquire additional observations at the cost of more unprofitable selections, whereas conservative policies such as matured greedy forgo substantial utility. SAIV instead achieves a more favorable utility–risk trade-off by improving which instances are selected, rather than by increasing the overall selection rate.

Finally, Figure 3c examines how the advantage evolves over the learning horizon. The paired utility gain over tuned survival is largest in the first 500 decisions, at approximately 6.6, and then declines steadily to below 1 in the later blocks. Nevertheless, the gain remains positive across all eight non-overlapping 500-decision blocks. This pattern indicates that the benefit of SAIV is strongest early in learning, when feedback is relatively scarce and additional information is most valuable, while a smaller advantage persists as posterior uncertainty is gradually resolved.



(a) Timing-induced actions. **(b)** Fixed-horizon utility. **(c)** Memorylessness interpolation.

Figure 4: Mechanism of scheduled information. Error bars are seed-level 95% confidence intervals, and zero lines denote no incremental effect.

3.3 When and Why Scheduled Information Helps

We examine whether SAIV's improvement arises from the elapsed times of unresolved feedback through matched mechanism experiments. In each experiment, the treatment and control policies start from the same posterior state and pending-feedback queue and face identical future realizations. The control uses the matched CAIV policy, whereas the treatment additionally activates the timing-specific increment $\gamma^j \kappa \Delta I_{G_j}(s)$. Hence, differences between the two trajectories isolate the contribution of elapsed-time information.

Does $\Delta I_{G_j}(s)$ change decisions? For each of 40 auxiliary seeds, we take system snapshots at $t_0 \in \{500, 1000, 1500, 2000\}$ and create two identical copies. The timing term is activated in the treatment copy for the first $H_{\text{int}} \in \{0, 1, 10, 25, 50, 100\}$ decisions, after which both copies are followed over the same 500-decision window. Figure 4a reports the fraction of paired trajectories with at least one differing action during the intervention. The divergence rate is zero when timing is disabled, rises to 5.63% after one timing-aware decision, and reaches 37.50% at $H_{\text{int}} = 100$. Thus, the timing-specific increment materially changes realized decisions.

Do timing-induced decisions improve subsequent performance? We next compare cumulative utility and pseudo-regret over the full 500-decision window. The policies differ only during the first H_{int} decisions; afterward, both follow the same CAIV rule. However, early timing-induced selections can alter the pending queue, subsequent feedback arrivals, and posterior evolution, so their effects may persist beyond the intervention period.

Figure 4b shows that gains become more visible as timing-aware decisions accumulate. At $H_{\text{int}} = 100$, the treatment increases 500-decision utility by 0.402 on average, with a statistically significant improvement, accompanied by a significant reduction in pseudo-regret. These results indicate that the timing increment improves performance through its effect on subsequent learning and decisions.

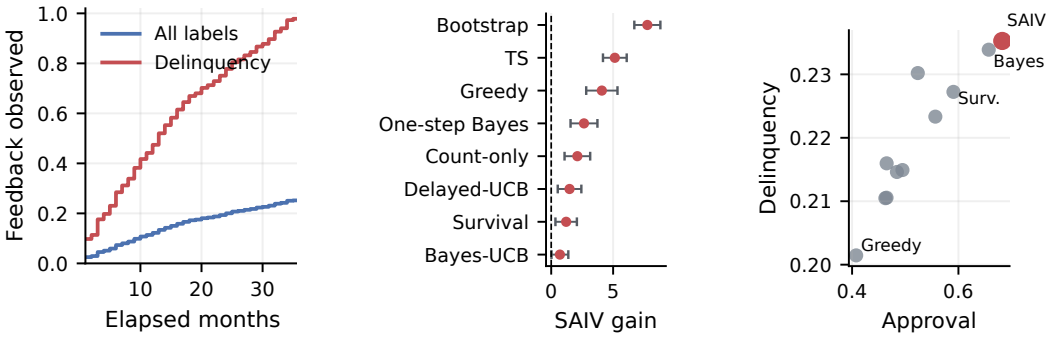
Does the timing advantage disappear when elapsed time is uninformative? Finally, we test the memoryless-delay implication of Proposition 2. Under a memoryless delay process, residual maturation probabilities are independent of elapsed time, so timing provides no additional information. We interpolate between an uncapped geometric distribution and the non-memoryless mixture used in the main experiment:

$$G_\lambda = (1 - \lambda) G_{\text{geo}} + \lambda G_{\text{mix}}, \qquad \lambda \in \{0, 0.25, 0.5, 0.75, 1\}. \tag{12}$$

At $\lambda = 0$, the scheduled-information increment is identically zero and SAIV coincides with CAIV. Figure 4c accordingly shows zero paired improvement. As the delay process departs from memorylessness, positive utility gains and regret reductions emerge and increase, supporting the interpretation that SAIV benefits specifically from heterogeneous residual maturation schedules.

4 Application on Delayed Credit Decisions

We evaluate SAIV using longitudinal credit records from a large commercial bank in China. The dataset contains 7,157 customers with monthly records from January 2016 to May 2022 and 37 application-level features. Delinquency outcomes are observed when the first delinquency occurs. Additional details are provided in Appendix G.



(a) Observed feedback over elapsed time.

(b) Paired utility gains.

(c) Approval–delinquency trade-off.

Figure 5: Experiment on Bank replay over 100 paired seeds.

The delay is substantial. Figure 5a shows the fraction of outcomes observed after different elapsed times. After 12 months, fewer than half of eventual delinquency outcomes and only about one tenth of all outcomes are available, leaving many recent decisions unresolved. This setting naturally exhibits delayed selective feedback where approval determines whether a repayment outcome can enter the learning process, while the outcome may remain unresolved for an extended period. In the sequential replay, each policy generates its own feedback path, where only approved cases produce future labels and only matured outcomes update the posterior.

We split the customers into disjoint historical, validation, and test samples, each serving a different purpose. The historical sample is used for feature preprocessing, construction of customer profiles, initialization of the profile-level priors, and estimation of the delay distribution. The validation sample is used to choose the SAIV specification and its policy parameters. At last, the test sample is used only for the final comparison between SAIV and the competing policies.

Figure 5b compares SAIV with competing policies over 100 paired replay seeds. SAIV outperforms all baseline methods significantly on the same replay environment. Specifically, SAIV achieves utility 44.68 per $1,000$ decisions, compared with 43.98 for Bayes-UCB. Figure 5c shows that the gain is not driven by accepting substantially riskier customers. SAIV approves 68.24% of customers, compared with 65.70% for Bayes-UCB, while the delinquency rates among approved customers are similar (23.53% versus 23.39%).

5 Conclusion

We study sequential decision making under delayed selective labels, where actions determine whether outcomes can ever be observed and selected outcomes may remain unresolved for substantial periods. Our key observation is that the pending queue contains more information than its size alone where elapsed waiting times determine how much feedback is already likely to arrive. We introduce scheduled information value to value a new selection relative to this scheduled feedback, and develop SAIV by combining timing-specific information with a survival-based uncertainty baseline. We derive its local Bayes-risk reduction, characterize memoryless delays as the case in which timing adds no value, and establish stability to delay-kernel estimation. Simulations and a sequential lending replay show improved decision performance, while targeted mechanism tests link these gains to the timing structure of unresolved feedback.

Several directions remain open. Our theory assumes outcome-ignorable delays, whereas real applications may exhibit informative maturation; jointly modeling outcomes and delays is therefore a natural extension. The current formulation also relies on fixed profile representations, motivating joint representation and acquisition learning in richer contextual settings. Finally, SAIV is a local Bellman-motivated rule rather than a globally optimal policy, leaving open regret analysis and more general dynamic-programming approximations, as well as principled evaluation when outcomes for historically rejected instances are unavailable.

REPRODUCIBILITY STATEMENT

We provide code, data-processing scripts, experiment configurations, random seeds, and commands for reproducing the reported results, tables, and figures in `https://anonymous.4open.science/r/Delayed-Selective-Label-4C2D`.

ETHICS STATEMENT

Automated lending affects access to financial services. The credit study uses licensed, de-identified longitudinal records from a large bank in China for a counterfactual replay; it is not a deployed decision system. Direct identifiers and protected attributes are excluded from modeling and released artifacts. Any deployment would require population-level validation, privacy and fairness assessment, adverse-action procedures, monitoring, and human oversight.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation, polishing language, and providing feedback on research methodology or experiments. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, cleaning datasets, generating synthetic data sets, interpreting results, and implementing methods, and proposing or refining hypotheses and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for identifying relevant literature and summarizing or analysing existing literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 39.1–39.26. PMLR, 2012.

Dimitri P Bertsekas and David A Castanon. Rollout algorithms for stochastic scheduling problems. *Journal of Heuristics*, 5(1):89–108, 1999.

Dimitri P Bertsekas, John N Tsitsiklis, and Cynara Wu. Rollout algorithms for combinatorial optimization. *Journal of heuristics*, 3(3):245–262, 1997.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, pp. 265–272, 1953.

Trenton Chang and Jenna Wiens. From biased selective labels to pseudo-labels: An expectation-maximization framework for learning from biased decisions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6286–6324. PMLR, 2024.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1097–1105, 2014.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Jian Chen, Zhehao Li, and Xiaojie Mao. Learning with selectively labeled data from multiple decision-makers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8480–8519. PMLR, 2025.

Stephen E. Chick, Jürgen Branke, and Christian Schmidt. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing*, 22(1):71–80, 2010.

Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency, 2018.

Henrique De Oliveira. Blackwell's informativeness theorem using diagrams. *Games and Economic Behavior*, 109:126–131, 2018.

Chenlu Ding, Jiancan Wu, Yancheng Yuan, Cunchun Li, Xiang Wang, Dingxian Wang, Frank Yang, and Andrew Rabinovich. Delayed feedback modeling with influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14648–14656, 2026.

Peter I. Frazier, Warren B. Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

Marco Gigli and Fabio Stella. Bootstrap your conversions: Thompson sampling for partially observable delayed rewards. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1438–1452. PMLR, 2024.

András György and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3988–3997. PMLR, 2021.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in generalised linear bandits revisited. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6095–6119. PMLR, 2023.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1453–1461. PMLR, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pp. 592–600. PMLR, 2012.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

Orin Levy, Liad Erez, Alon Peled-Cohen, and Yishay Mansour. Regret bounds for adversarial contextual bandits with general function approximation and delayed feedback. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Bingcong Li, Tianyi Chen, and Georgios B. Giannakis. Bandit online learning with unknown delays. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 993–1002. PMLR, 2019.

Zhongxuan Liu, Yue Kang, and Thomas C. M. Lee. Lipschitz bandits with stochastic delayed feedback. In *International Conference on Learning Representations*, 2026.

Anne Gael Manegueu, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3348–3356. PMLR, 2020.

Timothy Arthur Mann, Sven Gowal, András György, Huiyi Hu, Ray Jiang, Balaji Lakshminarayanan, and Prav Srinivasan. Learning from delayed outcomes via proxies with applications to recommender systems. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4324–4332. PMLR, 2019.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvári, and Steffen Grünewälder. Bandits with delayed, aggregated anonymous feedback. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4105–4113. PMLR, 2018.

Giovani Valdrighi, Isabel Valera, and Marcos M. Raimundo. Long-term fairness with selective labels. In *Forty-third International Conference on Machine Learning*, 2026.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11035–11046. PMLR, 2021.

Han Wu and Stefan Wager. Thompson sampling with unrestricted delays. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 937–955, 2022.

Jia-Qi Yang, Xiang Li, Shuguang Han, Tao Zhuang, De-Chuan Zhan, Xiaoyi Zeng, and Bin Tong. Capturing delayed feedback in conversion rate prediction via elapsed-time sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4582–4589, 2021.

Shota Yasui, Gota Morishita, Komei Fujita, and Masashi Shibata. A feedback shift correction in predicting conversion rates under delayed feedback. In *Proceedings of The Web Conference 2020*, pp. 2740–2746, 2020.

Tianrun Zhao, Xiaojie Mao, and Yong Liang. Online strategic classification with noise and partial feedback. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Zhengyuan Zhou, Renyuan Xu, and José Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lukas Zierahn, Dirk van der Hoeven, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and mdps. *Journal of Machine Learning Research*, 26(104):1–60, 2025.

A Related Work

Selective-label learning studies settings in which historical decisions determine which outcomes become observable. Early work by Lakkaraju et al. (2017) formalized the resulting evaluation problem and showed how heterogeneity across human decision makers can be exploited when outcomes are selectively observed. De-Arteaga et al. (2018) further used expert consistency to enlarge the region in which labels can be learned from, while subsequent work developed intervention- and pseudo-label-based approaches to mitigate selection bias (Wei, 2021; Chang & Wiens, 2024). More recent work has moved toward principled identification and sequential decision making. Chen et al. (2025) exploit heterogeneous historical decision makers as instruments, characterize exact and partial identification of classification risk, and develop a cost-sensitive learning procedure under selective labels. Zhao et al. (2025) study online classification with partial label feedback, where labels are observed only for positively classified agents, and derive sublinear-regret learning guarantees in the presence of strategic behavior. Related recent work also considers the long-run interaction between selective labels and sequential decision policies (Valdrighi et al., 2026). In general, this stream of literature addresses the endogeneity of label observation, but generally does not model the heterogeneous maturation schedules of labels generated by selected actions.

A parallel literature studies learning when feedback generated by an action is observed only after a delay. Joulani et al. (2013) provide a general analysis of delayed online learning, followed by work on stochastic and censored conversion delays, anonymously aggregated rewards, unknown delays, and arm-dependent delay distributions (Vernade et al., 2017; Pike-Burke et al., 2018; Li et al., 2019; Zhou et al., 2019; Manegueu et al., 2020; György & Joulani, 2021; Wu & Wager, 2022; Howson et al., 2023). Recent work extends delayed-feedback analysis to broader online-learning settings. Zierahn et al. (2025) provide a unified treatment of delayed feedback for combinatorial semi-bandits, linear bandits, and MDPs; Levy et al. (2025) study adversarial contextual bandits with general function approximation; and Liu et al. (2026) establish delay-aware algorithms and regret guarantees for Lipschitz bandits. A related delayed-outcome literature uses elapsed time, survival structure, or intermediate signals to predict outcomes before full maturity or to correct bias induced by incomplete feedback (Chapelle, 2014; Mann et al., 2019; Yasui et al., 2020; Yang et al., 2021; Ding et al., 2026). These methods use the delay process to improve learning from immature feedback rather than to value the future arrival schedule itself.

Our setting combines selective observation and delayed maturation, but the role of delay differs from either literature alone. Rejecting an instance prevents its label from ever being observed, whereas selecting it generates a label that enters a pending state and becomes available only after its delay has elapsed. Consequently, the learner faces a queue of unresolved labels with heterogeneous elapsed times and therefore heterogeneous residual delivery distributions. We do not impute unresolved outcomes or update the posterior before their arrival. Instead, elapsed time determines how much feedback is already likely to arrive within the planning horizon. Thus, two states with the same posterior and the same number of pending labels may have different prospective information because different amounts of feedback are already on the way.

B Illustration of the Timing Mechanism in SAIV

Figure 2 provides a schematic illustration of the key mechanism underlying SAIV, which aims to isolate the central insight that even when two decision states have the same current posterior and the same number of unresolved labels, they may imply different values of additional selection because the pending labels have different maturation schedules.

The top panel in Figure 2 shows the *count-aware queue*, which serves as the timing-free reference. The posterior mean $\mu_t^j(\cdot)$ and the unresolved queue length L_t are fixed across the subsequent comparisons. A count-aware representation uses only the number of pending labels and assigns each of them the same maturation probability as a newly selected label. The dashed score curve represents the CAIV baseline

$$S_{G_j}^{\mathrm{cnt}}(\rho^j; s) = \mu_t^j + \rho^j B_{G_j}(s).$$

The middle panel in Figure 2 considers a *fast-maturing queue*. In the illustrated setting, the pending labels have relatively high conditional maturation probabilities, so more existing feedback is expected to arrive within the planning window. The corresponding scheduled maturation distribu-

tion therefore shifts toward larger values of m. Because the marginal information value $h_t^j(m; s)$ decreases with m, we obtain

$$I_{G_j}^{\mathrm{sch}}(s) \le I_{G_j}^{\mathrm{cnt}}(s), \qquad \Delta I_{G_j}(s) = \left[\sqrt{I_{G_j}^{\mathrm{sch}}(s)} - \sqrt{I_{G_j}^{\mathrm{cnt}}(s)}\right]_+ = 0.$$

Hence, no positive timing correction is added and substantial feedback is already expected to arrive from the existing queue.

Finally, the bottom panel in Figure 2 considers a *slow-maturing queue*. Here, the pending labels have relatively low conditional maturation probabilities, so less existing feedback is expected to arrive within the planning window. The scheduled maturation distribution therefore shifts toward smaller values of m, and the direction is reversed:

$$I_{G_j}^{\mathrm{sch}}(s) > I_{G_j}^{\mathrm{cnt}}(s), \qquad \Delta I_{G_j}(s) > 0.$$

The resulting timing lift raises the full SAIV score

$$S_{G_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s) = S_{G_j}^{\mathrm{cnt}}(\rho^j; s) + \gamma^j \kappa \Delta I_{G_j}(s).$$

In the schematic, this lift moves the score above the decision threshold c_t, so timing information changes the current decision.

C Derivation of the Information Quantities

This appendix provides the mathematical justification for the information quantities introduced in Section 2.2. The key distinction is between information that is already scheduled to arrive from pending labels and information generated by a new selection. We first give the mathematical definitions and then derive their explicit forms under the Beta–Bernoulli model.

C.1 Count-aware Information

The scheduled-information quantity $I_{G_j}^{\mathrm{sch}}(s)$ retains the individual elapsed times of the currently pending labels through their profile-specific maturation probabilities $a_{t,G_j}^{j,i}(s)$. To isolate the incremental value of this timing information, we construct a count-aware reference that preserves the number of pending labels but ignores their individual elapsed-time configuration.

Under this reference, each of the L_t pending labels is assigned the same probability $a_{t,G_j}^{j,0}(s) = \mathbb{P}_{G_j}(D \le s) = 1 - \mathrm{Sur}_{G_j}(s)$ of being delivered within the next s time units. Let $M_t^{j,\mathrm{cnt}}(s)$ denote the number of pending labels delivered under this count-aware reference. Then $M_t^{j,\mathrm{cnt}}(s) \sim \mathrm{Binomial}\left(L_t, a_{t,G_j}^{j,0}(s)\right)$, and hence

$$\mathbb{P}\left(M_t^{j,\mathrm{cnt}}(s) = m \mid \mathcal{F}_t\right) = \binom{L_t}{m} a_{t,G_j}^{j,0}(s)^m \left(1 - a_{t,G_j}^{j,0}(s)\right)^{L_t - m}, \quad m = 0, \dots, L_t. \tag{13}$$

Let $Y_t^{\mathrm{cnt}}(s; M_t^{j,\mathrm{cnt}}(s))$ denote the outcomes of $M_t^{j,\mathrm{cnt}}(s)$ pending labels that are delivered under this count-aware reference. Analogously to the scheduled-information case, we define the marginal information generated by one additional selection as

$$I_{G_j}^{\mathrm{cnt}}(s) := a_{t,G_j}^{j,0}(s)\, \mathbb{E}\left[\begin{array}{l}\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{cnt}}(s; M_t^{j,\mathrm{cnt}}(s))\right) \\ - \mathbb{E}\left[\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{cnt}}(s; M_t^{j,\mathrm{cnt}}(s)), Y_t^{\mathrm{new}}\right) \middle| \mathcal{F}_t, Y_t^{\mathrm{cnt}}(s; M_t^{j,\mathrm{cnt}}(s))\right]\end{array} \middle| \mathcal{F}_t\right], \tag{14}$$

where Y_t^{new} denotes the outcome associated with the newly selected instance.

The definition above is identical in structure to $I_{G_j}^{\mathrm{sch}}(s)$. The only difference is the distribution of future feedback from the existing pending queue. The scheduled quantity uses the heterogeneous maturation probabilities $a_{t,G_j}^{j,i}(s)$ implied by the individual elapsed delays, whereas the count-aware reference assigns every pending label the common probability $a_{t,G_j}^{j,0}(s)$.

To evaluate equation 14, condition on $M_t^{j,\mathrm{cnt}}(s) = m$. Then $Y_t^{\mathrm{cnt}}(s; M_t^{j,\mathrm{cnt}}(s)) = Y_t^{\mathrm{cnt}}(s; m)$. Following similar arguments presented Section C.3, we conclude that

$$I_{G_j}^{\mathrm{cnt}}(s) = a_{t,G_j}^{j,0}(s) V_t^j \sum_{m=0}^{L_t} \binom{L_t}{m} a_{t,G_j}^{j,0}(s)^m \left(1 - a_{t,G_j}^{j,0}(s)\right)^{L_t - m} \frac{n_t^j}{(n_t^j + m)(n_t^j + m + 1)}.$$

Thus, $I_{G_j}^{\mathrm{cnt}}(s)$ measures the marginal information provided by a new selection when the existing pending queue is represented only through its size L_t, without using the elapsed-time configuration of the individual pending labels.

Comparing $I_{G_j}^{\mathrm{sch}}(s)$ and $I_{G_j}^{\mathrm{cnt}}(s)$ therefore isolates the value of elapsed-time information. The two quantities use the same posterior state, the same number of pending labels, and the same arrival probability for the newly selected label; they differ only in the distribution assigned to the maturation of the existing pending queue. Accordingly, $\left[\sqrt{I_{G_j}^{\mathrm{sch}}(s)} - \sqrt{I_{G_j}^{\mathrm{cnt}}(s)}\right]_+$ captures the additional information value attributable to the timing configuration of unresolved feedback beyond what is implied by queue size alone.

C.2 Scheduled-aware Information

In this section, we introduce the quantity *scheduled information* that are useful in determining $I_{G_j}^{\mathrm{sch}}(s)$. To start, let us consider profile j with posterior $p^j \mid \mathcal{F}_t \sim \mathrm{Beta}(\alpha_t^j, \beta_t^j)$ at time t. Recall that $n_t^j = \alpha_t^j + \beta_t^j$, $\mu_t^j = \alpha_t^j / n_t^j$, and $V_t^j = \mathrm{Var}(p^j \mid \mathcal{F}_t) = \frac{\mu_t^j(1-\mu_t^j)}{n_t^j+1} = \frac{\alpha_t^j \beta_t^j}{(n_t^j)^2(n_t^j+1)}$.

Suppose that L_t previously selected labels remain pending. For the i-th pending label, let e_i denote its elapsed delay. Its conditional probability of being delivered within the next s time units is

$$a_{t,G_j}^{j,i}(s) = \mathbb{P}_{G_j}(D \le e_i + s \mid D > e_i) = 1 - \frac{\mathrm{Sur}_{G_j}(e_i + s)}{\mathrm{Sur}_{G_j}(e_i)}.$$

Let $\mathbf{a}_{t,G_j}^j(s) = \left(a_{t,G_j}^{j,1}(s), \ldots, a_{t,G_j}^{j,L_t}(s)\right)$, and let $M_t^{j,\mathrm{sch}}(s)$ denote the number of these pending labels that are delivered within the next s time units. Under conditional independence of feedback delays across selected instances,

$$\mathbb{P}\left(M_t^{j,\mathrm{sch}}(s) = m \mid \mathcal{F}_t\right) = w_{m,t}^j\left(\mathbf{a}_{t,G_j}^j(s); s\right), \qquad m = 0, \ldots, L_t.$$

We define the scheduled posterior variance as the expected posterior uncertainty remaining after the labels that mature within the planning window have been observed:

$$V_{G_j}^{\mathrm{sch}}(s) := \mathbb{E}\left[\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; M_t^{j,\mathrm{sch}}(s))\right) \mid \mathcal{F}_t\right], \tag{15}$$

where $Y_t^{\mathrm{sch}}(s; M_t^{j,\mathrm{sch}}(s))$ denotes the outcomes of $M_t^{j,\mathrm{sch}}(s)$ pending labels that are delivered within the next s time units. To be concrete, when $M_t^{j,\mathrm{sch}}(s) = m$, then $Y_t^{\mathrm{sch}}(s; M_t^{j,\mathrm{sch}}(s)) = Y_t^{\mathrm{sch}}(s; m) = (Y_{t,1}^{\mathrm{sch}}(s), \cdots, Y_{t,m}^{\mathrm{sch}}(s))$.

By total law of expectation, equation 15 becomes:

$$V_{G_j}^{\mathrm{sch}}(s) = \sum_{m=0}^{L_t} w_{m,t}^j\left(\mathbf{a}_{t,G_j}^j(s); s\right) \mathbb{E}\left[\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; m)\right) \mid \mathcal{F}_t, M_t^{j,sch}(s) = m\right] \tag{16}$$

To evaluate equation 15, first condition on $M_t^{j,\mathrm{sch}}(s) = m$. Define $S_m = \sum_{r=1}^m Y_{t,r}^{\mathrm{sch}}(s)$.

Beta–Bernoulli conjugacy gives $p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; m) \sim \mathrm{Beta}\left(\alpha_t^j + S_m,\ \beta_t^j + m - S_m\right)$ after observing $Y_t^{\mathrm{sch}}(s; m)$. Therefore, directly using the variance of a Beta distribution,

$$\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; m)\right) = \frac{(\alpha_t^j + S_m)(\beta_t^j + m - S_m)}{(n_t^j + m)^2(n_t^j + m + 1)}.$$

Taking the posterior-predictive expectation over $Y_t^{\mathrm{sch}}(s; m)$ yields

$$\mathbb{E}\left[\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; m)\right) \middle| \mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m\right] = \frac{\mathbb{E}\left[(\alpha_t^j + S_m)(\beta_t^j + m - S_m) \mid \mathcal{F}_t\right]}{(n_t^j + m)^2(n_t^j + m + 1)}. \tag{17}$$

For the Beta–binomial predictive distribution, we have

$$\mathbb{E}[S_m \mid \mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m] = \mathbb{E}[\mathbb{E}[S_m \mid p^j, \mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m] \mid \mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m] = \frac{m\alpha_t^j}{n_t^j},$$

and

$$\begin{aligned}
&\mathbb{E}[S_m^2 \mid \mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m] \\
&= \sum_{r\neq\ell} \mathbb{E}[Y_{t,r}^{\mathrm{sch}}(s) Y_{t,\ell}^{\mathrm{sch}}(s) \mid \mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m] + \sum_{r=1}^{m} \mathbb{E}[Y_{t,r}^{\mathrm{sch}}(s) \mid \mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m] \\
&= \sum_{r\neq\ell} \mathbb{E}[(p^j)^2 \mid \mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m] + \frac{m\alpha_t^j}{n_t^j} \\
&= \frac{m(m-1)\alpha_t^j\beta_t^j}{(n_t^j)^2(n_t^j+1)} + \frac{m(m-1)(\alpha_t^j)^2}{(n_t^j)^2} + \frac{m\alpha_t^j}{n_t^j} = \frac{m(m-1)\alpha_t^j(\alpha_t^j+1)}{(n_t^j)(n_t^j+1)} + \frac{m\alpha_t^j}{n_t^j}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[(\alpha_t^j + S_m)(\beta_t^j + m - S_m) \mid \mathcal{F}_t\right] = \alpha_t^j(\beta_t^j + m) + (\beta_t^j + m - \alpha_t^j)\frac{m\alpha_t^j}{n_t^j} - \frac{m(m-1)\alpha_t^j(\alpha_t^j+1)}{n_t^j(n_t^j+1)} - \frac{m\alpha_t^j}{n_t^j} \\
&= \frac{\alpha_t^j\beta_t^j(n_t^j+1)(m+n_t^j) + m\alpha_t^j(n_t^j+1)(n_t^j+m-\alpha_t^j-1) - m(m-1)\alpha_t^j(\alpha_t^j+1)}{n_t^j(n_t^j+1)} \\
&= \frac{\alpha_t^j\beta_t^j(n_t^j+1)(m+n_t^j) + m\alpha_t^j(n_t^j+1)(\beta_t^j+m-1) - m(m-1)\alpha_t^j(\alpha_t^j+1)}{n_t^j(n_t^j+1)} \\
&= \frac{\alpha_t^j\beta_t^j(n_t^j+1)(m+n_t^j) + m\alpha_t^j(n_t^j+1)\beta_t^j + m(m-1)\alpha_t^j\beta_t^j}{n_t^j(n_t^j+1)} = \frac{\alpha_t^j\beta_t^j(n_t^j+m)(n_t^j+m+1)}{n_t^j(n_t^j+1)}.
\end{aligned}$$

Substituting this expression into equation 17 gives

$$\mathbb{E}\left[\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s;m)\right)\middle|\mathcal{F}_t, M_t^{j,\mathrm{sch}}(s) = m\right] = \frac{\alpha_t^j\beta_t^j}{n_t^j(n_t^j+1)(n_t^j+m)} = V_t^j\frac{n_t^j}{n_t^j+m}. \tag{18}$$

Substituting the above equation back to 16, we have

$$V_{G_j}^{\mathrm{sch}}(s) = V_t^j \sum_{m=0}^{L_t} w_{m,t}^j \left(\mathbf{a}_{t,G_j}^j(s); s\right) \frac{n_t^j}{n_t^j+m}.$$

Thus, $V_{G_j}^{\mathrm{sch}}(s)$ is the expected posterior uncertainty remaining after accounting for the feedback already scheduled to arrive within the planning window.

C.3 Information from a New Selection

The scheduled labels describe feedback that may arrive independently of whether the current instance is selected. We now quantify the marginal reduction in posterior uncertainty generated by selecting one additional instance.

For a newly selected instance in profile j, let $a_{t,G_j}^{j,0}(s) = \mathbb{P}_{G_j}(D \le s) = 1 - \mathrm{Sur}_{G_j}(s)$ denote the probability that its label is delivered within the next s time units. Let Y_t^{new} denote the outcome if the instance at t is selected. Because feedback timing is conditionally independent of the unseen outcome, the new observation contributes to posterior learning with probability $a_{t,G_j}^{j,0}(s)$.

We define the scheduled information from one additional selection as

$$\begin{aligned}
&I_{G_j}^{\mathrm{sch}}(s) \\
&:= a_{t,G_j}^{j,0}(s)\,\mathbb{E}\left[\begin{array}{l}\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; M_t^{j,\mathrm{sch}}(s))\right) \\ -\,\mathbb{E}\left[\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; M_t^{j,\mathrm{sch}}(s)), Y_t^{\mathrm{new}}\right)\middle|\mathcal{F}_t, Y_t^{\mathrm{sch}}(s; M_t^{j,\mathrm{sch}}(s))\right]\end{array}\middle|\mathcal{F}_t\right].
\end{aligned} \tag{19}$$

Thus, $I_{G_j}^{\mathrm{sch}}(s)$ measures the expected reduction in posterior variance produced by the new label, relative to the uncertainty that would remain after incorporating the feedback already scheduled to

arrive. Note that $\mathbb{E}\left[\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; M_t^{j,\mathrm{sch}}(s))\right)\middle|\mathcal{F}_t\right]$ in equation 19 is computed in Section C.2. We aim to compute the second quantity on the r.h.s. in equation 19.

If the newly selected label is also delivered within the planning window, its outcome provides one additional posterior-predictive Bernoulli observation. Applying the same Beta–Bernoulli calculation with $m+1$ observations gives

$$\mathbb{E}\left[\mathbb{E}\left[\mathrm{Var}\left(p^j \mid \mathcal{F}_t, Y_t^{\mathrm{sch}}(s; M_t^{j,\mathrm{sched}}(s)), Y_t^{\mathrm{new}}\right)\middle|\mathcal{F}_t, M_t^{j,\mathrm{sch}}(s)=m\right]\middle|\mathcal{F}_t\right] = V_t^j \frac{n_t^j}{n_t^j+m+1}.$$

Hence, conditional on m scheduled labels being delivered and the new label also being observed, the expected reduction in posterior variance is

$$V_t^j \frac{n_t^j}{n_t^j+m} - V_t^j \frac{n_t^j}{n_t^j+m+1} = V_t^j \frac{n_t^j}{(n_t^j+m)(n_t^j+m+1)}.$$

The new label is delivered within the planning window with probability $a_{t,G_j}^{j,0}(s)$. By total law of expectation, equation 19 can be expressed as

$$I_{G_j}^{\mathrm{sch}}(s) = a_{t,G_j}^{j,0}(s) V_t^j \sum_{m=0}^{L_t} w_{m,t}^j \left(\mathbf{a}_{t,G_j}^j(s); s\right) \frac{n_t^j}{(n_t^j+m)(n_t^j+m+1)}. \tag{20}$$

This expression shows that the information value of a new selection depends on the feedback already scheduled to arrive. When more scheduled outcomes are expected to mature within the planning window, the posterior is expected to be more concentrated after incorporating the pending labels delivered within the window, thereby reducing the marginal information provided by that additional observation.

D Proofs

D.1 Bellman Characterization of the Optimal Policy

The objective in 1 is a finite-horizon sequential decision problem. Because a selection affects not only the current utility but also the future feedback available to the learner, the exact optimal policy is characterized by the continuation value of the evolving information state.

Define

$$\mathcal{V}_t(\mathcal{F}_t) = \sup_\pi \mathbb{E}\left[\sum_{s=t}^T \pi_s(\mathcal{F}_s)(Y_s - c_s)\middle|\mathcal{F}_t\right], \quad \textit{with terminal condition: } \mathcal{V}_{T+1}(\mathcal{F}_{T+1}) = 0.$$

Theorem 2 (Bellman characterization of the optimal policy)**.** *Suppose that $\mathcal{F}_t$ is a sufficient information state for the finite-horizon delayed selective-feedback problem. Denote $\mathcal{V}_t(\mathcal{F}_t) = \sup_\pi \mathbb{E}_\pi\left[\sum_{s=t}^T A_s(Y_s - c_s)\middle|\mathcal{F}_t\right]$. Then the optimal value function satisfies*

$$\mathcal{V}_t(\mathcal{F}_t) = \max_{a\in\{0,1\}} \left\{a(\mu_t^j - c_t) + \mathbb{E}[\mathcal{V}_{t+1}(\mathcal{F}_{t+1}) \mid \mathcal{F}_t, A_t = a]\right\}.$$

Consequently, an optimal decision for an arriving instance with profile j is $A_t^{*,j} = \mathbf{1}\{\Delta_t^j(\mathcal{F}_t) > c_t\}$, where $\Delta_t^j(\mathcal{F}_t) = \mu_t^j + \mathbb{E}[\mathcal{V}_{t+1}(\mathcal{F}_{t+1}) \mid \mathcal{F}_t, Z_t = j, A_t = 1] - \mathbb{E}[\mathcal{V}_{t+1}(\mathcal{F}_{t+1}) \mid \mathcal{F}_t, Z_t = j, A_t = 0]$.

Proof. We prove the result by backward induction. At the terminal decision time T, there is no future continuation value. Conditional on $\mathcal{F}_T$ and $Z_T = j$, the posterior mean satisfies $\mathbb{E}[Y_T \mid \mathcal{F}_T, Z_T = j] = \mu_T^j$. Therefore, for an action $a \in \{0,1\}$, $\mathbb{E}[a(Y_T - c_T) \mid \mathcal{F}_T, Z_T = j] = a(\mu_T^j - c_T)$. Since $\mathcal{V}_{T+1} = 0$, $\mathcal{V}_T(\mathcal{F}_T; j) = \max_{a\in\{0,1\}} a(\mu_T^j - c_T)$, an optimal terminal action is $A_T^* = \mathbf{1}\{\mu_T^j > c_T\}$.

Now suppose that $\mathcal{V}_{t+1}$ is the optimal continuation value from time $t+1$ onward. Conditional on $\mathcal{F}_t$, taking action $A_t = a$ gives the immediate expected utility $a\,\mathbb{E}[Y_t - c_t \mid \mathcal{F}_t, Z_t = j] = a(\mu_t^j - c_t)$. The state at time $t+1$ depends on the current action because selecting the instance creates a future feedback label, whereas rejecting it does not. Thus the expected total value of action a is

$$Q_t(\mathcal{F}_t, a) = a(\mu_t^j - c_t) + \mathbb{E}[\mathcal{V}_{t+1}(\mathcal{F}_{t+1}) \mid \mathcal{F}_t, Z_t = j, A_t = a].$$

By the induction hypothesis, $\mathcal{V}_{t+1}$ already gives the maximum attainable future value after the state transition. Hence the optimal value at time t is $\mathcal{V}_t(\mathcal{F}_t; j) = \max_{a\in\{0,1\}} Q_t(\mathcal{F}_t, a)$, which gives the stated Bellman recursion.

Because there are only two possible actions, selecting $A_t = 1$ is optimal exactly when $Q_t(\mathcal{F}_t, 1) - Q_t(\mathcal{F}_t, 0) > 0$. The difference is

$$\begin{aligned} &Q_t(\mathcal{F}_t, 1) - Q_t(\mathcal{F}_t, 0) \\ &= \underbrace{\mu_t^j + \mathbb{E}[\mathcal{V}_{t+1}(\mathcal{F}_{t+1}) \mid \mathcal{F}_t, Z_t = j, A_t = 1] - \mathbb{E}[\mathcal{V}_{t+1}(\mathcal{F}_{t+1}) \mid \mathcal{F}_t, Z_t = j, A_t = 0]}_{:=\Delta_t^j(\mathcal{F}_t)} - c_t. \end{aligned}$$

Therefore, $A_t^{*;j} = \mathbf{1}\{\Delta_t^j(\mathcal{F}_t) > c_t\}$. Applying the same argument recursively from T backward to $t = 1$ proves that the resulting policy maximizes the expected cumulative utility over the finite horizon. □

Proof of Proposition 1. Fix the initial distribution, the model laws, and the deterministic cost sequence.

First, choose $b_t^{\mathrm{cnt}} \equiv 0$ for every $t = 1, \ldots, T$. The resulting count-aware policy satisfies

$$\mathbf{1}\left\{\mu_t^j + b_t^{\mathrm{cnt}}(\xi_t^{\mathrm{cnt}}) > c_t\right\} = \mathbf{1}\left\{\mu_t^j > c_t\right\} = \pi_t^{\mathrm{post}}(\mathcal{F}_t)$$

at every feasible history. Therefore, $\pi^{\mathrm{post}} \in \Pi^{\mathrm{cnt}}$, and the singleton definition of Π^{post} gives $\Pi^{\mathrm{post}} \subseteq \Pi^{\mathrm{cnt}}$.

Next, take any $\pi \in \Pi^{\mathrm{cnt}}$, represented by measurable adjustments $(b_t^{\mathrm{cnt}})_{t=1}^T$. Define schedule-aware adjustments by

$$\widetilde{b}_t^{\mathrm{sch}}\left(\xi_t^{\mathrm{cnt}}, (e_i)_{i=1}^{L_t}\right) = b_t^{\mathrm{cnt}}(\xi_t^{\mathrm{cnt}}), \qquad t = 1, \ldots, T.$$

Each $\widetilde{b}_t^{\mathrm{sch}}$ is measurable because it is the composition of b_t^{cnt} with the coordinate projection that discards elapsed times. Moreover,

$$\mathbf{1}\left\{\mu_t^j + \widetilde{b}_t^{\mathrm{sch}}(\xi_t^{\mathrm{sch}}) > c_t\right\} = \mathbf{1}\left\{\mu_t^j + b_t^{\mathrm{cnt}}(\xi_t^{\mathrm{cnt}}) > c_t\right\}$$

at every feasible history. Thus the same policy admits a schedule-aware implementation, establishing $\Pi^{\mathrm{cnt}} \subseteq \Pi^{\mathrm{sch}}$.

These embeddings preserve the decision map on full histories, not merely on histories realized under a particular policy. Starting from the common initial distribution, the embedded policy therefore induces the same action at the first decision. Given any common history, it also induces the same next action; the common transition and observation laws then yield the same distribution of the subsequent history. Induction over decision times establishes equality of the entire induced trajectory distributions, including the policy-dependent posterior and pending-queue processes. Hence each embedding preserves $\mathcal{V}^\pi$.

Finally, taking the supremum of the same payoff functional over the nested classes yields

$$\sup_{\pi\in\Pi^{\mathrm{post}}} \mathcal{V}^\pi \le \sup_{\pi\in\Pi^{\mathrm{cnt}}} \mathcal{V}^\pi \le \sup_{\pi\in\Pi^{\mathrm{sch}}} \mathcal{V}^\pi.$$

Since $\Pi^{\mathrm{post}} = \{\pi^{\mathrm{post}}\}$, its supremum equals $\mathcal{V}^{\pi^{\mathrm{post}}}$, proving equation 7. □

D.2 Proof of Proposition 2

We prove the two claims in Proposition 2 separately.

Proof of Claim 1 in Proposition 2. Subtracting $S_{G_j}^{\mathrm{cnt}}(\rho^j; s)$ from $S_{G_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s)$ yields $S_{G_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s) - S_{G_j}^{\mathrm{cnt}}(\rho^j; s) = \gamma^j \kappa \left[\sqrt{I_{G_j}^{\mathrm{sch}}(s)} - \sqrt{I_{G_j}^{\mathrm{cnt}}(s)}\right]_+$. Since $I_{G_j}^{\mathrm{sch}}(s) \ge 0$ and $I_{G_j}^{\mathrm{cnt}}(s) \ge 0$, the positive-part operator satisfies $\left[\sqrt{I_{G_j}^{\mathrm{sch}}(s)} - \sqrt{I_{G_j}^{\mathrm{cnt}}(s)}\right]_+ \ge 0$. Moreover, $\gamma^j \ge 0$ and $\kappa > 0$. Hence $S_{G_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s) - S_{G_j}^{\mathrm{cnt}}(\rho^j; s) \ge 0$, which proves $S_{G_j}^{\mathrm{sch}}(\rho^j, \gamma^j, \kappa; s) \ge S_{G_j}^{\mathrm{cnt}}(\rho^j; s)$.

Proof of Claim 2 in Proposition 2. Suppose that G_j is an uncapped geometric delay distribution with per-period delivery probability r^j. For a pending label with elapsed delay e_i, the probability that it matures within the next s time units is

$$a^{j,i}_{t,G_j}(s) = \mathbb{P}_{G_j}(D \le e_i + s \mid D > e_i) = 1 - \frac{\mathbb{P}_{G_j}(D > e_i + s)}{\mathbb{P}_{G_j}(D > e_i)}.$$

For an uncapped geometric distribution, $\mathbb{P}_{G_j}(D > e) = (1 - r^j)^e$. Therefore,

$$a^{j,i}_{t,G_j}(s) = 1 - \frac{(1-r^j)^{e_i+s}}{(1-r^j)^{e_i}} = 1 - (1-r^j)^s.$$

On the other hand, the probability that a newly selected label is observed within the same planning window is $a^{j,0}_{t,G_j}(s) = 1 - \mathrm{Sur}_{G_j}(s) = 1 - (1-r^j)^s$. Thus, $a^{j,i}_{t,G_j}(s) = a^{j,0}_{t,G_j}(s)$, for all $i = 1, \dots, L_t$. Consequently, the vector of residual maturation probabilities is $\mathbf{a}^j_{t,G_j}(s) = \big(a^{j,0}_{t,G_j}(s), \dots, a^{j,0}_{t,G_j}(s)\big)$. Hence the number of scheduled labels that mature during the next s time units, i.e., $M^{j,\mathrm{sched}}_t(s)$, follows a binomial distribution $\mathrm{Binomial}\Big(L_t, a^{j,0}_{t,G_j}(s)\Big)$. Its probability mass function is therefore

$$\binom{L_t}{m} a^{j,0}_{t,G_j}(s)^m \big(1 - a^{j,0}_{t,G_j}(s)\big)^{L_t - m}, \qquad m = 0, \dots, L_t,$$

which is exactly the probability mass function used by the count-only reference model. Thus,

$$w^j_{m,t}\Big(\mathbf{a}^j_{t,G_j}(s); s\Big) = \binom{L_t}{m} a^{j,0}_{t,G_j}(s)^m \big(1 - a^{j,0}_{t,G_j}(s)\big)^{L_t - m}.$$

Substituting this equality into the definition of the scheduled marginal information gives

$$\begin{aligned} I^{\mathrm{sch}}_{G_j}(s) &= a^{j,0}_{t,G_j}(s) V^j_t \sum_{m=0}^{L_t} w^j_{m,t}\Big(\mathbf{a}^j_{t,G_j}(s); s\Big) \frac{n^j_t}{(n^j_t + m)(n^j_t + m + 1)} \\ &= a^{j,0}_{t,G_j}(s) V^j_t \sum_{m=0}^{L_t} \binom{L_t}{m} a^{j,0}_{t,G_j}(s)^m \big(1 - a^{j,0}_{t,G_j}(s)\big)^{L_t - m} \frac{n^j_t}{(n^j_t + m)(n^j_t + m + 1)}. \end{aligned}$$

The expression on the right-hand side is exactly $I^{\mathrm{cnt}}_{G_j}(s)$. Therefore, $I^{\mathrm{sch}}_{G_j}(s) = I^{\mathrm{cnt}}_{G_j}(s)$. It follows immediately that $\Big[\sqrt{I^{\mathrm{sch}}_{G_j}(s)} - \sqrt{I^{\mathrm{cnt}}_{G_j}(s)}\Big]_+ = 0$. Substituting this equality into the SF-SIV score yields

$$S^{\mathrm{sch}}_{G_j}(\rho^j, \gamma^j, \kappa; s) = \mu^j_t + \rho^j B_{G_j}(s) + \gamma^j \kappa(0) = \mu^j_t + \rho^j B_{G_j}(s) = S^{\mathrm{cnt}}_{G_j}(\rho^j; s).$$

Thus, under an uncapped geometric delay distribution, the timing-specific correction vanishes pathwise, proving the second claim. □

D.3 Proof of Theorem 1

We divide the proofs into several parts:

Proof of bounding $|B_{G_j}(s) - B_{\widehat{G}_j}(s)|$: Recall that $\bar{m}^j_{t,G_j}(s) = \sum_{i=1}^{L_t} a^{j,0}_{t,G_j}(s)$ and $\bar{m}^j_{t,\widehat{G}_j}(s) = \sum_{i=1}^{L_t} a^{j,0}_{t,\widehat{G}_j}(s)$. By the triangle inequality,

$$\Big|\bar{m}^j_{t,G_j}(s) - \bar{m}^j_{t,\widehat{G}_j}(s)\Big| = \left|\sum_{i=1}^{L_t} \Big(a^{j,0}_{t,G_j}(s) - a^{j,0}_{t,\widehat{G}_j}(s)\Big)\right| \le L_t \epsilon^j_0(s). \tag{21}$$

For fixed μ^j_t and n^j_t, define $f(m) = \sqrt{\frac{\mu^j_t(1-\mu^j_t)}{n^j_t + m + 1}}$. Then $B_{G_j}(s) = f\Big(\bar{m}^j_{t,G_j}(s)\Big)$, $B_{\widehat{G}_j}(s) = f\Big(\bar{m}^j_{t,\widehat{G}_j}(s)\Big)$, and $|f'(m)| = \frac{\sqrt{\mu^j_t(1-\mu^j_t)}}{2(n^j_t+m+1)^{3/2}} \le \frac{\sqrt{\mu^j_t(1-\mu^j_t)}}{2(n^j_t+1)^{3/2}}$, $m \ge 0$. Hence, by the mean-value theorem and equation 21, we have

$$\Big|B_{G_j}(s) - B_{\widehat{G}_j}(s)\Big| \le \frac{\sqrt{\mu^j_t(1-\mu^j_t)}}{2(n^j_t+1)^{3/2}} \Big|\bar{m}^j_{t,G_j}(s) - \bar{m}^j_{t,\widehat{G}_j}(s)\Big| \le \frac{\sqrt{\mu^j_t(1-\mu^j_t)}}{2(n^j_t+1)^{3/2}} L_t \epsilon^j_0(s) = \Delta^j_{B,t}(s).$$

Proof of bounding $|I^{\mathrm{sch}}_{G_j}(s) - I^{\mathrm{sch}}_{\widehat{G}_j}(s)|$: Recall that

$$I^{\mathrm{sch}}_{G_j}(s) = a^{j,0}_{t,G_j}(s) V^j_t \sum_{m=0}^{L_t} w^j_{m,t}(\mathbf{a}^j_{t,G_j}(s); s) \frac{n^j_t}{(n^j_t + m)(n^j_t + m + 1)}.$$

Define $g^j_t(m) = \frac{n^j_t}{(n^j_t+m)(n^j_t+m+1)}$. Since $0 \le g^j_t(m) \le 1/(n^j_t + 1)$, we have

$$\begin{aligned}\left| I^{\mathrm{sch}}_{G_j}(s) - I^{\mathrm{sch}}_{\widehat{G}_j}(s) \right| \le & V^j_t \left| a^{j,0}_{t,G_j}(s) - a^{j,0}_{t,\widehat{G}_j}(s) \right| \sum_{m=0}^{L_t} w^j_{m,t}(\mathbf{a}^j_{t,G_j}(s); s) g^j_t(m) \\ & + V^j_t a^{j,0}_{t,\widehat{G}_j}(s) \left| \sum_{m=0}^{L_t} \left[w^j_{m,t}(\mathbf{a}^j_{t,G_j}(s); s) - w^j_{m,t}(\mathbf{a}^j_{t,\widehat{G}_j}(s); s) \right] g^j_t(m) \right|.\end{aligned}$$

The first term is bounded by $\frac{V^j_t c^j_0(s)}{n^j_t + 1}$. We now consider the second term. Denote

$$w^j_{m,t}\left(\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \cdots, \mathbf{a}^{j,i+1:L_t}_{t,G_j}(s); s \right) = w^j_{m,t}\left(a^{j,1}_{t,\widehat{G}_j}(s), \cdots, a^{j,i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \cdots, a^{j,L_t}_{t,G_j}(s); s \right)$$

for $0 \le i \le L_t$ with $w^j_{m,t}\left(\mathbf{a}^{j,1:0}_{t,\widehat{G}_j}(s), \cdots, \mathbf{a}^{j,1:L_t}_{t,G_j}(s); s \right) = w^j_{m,t}\left(a^{j,1}_{t,G_j}(s), \cdots, a^{j,L_t}_{t,G_j}(s); s \right)$ and $w^j_{m,t}\left(\mathbf{a}^{j,1:L_t}_{t,\widehat{G}_j}(s), \cdots, \mathbf{a}^{j,L_t+1:L_t}_{t,G_j}(s); s \right) = w^j_{m,t}\left(a^{j,1}_{t,\widehat{G}_j}(s), \cdots, a^{j,L_t}_{t,\widehat{G}_j}(s); s \right)$. Then we can write the second as follows:

$$\begin{aligned}& \left| \sum_{m=0}^{L_t} \left[w^j_{m,t}(\mathbf{a}^j_{t,G_j}(s); s) - w^j_{m,t}(\mathbf{a}^j_{t,\widehat{G}_j}(s); s) \right] g^j_t(m) \right| \\ &= \left| \sum_{m=0}^{L_t} \left[\sum_{i=0}^{L_t-1} w^j_{m,t}(\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+1:L_t}_{t,G_j}(s); s) - \sum_{i=1}^{L_t} w^j_{m,t}(\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+1:L_t}_{t,G_j}(s); s) \right] g^j_t(m) \right| \\ &= \left| \sum_{m=0}^{L_t} \left[\sum_{i=0}^{L_t-1} \left\{ w^j_{m,t}(\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+1:L_t}_{t,G_j}(s); s) - w^j_{m,t}(\mathbf{a}^{j,1:i+1}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s); s) \right\} \right] g^j_t(m) \right| \\ &= \left| \sum_{i=0}^{L_t-1} \left[\sum_{m=0}^{L_t} \left\{ \begin{array}{l} w^j_{m,t}(\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s); s) \\ - w^j_{m,t}(\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s); s) \end{array} \right\} g^j_t(m) \right] \right|.\end{aligned}$$

Denote M as a random variable which is the Poisson-binomial distributed in a collection of L_t independent yes/no experiments with success probabilities $p^1, \cdots, p^{L_t}$. We can then write $M = \sum_{i=1}^{L_t} M_i$ with M_i is a Bernoulli random variable with success probability p^i. Note that $\sum_{m=0}^{L_t} w^j_{m,t}(\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s); s) g^j_t(m)$ equals the expectation of $g^j_t(M)$, where M follows Poisson-binomial distribution with success probabilities $\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)$ such that $\mathbb{P}(M = m) = w^j_{m,t}(\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s); s)$. Let the corresponding expectation be $\mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(M); w^j_{m,t}(\cdot; s) \right]$. Then we have

$$\begin{aligned}& \left| \sum_{m=0}^{L_t} \left[w^j_{m,t}(\mathbf{a}^j_{t,G_j}(s); s) - w^j_{m,t}(\mathbf{a}^j_{t,\widehat{G}_j}(s); s) \right] g^j_t(m) \right| \\ & \le \sum_{i=0}^{L_t-1} \left| \begin{array}{l} \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(M); w^j_{m,t}(\cdot; s) \right] \\ - \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(M); w^j_{m,t}(\cdot; s) \right] \end{array} \right|.\end{aligned}$$

In the above equation, the Poisson-binomial variable M in the first expectation on the r.h.s. can be expressed as $\sum_{\ell=1}^{i} M_\ell + M_{i+1} + \sum_{\ell=i+2}^{L_t} M_\ell$ with $M_\ell \sim \mathrm{Bern}(a^{j,\ell}_{t,\widehat{G}_j})$ for $1 \le \ell \le i$ and $M_\ell \sim \mathrm{Bern}(a^{j,\ell}_{t,G_j})$ for $i+1 \le \ell \le L_t$, while the Poisson-binomial variable M in the

second expectation on the r.h.s. can be expressed as $\sum_{\ell=1}^{i} M_\ell + M_{i+1} + \sum_{\ell=i+2}^{L_t} M_\ell$ with $M_\ell \sim \text{Bern}(a^{j,\ell}_{t,\widehat{G}_j})$ for $1 \le \ell \le i+1$ and $M_\ell \sim \text{Bern}(a^{j,\ell}_{t,G_j})$ for $i+2 \le \ell \le L_t$. Consider $\mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(M); w^j_{m,t}(\cdot; s) \right]$. We have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(M); w^j_{m,t}(\cdot; s) \right] \\
&= \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[\mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(M); w^j_{m,t}(\cdot; s) \middle| M_{i+1} \right] \right] \\
&= \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[\mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,G_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(\sum_{\substack{\ell=1,\\ \ell\neq i+1}}^{L_t} M_\ell + M_{i+1}); w^j_{m,t}(\cdot; s) \middle| M_{i+1} \right] \right] \\
&= (1 - a^{j,i+1}_{t,G_j}(s)) \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(\sum_{\substack{\ell=1,\\ \ell\neq i+1}}^{L_t} M_\ell); w^j_{m,t}(\cdot; s) \right] \\
&\quad + a^{j,i+1}_{t,G_j}(s) \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(\sum_{\substack{\ell=1,\\ \ell\neq i+1}}^{L_t} M_\ell + 1); w^j_{m,t}(\cdot; s) \right].
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), a^{j,i+1}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(M); w^j_{m,t}(\cdot; s) \right] \\
&= (1 - a^{j,i+1}_{t,\widehat{G}_j}(s)) \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(\sum_{\substack{\ell=1,\\ \ell\neq i+1}}^{L_t} M_\ell); w^j_{m,t}(\cdot; s) \right] \\
&\quad + a^{j,i+1}_{t,\widehat{G}_j}(s) \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[g^j_t(\sum_{\substack{\ell=1,\\ \ell\neq i+1}}^{L_t} M_\ell + 1); w^j_{m,t}(\cdot; s) \right].
\end{aligned}
$$

As such, we have:

$$
\begin{aligned}
&\left| \sum_{m=0}^{L_t} \left[w^j_{m,t}(\mathbf{a}^j_{t,G_j}(s); s) - w^j_{m,t}(\mathbf{a}^j_{t,\widehat{G}_j}(s); s) \right] g^j_t(m) \right| \\
&\le \sum_{i=0}^{L_t - 1} \left| a^{j,i+1}_{t,G_j}(s) - a^{j,i+1}_{t,\widehat{G}_j}(s) \right| \mathbb{E}_{\mathbf{a}^{j,1:i}_{t,\widehat{G}_j}(s), \mathbf{a}^{j,i+2:L_t}_{t,G_j}(s)} \left[\left| g^j_t(M) - g^j_t(M+1) \right|; w^j_{m,t}(\cdot; s) \right] \\
&\le \frac{2}{(n^j_t + 1)(n^j_t + 2)} \sum_{i=1}^{L_t} \left| a^{j,i}_{t,G_j}(s) - a^{j,i}_{t,\widehat{G}_j}(s) \right|.
\end{aligned}
$$

Thus, we have

$$
\left| I^{\text{sch}}_{G_j}(s) - I^{\text{sch}}_{\widehat{G}_j}(s) \right| \le \frac{V^j_t \epsilon^j_0(s)}{n^j_t + 1} + \frac{2 V^j_t E^j_t(s)}{(n^j_t + 1)(n^j_t + 2)} = \Delta^j_{A,t}(s).
$$

Proof of bounding $|I^{\text{cnt}}_{G^j}(s) - I^{\text{cnt}}_{\widehat{G}_j}(s)|$: The count-aware information term depends on the delay kernel only through the common arrival probability $a^{j,0}_{t,G_j}(s)$. Therefore, replacing $a^{j,0}_{t,G_j}(s)$ by $a^{j,0}_{t,\widehat{G}_j}(s)$ changes both the probability of the new label arriving and the binomial distribution of the scheduled labels.

Let $b_t^j(m;q) = \binom{L_t}{m} q^m (1-q)^{L_t-m}$. Then $I^{\mathrm{cnt}}_{G_j}(s) = a^{j,0}_{t,G_j}(s) V_t^j \sum_{m=0}^{L_t} b_t^j(m; a^{j,0}_{t,G_j}(s)) g_t^j(m)$. Applying similar arguments presented in bounding $\left|I^{\mathrm{sch}}_{G_j}(s) - I^{\mathrm{sch}}_{\widehat{G}_j}(s)\right|$, we conclude that

$$\left|I^{\mathrm{cnt}}_{G_j}(s) - I^{\mathrm{cnt}}_{\widehat{G}_j}(s)\right| \le \frac{V_t^j \epsilon_0^j(s)}{n_t^j + 1} + \frac{2V_t^j L_t \epsilon_0^j(s)}{(n_t^j+1)(n_t^j+2)} = \Delta^j_{0,t}(s).$$

Proof of bounding $|S^{\mathrm{sch}}_{G_j}(\rho^j,\gamma^j,\kappa;s) - S^{\mathrm{sch}}_{\widehat{G}_j}(\rho^j,\gamma^j,\kappa;s)|$: Using the definition of the SAIV score, we have

$$\left|S^{\mathrm{sch}}_{G_j}(\rho^j,\gamma^j,\kappa;s) - S^{\mathrm{sch}}_{\widehat{G}_j}(\rho^j,\gamma^j,\kappa;s)\right| \le \rho^j \left|B_{G_j}(s) - B_{\widehat{G}_j}(s)\right| + \gamma^j \kappa \left|\Delta I_{G_j}(s) - \Delta I_{\widehat{G}_j}(s)\right|.$$

The map $x \mapsto [x]_+$ is 1-Lipschitz, and for $x, y \ge 0$, $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$. Therefore,

$$\begin{aligned}\left|\Delta I_{G_j}(s) - \Delta I_{\widehat{G}_j}(s)\right| &\le \left|\sqrt{I^{\mathrm{sch}}_{G_j}(s)} - \sqrt{I^{\mathrm{sch}}_{\widehat{G}_j}(s)}\right| + \left|\sqrt{I^{\mathrm{cnt}}_{G_j}(s)} - \sqrt{I^{\mathrm{cnt}}_{\widehat{G}_j}(s)}\right| \\ &\le \sqrt{\left|I^{\mathrm{sch}}_{G_j}(s) - I^{\mathrm{sch}}_{\widehat{G}_j}(s)\right|} + \sqrt{\left|I^{\mathrm{cnt}}_{G_j}(s) - I^{\mathrm{cnt}}_{\widehat{G}_j}(s)\right|} \le \sqrt{\Delta^j_{A,t}(s)} + \sqrt{\Delta^j_{0,t}(s)}.\end{aligned}$$

As a result, we have

$$\left|S^{\mathrm{sch}}_{G_j}(\rho^j,\gamma^j,\kappa;s) - S^{\mathrm{sch}}_{\widehat{G}_j}(\rho^j,\gamma^j,\kappa;s)\right| \le \rho^j \Delta^j_{B,t}(s) + \gamma^j \kappa \left(\sqrt{\Delta^j_{A,t}(s)} + \sqrt{\Delta^j_{0,t}(s)}\right).$$

The claim is proved.

Remark 1. *We give a sufficient condition for $E_t^j(s)$. For a pending label with elapsed delay e_i,*

$$a^{j,i}_{t,G_j}(s) = 1 - \frac{\mathrm{Sur}_{G_j}(e_i+s)}{\mathrm{Sur}_{G_j}(e_i)}, \qquad a^{j,i}_{t,\widehat{G}_j}(s) = 1 - \frac{\mathrm{Sur}_{\widehat{G}_j}(e_i+s)}{\mathrm{Sur}_{\widehat{G}_j}(e_i)}.$$

Let $S_i = \mathrm{Sur}_{G_j}(e_i)$, $\widehat{S}_i = \mathrm{Sur}_{\widehat{G}_j}(e_i)$. Define $S_i^+ = \mathrm{Sur}_{G_j}(e_i+s)$ and $\widehat{S}_i^+ = \mathrm{Sur}_{\widehat{G}_j}(e_i+s)$. Suppose $S_i > s_{\min} > \delta > 0$. Under the assumed uniform survival-function bound, $|S_i - \widehat{S}_i| \le \delta$, $|S_i^+ - \widehat{S}_i^+| \le \delta$. Furthermore, $S_i \ge s_{\min}$, $\widehat{S}_i \ge S_i - \delta \ge s_{\min} - \delta > 0$. Therefore,

$$|a^{j,i}_{t,G_j}(s) - a^{j,i}_{t,\widehat{G}_j}(s)| = \left|\frac{S_i^+}{S_i} - \frac{\widehat{S}_i^+}{\widehat{S}_i}\right| \le \frac{|S_i^+ - \widehat{S}_i^+|}{S_i} + \frac{\widehat{S}_i^+ |S_i - \widehat{S}_i|}{S_i \widehat{S}_i} \le \frac{\delta}{s_{\min}} + \frac{\delta}{s_{\min}} = \frac{2\delta}{s_{\min}}.$$

Since probabilities differ by at most one,

$$|a^{j,i}_{t,G_j}(s) - a^{j,i}_{t,\widehat{G}_j}(s)| \le \min\left\{1, \frac{2\delta}{s_{\min}}\right\}.$$

Summing over the L_t pending labels gives

$$E_t^j(s) = \sum_{i=1}^{L_t} |a^{j,i}_{t,G_j}(s) - a^{j,i}_{t,\widehat{G}_j}(s)| \le L_t \min\left\{1, \frac{2\delta}{s_{\min}}\right\}.$$

This proves the stated sufficient condition and completes the proof.

D.4 Delay-Kernel Estimation

The online procedure requires an estimate of the delay survival function for each profile. We obtain this estimate from a historical calibration sample using a nonparametric product-limit estimator, which naturally handles feedback delays that are not yet observed at the end of the calibration period.

The calibration data record the elapsed time until feedback is observed and, when feedback has not yet arrived by the end of the calibration period, the corresponding observation is treated as right-censored. For profile j, suppose the calibration sample contains N^j_{cal} selected instances. For

calibration instance r, let $U_r^j = \min\{D_r^j, C_r^j\}$, $\Delta_r^j = \mathbf{1}\{D_r^j \le C_r^j\}$, where D_r^j is the true feedback delay and C_r^j is the amount of follow-up available for that instance in the calibration data. Thus, $\Delta_r^j = 1$ indicates that the feedback delay is observed, whereas $\Delta_r^j = 0$ indicates right censoring at C_r^j.

Let $0 < u_1^j < \cdots < u_{K_j}^j$ denote the distinct observed delay times among instances with $\Delta_r^j = 1$. Define $d_k^j = \sum_{r=1}^{N_{\text{cal}}^j} \mathbf{1}\{U_r^j = u_k^j, \Delta_r^j = 1\}$ as the number of feedback events at u_k^j, and $Y_k^j = \sum_{r=1}^{N_{\text{cal}}^j} \mathbf{1}\{U_r^j \ge u_k^j\}$ as the number of instances still at risk immediately before u_k^j.

The estimated survival function is then $\mathrm{Sur}_{\widehat{G}_j}(e) = \prod_{k:u_k^j \le e} \left(1 - \frac{d_k^j}{Y_k^j}\right)$. Equivalently, the estimated delay distribution is obtained as $\widehat{G}_j(e) = 1 - \mathrm{Sur}_{\widehat{G}_j}(e)$. The estimator is evaluated only at elapsed times required by the online decision rule. In particular, for a pending label with elapsed delay e_i, the conditional maturation probability is computed as $a_{t,\widehat{G}_j}^{j,i}(s) = 1 - \frac{\mathrm{Sur}_{\widehat{G}_j}(e_i+s)}{\mathrm{Sur}_{\widehat{G}_j}(e_i)}$, provided that $\mathrm{Sur}_{\widehat{G}_j}(e_i) > 0$. For a newly selected instance, $a_{t,\widehat{G}_j}^{j,0}(s) = 1 - \mathrm{Sur}_{\widehat{G}_j}(s)$.

If the delay distribution is assumed to be common across profiles, the same procedure is applied to the pooled calibration sample to obtain a single $\widehat{G}$. If delay behavior is profile-specific, the procedure is applied separately for each j to obtain $\widehat{G}_j$. The calibration sample is fixed before online evaluation. Consequently, the online procedure does not re-estimate the delay kernel after every decision; it only evaluates the fitted survival function at the elapsed times required by the current feedback queue.

E ALGORITHM

This section gives the algorithm of SAIV proposed in the paper. Please see Algorithm 1.

Algorithm 1 Scheduled-aware Information Value (SAIV)

Require: Calibration data, initial Beta parameters $\{(\alpha_0^j, \beta_0^j)\}_{j=1}^K$, cost sequence $\{c_t\}_{t=1}^T$, planning horizon s, and score parameters $\{\rho^j, \gamma^j\}_{j=1}^K, \kappa$

Ensure: Decisions $\{A_t\}_{t=1}^T$

1: Estimate $\widehat{G}_j$ from the calibration data for each profile j using Appendix D.4.
2: Initialize the posterior parameters (α_0^j, β_0^j) and the pending feedback queues.
3: **for** $t = 1, \ldots, T$ **do**
4: Process all feedback that arrives at time t and update the corresponding Beta posterior parameters.
5: Observe the arriving instance and assign it to profile $Z_t = j$.
6: Retrieve the L_t pending labels for profile j and their elapsed delays $e_1, \ldots, e_{L_t}$.
7: Compute $a_{t,\widehat{G}_j}^{j,i}(s)$ for $i = 1, \ldots, L_t$ and $a_{t,\widehat{G}_j}^{j,0}(s)$ from $\mathrm{Sur}_{\widehat{G}_j}$.
8: Initialize $w_{0,t}^{j,(0)}(\widehat{G}_j; s) \leftarrow 1$ and $w_{m,t}^{j,(0)}(\widehat{G}_j; s) \leftarrow 0$ for $m > 0$.
9: **for** $i = 1, \ldots, L_t$ **do**
10: Update $w_{m,t}^{j,(i)}(\widehat{G}_j; s)$ using equation 10.
11: Compute $I_{\widehat{G}_j}^{\text{sch}}(s)$, $I_{\widehat{G}_j}^{\text{cnt}}(s)$, and $B_{\widehat{G}_j}(s)$ using the definitions in Section 2.2 and 2.3.
12: Compute $S_{\widehat{G}_j}^{\text{sch}}(\rho^j, \gamma^j, \kappa; s)$ using Section 2.3.
13: Set $A_t \leftarrow \mathbf{1}\left\{ S_{\widehat{G}_j}^{\text{sch}}(\rho^j, \gamma^j, \kappa; s) > c_t \right\}$.
14: **if** $A_t = 1$ **then**
15: Add the selected label to the appropriate pending feedback queue.

F Additional Simulation Experimental Details

F.1 Simulation DGP Details

The main simulation uses $K = 20$ profiles. Their prior means and true success probabilities are specified as follows.

For the six low-success profiles, $m_{0,\text{low}} = (0.48, 0.52, 0.56, 0.60, 0.62, 0.64)$, and the true probabilities satisfy $p_{\text{low}} = m_{0,\text{low}}$.

For the six high-success profiles, $m_{0,\text{high}} = (0.82, 0.84, 0.86, 0.88, 0.90, 0.92)$, with $p_{\text{high}} = m_{0,\text{high}}$.

For the remaining eight middle profiles, the historical prior means are evenly spaced over $m_{0,\text{mid}} \in [0.668, 0.682]$, whereas their true success probabilities are evenly spaced over $p_{\text{mid}} \in [0.793, 0.807]$.

Decision costs is 0.70. Hence, the middle-profile priors initially lie below the decision threshold while their true success probabilities lie above it, generating the prior misspecification used to induce exploration under selective feedback.

F.2 Introduction to baselines

The comparator set spans exploitation, posterior optimism, posterior sampling, bootstrap exploration, myopic value of information, and delay-aware scoring. Several policies are deliberately simplified implementations of these ideas so that every method uses the same Beta–Bernoulli state and differs only in how it acts on delivered or pending feedback.

Matured greedy.
Selects when the current posterior mean exceeds cost; pending labels have no decision role. It is the delayed-feedback analogue of pure greedy exploitation.

CAIV.
Using equation 8, this constructed ablation represents delay-aware methods that summarize outstanding feedback by its count rather than its residual schedule.

Bayes-UCB.
Uses the 0.75 posterior quantile, following the Bayesian upper-confidence principle of Kaufmann et al. (2012). The fixed, non-oracle quantile provides moderate optimism.

Delayed-UCB.
Adds a 0.05 posterior uncertainty bonus while updating only from delivered outcomes, paralleling UCB adaptations for delayed feedback (Joulani et al., 2013).

Tempered TS.
Samples from the delayed posterior with temperature 0.25; five independent policy streams are averaged within seed. This is a tempered implementation of Thompson sampling (Chapelle & Li, 2011; Agrawal & Goyal, 2012) under delayed observations (Wu & Wager, 2022).

Online bootstrap.
Uses 101 selective bootstrap members and five policy repeats, updating only on delivered selected labels. It instantiates bootstrap posterior sampling for partially observed delayed rewards (Gigli & Stella, 2024).

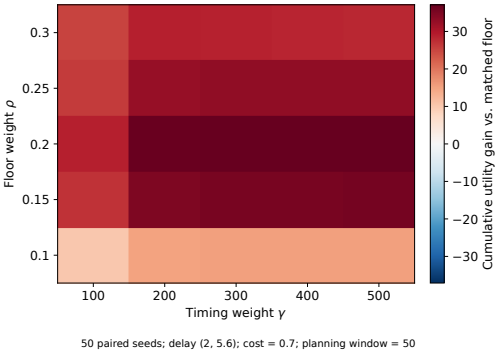
One-step Bayes.
Uses the exact immediate Beta–Bernoulli value of one label with coefficient four, without modeling its delivery schedule. It is a myopic value-of-information comparator (Frazier et al., 2008; Chick et al., 2010).

Tuned survival.
Uses expected scheduled arrivals $\bar{m}_G$ with exploration coefficient 0.6 but no elapsed-time increment. This constructed reference draws on survival-based modeling of pending delayed conversions (Chapelle, 2014; Vernade et al., 2017).

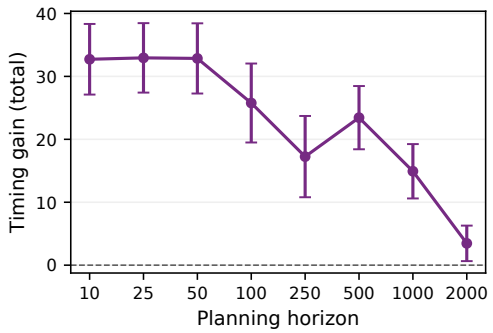
F.3 SENSITIVITY ANALYSES

We examine whether the performance of SAIV depends on a narrowly chosen coefficient pair, a particular planning horizon, or exact knowledge of the delay distribution.

Does the timing gain depend on a narrow choice of (ρ, γ)? We first vary the coefficients governing the count-aware uncertainty term and the scheduled-timing term. Specifically, we consider $\rho \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$ and $\gamma \in \{100, 200, 300, 400, 500\}$. For each pair (ρ, γ), we compare SAIV with a matched CAIV policy using the same value of ρ and the same environment seeds. Thus, each comparison isolates the incremental effect of the scheduled-timing term while holding the count-aware uncertainty adjustment fixed.



(a) Sensitivity to (ρ, γ).



(b) Sensitivity to the planning horizon s.

Figure 6: Sensitivity analyses on 50 paired auxiliary seeds.

Figure 6(a) shows that the timing-enabled policy achieves positive utility gains throughout the entire parameter grid. The gains are particularly large over a broad region with $\rho \in [0.15, 0.25]$ and $\gamma \geq 200$. Increasing the timing weight beyond this range produces relatively small additional changes. The timing advantage is robust to substantial variation in both coefficients, indicating that the benefit of scheduled information is not driven by a narrowly tuned (ρ, γ) specification.

Does the result depend on the planning horizon? We vary the planning horizon over $s \in \{10, 25, 50, 100, 250, 500, 1000, 2000\}$ while keeping all other policy parameters fixed. Figure 6(b) shows that the timing gain is large and stable for short planning horizons, remaining around 33 for $s = 10$, 25, and 50. It then decreases as the horizon expands. However, even at $s = 1000$, the timing gain remains positive and economically meaningful, while it becomes small only at the longest horizon, $s = 2000$.

This result is robust over a broad range of planning horizons but weakens when the window becomes very long. A moderate planning window preserves differences in the residual maturation schedules of pending outcomes, allowing elapsed-time information to affect the scheduled information term. As s becomes sufficiently large, an increasing share of pending outcomes can mature within the window regardless of their current elapsed times, reducing the distinction between different queue schedules. This explains the broad decline at long horizons.

Is SAIV robust to an estimated delay kernel? The preceding experiments use the true delay distribution. In practice, however, SAIV must compute scheduled information using an estimated delay kernel. We therefore estimate the lognormal mixture distribution from independent calibration samples of size N_{cal} and compare the resulting policy with an oracle-kernel policy that uses the true delay distribution.

As the calibration sample grows from $N_{\text{cal}} = 25$ to 1,000, the CDF error decreases from 0.1676 to 0.0276, while disagreement with the oracle-kernel policy falls from 0.94% to 0.11%. Thus, more accurate delay-kernel estimation produces a policy that is progressively closer to the oracle-kernel policy. At the same time, decision performance remains stable across calibration sizes. Utility stays

Table 2: Sensitivity to delay-kernel estimation.

N_{cal}	CDF error	Disagreement (%)	Utility	Regret
25	0.1676 ± 0.0500	0.94 ± 2.82	351.10 ± 28.12	13.28 ± 15.28
50	0.1125 ± 0.0384	0.48 ± 1.55	353.23 ± 25.57	11.40 ± 13.30
100	0.0840 ± 0.0293	0.40 ± 1.15	354.14 ± 24.58	10.16 ± 11.33
250	0.0520 ± 0.0171	0.09 ± 0.51	354.64 ± 23.82	9.77 ± 10.91
500	0.0379 ± 0.0113	0.11 ± 0.55	355.00 ± 24.56	9.63 ± 11.00
1000	0.0276 ± 0.0101	0.11 ± 0.55	355.05 ± 24.58	9.62 ± 11.00
oracle	0.0000 ± 0.0000	0.00 ± 0.00	355.15 ± 24.23	9.50 ± 10.98

between 351.10 and 355.05, and pseudo-regret between 13.28 and 9.62. This pattern is consistent with Theorem 1, which does not require realized utility or regret to improve monotonically with N_{cal}.

G Additional Empirical Experimental Details

The analysis sample contains 7,157 customers from a licensed longitudinal data set supplied by a large commercial bank in China. Each customer has monthly account records from January 2016 through May 2022. The decision-time representation contains 31 application and contract quantities and six account balances measured at baseline. The binary outcome is 1 if the customer remains non-delinquent and zero otherwise. Delinquent outcomes arrive in the first month in which delinquency is recorded.

Since all customers in the dataset were historically approved, their realized outcomes are available for offline evaluation. We replay lending decisions sequentially for each policy. If a policy approves a customer, that customer's outcome becomes available to the policy only after the corresponding delay has passed. If the policy rejects the customer, the outcome is never used to update that policy. Thus, although the offline evaluator observes the realized outcomes for all customers, each policy learns only from customers it approves and whose outcomes have already been observed.

Table 3 shows the delayed-feedback problem. Only 10.5% of labels have arrived by month 12 and 18.0% by month 24. Among customers who become delinquent, the median event month is 14 and the interquartile range is 6–24. Thus, even after many decisions have been made, a large share of accepted cases remains pending, and pending cases differ materially in how close they are to resolution.

Table 3: Outcome and feedback-time characteristics of the Bank analysis sample. The three partitions are customer-disjoint. "Feedback observed" is the fraction of outcomes available by the indicated elapsed month.

Panel A: outcomes by partition

Partition	n	Delinquent n	Delinquent (%)
Historical	2,862	677	23.65
Validation	2,862	648	22.64
Test	1,433	369	25.75
Total	7,157	1,694	23.67

Panel B: feedback times

Outcome	n	Mean	Q1	Median	Q3
Delinquent	1,694	15.46	6	14	24
Non-delinquent	5,463	36.00	36	36	36
All	7,157	31.14	36	36	36

Feedback observed	M6	M12	M24	M36
Delinquent (%)	25.4	44.4	75.9	100.0
All labels (%)	6.0	10.5	18.0	100.0

Table 4 describes the continuous variables most closely tied to exposure, pricing, maturity, collateral, and baseline balances. The full representation also includes product, purpose, guarantee, and account type indicators. Missing values are imputed using historical-partition medians; continuous variables are clipped at the historical 1st and 99th percentiles and standardized with historical moments. The 77-month records are used to determine outcome times, but post-baseline balances are not supplied to the decision rule. Its temporal state instead consists of the elapsed feedback time of each accepted, unresolved case.

Table 4: Summary statistics for representative decision-time covariates in the Bank analysis sample (n = 7,157). Monetary amounts are in thousands of RMB. The experiment uses 37 covariates in total; the remaining variables are application and contract indicators. IQR denotes the 25th–75th percentiles.

Feature	Unit	Mean	SD	Median	IQR
Installment fee rate	%	11.29	5.60	11.00	5–15
Installment amount	RMB k	125.94	64.58	100.00	100–150
Number of installments	Count	29.81	14.46	24.00	24–36
Contract amount	RMB k	294.23	237.35	260.00	200–330
Loan term	Months	215.95	54.34	240.00	180–240
Collateral value	RMB k	500.17	920.13	421.99	333.73–532.40
Collateral floor area	m^2	174.79	6204.25	95.00	84–115

Feature	Unit	Mean	SD	Median	IQR
Disbursement amount	RMB k	294.23	237.35	260.00	200–330
Benchmark interest rate	%	4.89	0.11	4.90	4.90–4.90
Applied interest rate	%	4.63	0.87	4.90	3.68–5.39
Total credit facility	RMB k	211.20	194.02	178.69	115.85–247.93
Mortgage balance at baseline	RMB k	251.97	202.80	224.32	159.49–293.55
On-balance-sheet balance	RMB k	41.43	349.26	1.88	0.05–11.40
Off-balance-sheet balance	RMB k	39.09	251.05	2.00	0.05–11.10

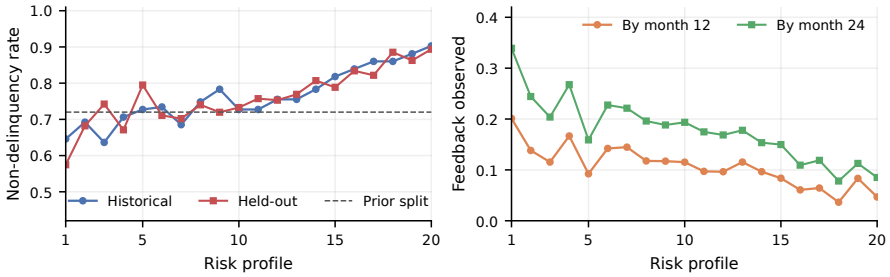


Figure 7: Risk and feedback-time heterogeneity across the 20 profiles used in the Bank replay.

Historical beliefs and profiles. We fit a regularized logistic model ($C = 0.01$, class-balanced observation weights) on the historical partition and use its scores only to group customers into 20 profiles. Score quantiles are estimated on the historical partition. Within profile j, the smoothed historical success rate is $\widehat{p}^j = (N_{j,1} + 1)/(N_j + 2)$ (N_j is the number of total historical outcomes and $N_{j,1}$ is the number of successful historical outcomes), and the policy is initialized with $(\alpha_0^j, \beta_0^j) = 4(\widehat{p}^j, 1 - \widehat{p}^j)$. No outcomes from the validation or test partitions enter these initial beliefs. Figure 7 shows that the score ordering remains informative out of sample while preserving meaningful heterogeneity in feedback arrival.

Policy specification and replay. Because the cost of exploratory approvals can differ across historical risk levels, we allow the uncertainty weight ρ^j in the CAIV component to vary across two prespecified risk strata. Profiles with $\widehat{p}^j < 0.72$ use $\rho^j = 0.8$, whereas the remaining profiles use $\rho^j = 1.0$. The threshold and the corresponding weights are selected using the validation partition and then fixed before test evaluation. All remaining SAIV parameters are likewise selected on the validation partition and held fixed throughout the test replay.

Within a seed, every method receives the same customer order, outcomes, and approval-cost sequence. The maturation kernel is estimated from 1,000 historical outcome times rather than from the frozen test partition. Offline outcomes score every policy, but only outcomes from that policy's accepted cases are scheduled for feedback and only matured outcomes update its posterior; rejected outcomes remain unavailable throughout the replay.

H Computation Infrastructure

All studies and experiments are run on Dell Precision 7920 Workstations with Intel(R) Xeon(R) Gold 6256 CPU at 3.60GHz and three sets of NVIDIA Quadro GV100 GPUs. All models are implemented in Python 3.8. The versions of the main packages of our code are NumPy: 1.19.2, Pandas: 1.1.3, Matplotlib: 3.3.2.