THE SURROGATE IS NOT THE REWARD: POST-SURROGATE PRIMARY-OUTCOME ACQUISITION IN CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study contextual bandits in which a surrogate is observed after the action but before the learner decides whether to acquire the primary outcome that defines action value and regret. The value of acquiring the primary outcome depends on both decision relevance—how much the current outcome matters for comparing policies—and the residual uncertainty after observing the surrogate. The Audited Surrogate Bandit (ASB) learns a contextual policy while allocating a budget of B primary-outcome acquisitions over T rounds. ASB sets a pre-surrogate acquisition level from current decision relevance and, after observing the surrogate, redistributes that level using an estimate of that residual uncertainty. For a finite class of N policies over K actions, ASB incurs $\widetilde{O}[\sqrt{KT\log N}\{1+\sqrt{T/B}\}]$ regret relative to the best policy in the class. In a two-action family where the surrogate does not reveal the better action, a learner that observes the surrogate before deciding whether to acquire can achieve bounded regret, whereas any learner that must decide before seeing the surrogate incurs $\Omega(T/B)$ worst-case regret under the same budget. Synthetic experiments show that both acquisition factors matter: ASB has lower regret than variants using only decision relevance or only residual uncertainty. On a KuaiRec benchmark of user–video interactions, the regret gap relative to relevance-only acquisition widens and then narrows as the budget grows.

1 INTRODUCTION

Post-action information can arrive too late to change the current action yet early enough to affect whether the primary outcome is acquired. We study contextual bandits in which a costly primary outcome Y_t defines action value and regret, but after selecting action A_t in context X_t, the learner first observes a surrogate S_t and only then decides whether to acquire Y_t. In mobile interventions, proximal activity can be observed automatically after intervention while broader health outcomes are harder to assess (Nahum-Shani et al., 2018; Klasnja et al., 2019). The round follows

$$X_t \longrightarrow A_t \longrightarrow S_t \longrightarrow \text{acquire } Y_t?.$$

By the time S_t arrives, the current action cannot be changed; the remaining decision is whether to measure Y_t. This leaves a distinct use for post-action information: it can help decide where to spend a limited measurement budget, while action value and regret remain tied to Y_t. The usefulness of measuring Y_t depends on both the current policy comparison and what remains uncertain after S_t. If competing policies choose the same action, observing Y_t does not distinguish them on that round; if they disagree but little residual variation remains after S_t, observing Y_t provides little information for distinguishing them. This points to two ingredients: *decision relevance*—how much the current outcome matters to unresolved policy comparisons—and the residual uncertainty estimated after the surrogate.

The Audited Surrogate Bandit (ASB) is an end-to-end contextual-bandit method that uses these two ingredients while learning a policy and deciding which primary outcomes to acquire. Before observing the surrogate, ASB sets an acquisition level from decision relevance; after observing it, ASB refines that level using estimated residual uncertainty. This timing also raises a broader question: what can be gained by allowing the acquisition decision to use the current surrogate? Our contributions are as follows:

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.

- We provide ASB, an end-to-end contextual-bandit method that learns a policy while allocating a hard budget of B primary-outcome acquisitions. In ASB, decision relevance sets the acquisition level before the surrogate is observed, and estimated residual uncertainty refines that level afterward. For a finite class of N policies over K actions, ASB acquires exactly B primary outcomes on every path and achieves $\widetilde{O}[\sqrt{KT\log N}\{1+\sqrt{T/B}\}]$ regret relative to the best policy in the class. A lower bound shows that the $\sqrt{T/B}$ dependence on the acquisition budget is unavoidable up to logarithmic factors on a subclass where an optimal policy is represented.
- We compare primary-outcome acquisition before and after observing the surrogate. For a fixed policy comparison under the same mean acquisition rate, allowing acquisition to use S_t can strictly lower the minimum variance of an unbiased estimate of the primary-outcome value difference. We show on a two-action family that this timing can change the attainable regret rate: post-surrogate acquisition can achieve bounded regret while any learner deciding before the surrogate incurs $\Omega(T/B)$ worst-case regret, even though S_t does not reveal the optimal action.
- We demonstrate in synthetic experiments that both parts of ASB's acquisition design matter for regret: removing either decision relevance or the post-surrogate residual-RMS refinement increases cumulative regret. These experiments also show that using the current surrogate for acquisition can reduce regret even when it does not reveal the optimal action. ASB has lower regret when more residual information lies where represented policies disagree rather than where they agree, even when the overall amount of residual information is fixed. We evaluate ASB against relevance-only acquisition on KuaiRec recommendation data across several acquisition budgets: the mean-regret gap first widens and then narrows as the budget grows.

Related work on acquisition design includes prediction-guided experimentation and selective ground-truth acquisition for inference or best-arm identification (Ao et al., 2024; 2026). Costly-feedback bandits study whether to observe reward (Efroni et al., 2021; Tucker et al., 2023), while intermediate- and post-action-feedback methods use information revealed after acting (Esposito et al., 2023; Wang et al., 2023). In ASB, the surrogate is observed after the action and used to decide whether to acquire the same round's primary outcome, while regret and subsequent policy learning remain tied to that primary outcome. Appendix H discusses broader connections to surrogate-based inference and limited-follow-up designs, active inference and feature acquisition, and surrogate-guided sequential decision methods.

2 Post-Surrogate Primary-Outcome Acquisition

For $t = 1, \dots, T$, the learner observes a context $X_t \in \mathcal{X}$ and selects an action $A_t \in \mathcal{A}$, where $K = |\mathcal{A}| < \infty$. The resulting surrogate S_t is then observed for free. The pre-surrogate and post-surrogate states are $U_t = (X_t, A_t)$ and $Z_t = (U_t, S_t)$, respectively. After observing S_t, the learner sets an acquisition probability $\rho_t(Z_t) \in [0,1]$, draws $Q_t \sim \mathrm{Bern}\{\rho_t(Z_t)\}$, and observes the primary outcome $Y_t \in [0,1]$ iff $Q_t = 1$. The acquisition decision only determines whether Y_t is observed; it does not change future contexts or outcomes. With $\mathcal{F}_{t-1}$ the history before round t, the timing is

$$(\mathcal{F}_{t-1}, X_t) \to A_t \to S_t \to \rho_t(Z_t) \to Q_t \to Q_t Y_t, \qquad Q_t \perp Y_t \mid (\mathcal{F}_{t-1}, Z_t).$$

The conditional mean of the primary outcome is denoted by $r^\star(x,a)$, so $\mathbb{E}(Y_t \mid \mathcal{F}_{t-1}, X_t, A_t) = r^\star(X_t, A_t)$. Cumulative regret is $R_T := \sum_{t=1}^{T}[\max_{a\in\mathcal{A}} r^\star(X_t,a) - r^\star(X_t,A_t)]$. The learner minimizes R_T subject to the hard budget $\sum_{t=1}^{T} Q_t \le B$ on every path, where $B \in \{1,\dots,T\}$ is the primary-outcome acquisition budget.

This timing raises two questions: how to learn a contextual policy while allocating the primary-outcome budget, and what changes when the learner decides whether to acquire the primary outcome before or after observing the current surrogate.

3 Audited Surrogate Bandit

We present ASB for the setting in Section 2. The method learns a contextual policy while deciding which primary outcomes to acquire, using a finite class Π of N deterministic policies. We first follow one online round from action selection through primary-outcome acquisition and policy updating, then examine why ASB uses decision relevance before the surrogate and residual uncertainty after it.

3.1 ONE ROUND OF ASB

Action selection and decision relevance. At round t, ASB begins with its current weights $w_t(\pi)$ for policies $\pi \in \Pi$. After observing X_t, it normalizes these weights to form a policy mixture $\mu_t(\pi) \propto w_t(\pi)$. This mixture induces the action distribution $\bar{p}_t(a) = \sum_{\pi \in \Pi} \mu_t(\pi) \mathbf{1}\{\pi(X_t) = a\}$ over $a \in \mathcal{A}$. Using exploration rate ε_t, ASB forms $p_t(a) = (1 - \varepsilon_t)\bar{p}_t(a) + \varepsilon_t u_t(a)$, where u_t is uniform on the support of $\bar{p}_t$, and draws $A_t \sim p_t$.

For any $a \in \mathcal{A}$ with $p_t(a) > 0$, define the *policy-comparison coefficient* $c_{t,\pi}(a)$ and $\ell_t(a)$ by

$$c_{t,\pi}(a) = \{\mathbf{1}\{\pi(X_t) = a\} - \bar{p}_t(a)\}/p_t(a), \qquad \ell_t(a) = \max_{\pi' \in \Pi} |c_{t,\pi'}(a)|. \tag{1}$$

The coefficient $c_{t,\pi}(a)$ measures how much the primary-outcome mean at action a contributes to the comparison between policy π and the current policy mixture μ_t, accounting for the probability $p_t(a)$ of selecting a. Its maximum magnitude across policies, $\ell_t(a)$, is the *decision-relevance score*.

Pre-surrogate acquisition level and post-surrogate refinement. Before observing S_t, ASB uses $\ell_t(A_t)$ to set a pre-surrogate acquisition level ρ_t^0. After observing S_t, it evaluates the fitted bridge $h_t(Z_t) \in [0, 1]$ for Y_t and a learned root-mean-square (RMS) residual score $\widehat{\tau}_t(Z_t)$ for $Y_t - h_t(Z_t)$. ASB combines decision relevance with residual RMS through the following basic acquisition form:

$$\rho_t^0 \propto \ell_t(A_t), \qquad \rho_t^1(Z_t) = \rho_t^0 \frac{\widehat{\tau}_t(Z_t)}{\widehat{\mathbb{E}}_t\{\widehat{\tau}_t(U_t, S) \mid U_t\}}, \qquad \rho_t(Z_t) = (1 - \gamma_t)\rho_t^0 + \gamma_t \rho_t^1(Z_t). \tag{2}$$

Here $\widehat{\mathbb{E}}_t(\cdot \mid U_t)$ is a fitted conditional expectation over S. The refinement ρ_t^1 redistributes the pre-surrogate level across surrogate values according to the learned residual RMS score. The resulting acquisition probability ρ_t combines the pre-surrogate level with this refinement, with $\gamma_t \in [0, 1]$, fixed before S_t, controlling how much of the redistribution is used. Appendix E gives one concrete online fitting specification for h_t, $\widehat{\tau}_t$, and $\widehat{\mathbb{E}}_t$.

Algorithm 2 in Appendix B turns the acquisition structure in Equation (2) into the hard-budget rule. It uses the remaining budget and past decision-relevance scores to set the scale of ρ_t^0. When acquisition is not fixed by the budget rule, the post-surrogate refinement is constrained to keep its fitted conditional mean at ρ_t^0 and, whenever $\ell_t(A_t) > 0$, to stay above a positive acquisition floor set before S_t without using the learned residual score. Its remaining-count boundary rules ensure $\sum_{t=1}^{T} Q_t = B$ on every path.

Augmented feedback and policy update. ASB then draws $Q_t \sim \mathrm{Bern}\{\rho_t(Z_t)\}$ and observes Y_t iff $Q_t = 1$. For $\rho_t(Z_t) > 0$, it forms the augmented feedback

$$\Phi_t = h_t(Z_t) + \frac{Q_t}{\rho_t(Z_t)}\{Y_t - h_t(Z_t)\}. \tag{3}$$

Using $Q_t \perp Y_t \mid (\mathcal{F}_{t-1}, Z_t)$, the correction in (3) preserves the primary-outcome conditional mean:

$$\mathbb{E}(\Phi_t \mid \mathcal{F}_{t-1}, Z_t) = \mathbb{E}(Y_t \mid \mathcal{F}_{t-1}, Z_t), \qquad \mathbb{E}(\Phi_t \mid \mathcal{F}_{t-1}, U_t) = r^\star(X_t, A_t). \tag{4}$$

For each policy π, ASB forms the *policy-comparison evidence* $G_{t,\pi} := c_{t,\pi}(A_t)\{\Phi_t - 1/2\}$, with $1/2$ centering the $[0, 1]$ outcome scale. If $\rho_t(Z_t) > 0$ whenever $\ell_t(A_t) > 0$, Equations (1) and (4) imply

$$\mathbb{E}(G_{t,\pi} \mid \mathcal{F}_{t-1}, X_t) = r^\star\{X_t, \pi(X_t)\} - \sum_a \bar{p}_t(a) r^\star(X_t, a). \tag{5}$$

Then $G_{t,\pi}$ is conditionally unbiased for the primary-outcome mean difference between policy π and the policy mixture μ_t. ASB then updates the policy weights using the second-order exponential-weights rule with learning rate η_t:

$$w_{t+1}(\pi) = w_t(\pi) \exp\{\eta_t G_{t,\pi} + \eta_t^2 G_{t,\pi}^2\}. \tag{6}$$

ASB also updates h_{t+1}, $\widehat{\tau}_{t+1}$, and $\widehat{\mathbb{E}}_{t+1}$, and the next round starts from the new policy weights w_{t+1}.

Algorithm 1 Audited Surrogate Bandit (ASB)

Require: horizon T, budget B, policy class Π; exploration/refinement schedules $(\varepsilon_t, \gamma_t)_{t=1}^T$; policy learning-rate rule for η_t; fitting procedures for h_t, $\widehat{\tau}_t$, and $\widehat{\mathbb{E}}_t$.

1: **Initialize:** set $w_1(\pi) = 1$ for every $\pi \in \Pi$.
2: **for** $t = 1, \ldots, T$ **do**
3: **Policy mixture:** observe X_t; set $\mu_t(\pi) = w_t(\pi)/\sum_{\pi'} w_t(\pi')$; $\bar{p}_t(a) = \sum_{\pi \in \Pi} \mu_t(\pi) \mathbf{1}\{\pi(X_t) = a\}$.
4: **Act:** set $p_t(a) = (1 - \varepsilon_t)\bar{p}_t(a) + \varepsilon_t u_t(a)$; draw $A_t \sim p_t$.
5: **Decision relevance:** compute $c_{t,\pi}(a)$ and $\ell_t(a)$ by (1).
6: **Acquisition probability:** use Algorithm 2 in Appendix B to set ρ_t^0 before observing S_t and $\rho_t^1(Z_t)$ after observing S_t. Set $\rho_t(Z_t) = (1 - \gamma_t)\rho_t^0 + \gamma_t \rho_t^1(Z_t)$.
7: **Acquire primary outcome:** draw $Q_t \sim \mathrm{Bern}\{\rho_t(Z_t)\}$; observe Y_t iff $Q_t = 1$.
8: **Augmented feedback:** if $\rho_t(Z_t) > 0$, set $\Phi_t = h_t(Z_t) + \{Q_t/\rho_t(Z_t)\}\{Y_t - h_t(Z_t)\}$ and $G_{t,\pi} = c_{t,\pi}(A_t)\{\Phi_t - 1/2\}$; otherwise set $G_{t,\pi} = 0$.
9: **Update:** set $w_{t+1}(\pi)$ by (6); update h_{t+1}, $\widehat{\tau}_{t+1}$, and $\widehat{\mathbb{E}}_{t+1}$.
10: **end for**

3.2 Acquisition structure: decision relevance and residual RMS

To explain why Equation (2) uses decision relevance before the surrogate and residual RMS afterward, consider how the acquisition probability ρ_t affects the one-round mean-squared error of the policy-comparison evidence $G_{t,\pi}$. Under the positive-support condition in Equation (5), this mean-squared error is its conditional variance, which decomposes as

$$\mathrm{Var}(G_{t,\pi} \mid \mathcal{F}_{t-1}, X_t) = \mathbb{E}\left[\left.\frac{c_{t,\pi}^2(A_t)\tau_t^2(Z_t)}{\rho_t(Z_t)}\right| \mathcal{F}_{t-1}, X_t\right] + \text{a term independent of } \rho_t. \tag{7}$$

Here $\tau_t(U_t, s)$ is the *true residual RMS* around the bridge h_t:

$$\tau_t^2(U_t, s) := \mathbb{E}\left[\{Y_t - h_t(U_t, s)\}^2 \mid \mathcal{F}_{t-1}, U_t, S_t = s\right].$$

Equation (7) shows why ASB uses the decision-relevance score $\ell_t(A_t) = \max_{\pi'} |c_{t,\pi'}(A_t)|$ for the pre-surrogate acquisition level and residual RMS $\tau_t(U_t, S_t)$ for post-surrogate redistribution.

Pre-surrogate acquisition level: decision relevance. To characterize the action dependence of ρ_t^0, consider a candidate pre-surrogate rule ϱ, with level $\varrho(a)$ when $A_t = a$. Because $A_t \sim p_t$, Equation (7) averages these action-specific terms with weights $p_t(a)$. Since $\tau_t^2 \le 1$ and $\ell_t(a)$ bounds $|c_{t,\pi}(a)|$ over $\pi \in \Pi$, the acquisition-dependent term is at most $\sum_{a:\, \ell_t(a) > 0} p_t(a)\ell_t^2(a)/\varrho(a)$ for every $\pi \in \Pi$. At a fixed average acquisition level, this calculation gives the proportional shape $\varrho(a) \propto \ell_t(a)$, motivating $\rho_t^0 \propto \ell_t(A_t)$ in Equation (2).

Post-surrogate redistribution: residual RMS. The pre-surrogate argument above determines how ρ_t^0 depends on the selected action. The post-surrogate question is how to redistribute that level across surrogate values for a fixed U_t. The residual RMS $\tau_t(U_t, S)$ may vary with S but is unknown, so ASB uses the learned residual RMS score $\widehat{\tau}_t(U_t, S)$. Among candidate redistribution rules $r(\cdot)$ with fitted conditional mean ρ_t^0, consider

$$\min_{r(\cdot) > 0} \widehat{\mathbb{E}}_t\left[\left.\frac{\widehat{\tau}_t^2(U_t, S)}{r(S)}\right| U_t\right] \quad \text{subject to} \quad \widehat{\mathbb{E}}_t\{r(S) \mid U_t\} = \rho_t^0. \tag{8}$$

When $\widehat{\tau}_t(U_t, S) > 0$ almost surely under the fitted law, its proportional solution gives the basic refinement in Equation (2); Appendix B additionally enforces the probability bounds and support floor. Surrogate values with larger learned residual RMS scores receive more acquisition probability. The hard-budget refinement keeps its fitted conditional mean at ρ_t^0, so the final acquisition probability preserves that level as well:

$$\widehat{\mathbb{E}}_t\{\rho_t(U_t, S) \mid U_t\} = (1 - \gamma_t)\rho_t^0 + \gamma_t \widehat{\mathbb{E}}_t\{\rho_t^1(U_t, S) \mid U_t\} = \rho_t^0.$$

Under the same support condition, $\rho_t(Z_t)$ may depend on S_t, yet Equations (4) and (5) show that the conditional mean of $G_{t,\pi}$ remains unchanged—the primary-outcome mean difference between policy π and the current policy mixture μ_t. Equation (7), by contrast, shows that the conditional variance of $G_{t,\pi}$ can change with ρ_t. Acquisition rules can therefore preserve that same primary-outcome mean difference while differing in whether ρ_t may depend on S_t.

4 PRIMARY-OUTCOME ACQUISITION BEFORE AND AFTER THE SURROGATE: ALLOCATION AND REGRET

We now isolate the effect of acquisition timing by comparing rules that choose ρ_t before versus after observing S_t. We first ask when allowing ρ_t to depend on S_t can reduce the minimum variance for a fixed policy comparison, and then whether the same timing difference can change the regret rate.

Throughout this section, we use PRE for acquisition rules whose ρ_t is fixed before S_t is observed, and POST for rules whose ρ_t may depend on S_t; every PRE rule is therefore also a POST rule.

4.1 ACQUISITION TIMING FOR A POLICY COMPARISON

Decision relevance describes how much a primary outcome helps distinguish competing policies. ASB quantifies it through comparisons with its current policy mixture; here we fix one policy comparison and compare PRE and POST acquisition rules.

Acquisition objective. Fix two deterministic policies π_1, π_2 and a common joint law for (X, A, S, Y). Under this law, $A \sim p(\cdot \mid X)$, where p is the action-selection distribution and $p(a \mid X) > 0$ for every action chosen by either policy. Write $U = (X, A)$ for the pre-surrogate state. For this fixed comparison, $|g(U)|$ captures decision relevance through

$$g(U) = \frac{\mathbf{1}\{\pi_1(X) = A\} - \mathbf{1}\{\pi_2(X) = A\}}{p(A \mid X)}.$$

Under an acquisition allocation ρ and a fixed bridge $h(U, S)$, the augmented feedback from Equation (3) is $\Phi = h(U, S) + \{Q/\rho(U, S)\}\{Y - h(U, S)\}$ whenever $\rho(U, S) > 0$. Combined with $g(U)$, Equation (4) gives

$$\mathbb{E}[g(U)\{\Phi - 1/2\}] = \mathbb{E}[r^\star\{X, \pi_1(X)\} - r^\star\{X, \pi_2(X)\}].$$

Thus $g(U)\{\Phi - 1/2\}$ is unbiased for the same primary-outcome value difference. Acquisition timing can therefore change its variance without changing its target.

To characterize this variance, fix the bridge h and define the residual RMS $\tau_h(U, S)$ by $\tau_h^2(U, S) := \mathbb{E}[\{Y - h(U, S)\}^2 \mid U, S]$. The part of the variance controlled by the allocation ρ is

$$\mathcal{V}(\rho) := \mathbb{E}\left[\frac{g^2(U)\tau_h^2(U, S)}{\rho(U, S)}\right]. \tag{9}$$

In $\mathcal{V}(\rho)$, decision relevance and residual RMS appear as the product $|g(U)|\tau_h(U, S)$. Here $|g(U)|$ is determined by U, whereas $\tau_h(U, S)$ may depend on S. The contribution vanishes whenever either is zero; finite $\mathcal{V}(\rho)$ requires $\rho(U, S) > 0$ wherever both are positive.

PRE and POST therefore minimize $\mathcal{V}(\rho)$ under different timing restrictions: PRE restricts ρ to $\rho(U)$, whereas POST allows $\rho(U, S)$. The theorem below gives the two minimizers and identifies when allowing ρ to depend on S strictly lowers the minimum variance.

Theorem 1 (PRE/POST variance characterization for a fixed policy comparison)**.** *For $\bar{\rho} \in (0, 1]$, consider minimizing $\mathcal{V}(\rho)$ subject to $\mathbb{E}\rho \le \bar{\rho}$. If $\Pr\{|g(U)|\tau_h(U, S) > 0\} > 0$, the PRE and POST minimizers, denoted $\rho^\star_{\mathrm{pre}}$ and $\rho^\star_{\mathrm{post}}$, are*

$$\begin{aligned} \rho^\star_{\mathrm{pre}}(U) &= \bar{\rho}\,\frac{|g(U)|\,\{\mathbb{E}[\tau_h^2(U, S) \mid U]\}^{1/2}}{\mathbb{E}\big[|g(U)|\,\{\mathbb{E}[\tau_h^2(U, S) \mid U]\}^{1/2}\big]}, \\ \rho^\star_{\mathrm{post}}(U, S) &= \bar{\rho}\,\frac{|g(U)|\,\tau_h(U, S)}{\mathbb{E}[|g(U)|\,\tau_h(U, S)]}, \end{aligned} \tag{10}$$

provided $\rho^\star_{\mathrm{pre}}(U) \le 1$ and $\rho^\star_{\mathrm{post}}(U, S) \le 1$ almost surely. Moreover, $\mathcal{V}(\rho^\star_{\mathrm{post}}) \le \mathcal{V}(\rho^\star_{\mathrm{pre}})$, with strict inequality exactly when $\Pr(|g(U)| > 0,\ \mathrm{Var}\{\tau_h(U, S) \mid U\} > 0) > 0$.

The decision-relevance factor $|g(U)|$ is the same in $\rho^\star_{\mathrm{pre}}$ and $\rho^\star_{\mathrm{post}}$. $\rho^\star_{\mathrm{pre}}(U)$ uses the U-conditional residual RMS $\{\mathbb{E}[\tau_h^2(U, S) \mid U]\}^{1/2}$, whereas $\rho^\star_{\mathrm{post}}(U, S)$ uses $\tau_h(U, S)$ after observing S. Hence POST strictly lowers the minimum variance when, at pre-surrogate states U that matter to the policy comparison, observing S adds information about residual uncertainty beyond that in U alone.

4.2 Regret-Rate Separation

For a fixed policy comparison, the preceding result shows that using the current surrogate can strictly reduce the minimum variance at the same average acquisition rate. We next ask whether this timing difference can also separate regret rates when the surrogate does not identify the optimal action.

For any fixed $q \in (0,1)$, consider two bandit instances indexed by $\vartheta \in \{-1,+1\}$. Both have a constant context $X_t \equiv x_0$, binary action set $\mathcal{A} = \{-1,+1\}$, and the same i.i.d. binary surrogate law $S_t \sim \mathrm{Bern}(q)$, with $S_t \perp A_t \mid \mathcal{F}_{t-1}$. Define the primary outcome by

$$Y_t = 1/2 \quad \text{if } S_t = 0, \qquad Y_t = \mathbf{1}\{A_t = \vartheta\} \quad \text{if } S_t = 1.$$

In instance ϑ, the primary-outcome mean is $\mathbb{E}_\vartheta(Y_t \mid A_t = a) = 1/2 + (q/2)\vartheta a$. Thus $a = \vartheta$ is optimal, while choosing $a = -\vartheta$ incurs regret q. The surrogate does not reveal the optimal action, but it does reveal whether the current primary outcome is informative: if $S_t = 1$, (A_t, Y_t) identifies ϑ; if $S_t = 0$, $Y_t = 1/2$ under both instances. For the fixed bridge $h \equiv 1/2$, the residual RMS is 0 at $S_t = 0$ and $1/2$ at $S_t = 1$.

Let $\mathfrak{L}_{\mathrm{pre}}(B)$ denote PRE strategies satisfying $\mathbb{E}_\vartheta \sum_{t=1}^T Q_t \le B$ for each ϑ. Every PRE strategy under the pathwise hard budget of Section 2 belongs to this class, so the lower bound also applies to the original hard-budget problem. Write $(x)_+ = \max\{x, 0\}$.

Theorem 2 (Regret separation from acquisition timing)**.** *For any $q \in (0,1)$ and $B \in \{1,\dots,T\}$, every $\mathcal{L} \in \mathfrak{L}_{\mathrm{pre}}(B)$ satisfies*

$$\max_{\vartheta \in \{-1,+1\}} \mathbb{E}_\vartheta R_T \ge (qT/2)(1 - qB)_+. \tag{11}$$

In contrast, there exists a POST strategy $\mathcal{L}_{\mathrm{post}}$ such that $\sum_{t=1}^T Q_t = B$ on every path and

$$\max_{\vartheta \in \{-1,+1\}} \mathbb{E}_\vartheta R_T \le [1 - (1-q)^T]/2 \le 1/2. \tag{12}$$

PRE obtains informative evidence only when an acquisition coincides with $S_t = 1$, so the expected number of informative acquisitions is at most qB. With $q = 1/(2B)$, the probability of no informative acquisition is at least $1/2$. POST can instead acquire the first $S_t = 1$ round, which identifies ϑ. Thus

$$\begin{aligned} &\max_\vartheta \mathbb{E}_\vartheta R_T \le 1/2 && \text{for } \mathcal{L}_{\mathrm{post}}, \\ &\max_\vartheta \mathbb{E}_\vartheta R_T \ge T/(8B) && \text{for every } \mathcal{L} \in \mathfrak{L}_{\mathrm{pre}}(B). \end{aligned} \tag{13}$$

For $B = o(T)$, every $\mathcal{L} \in \mathfrak{L}_{\mathrm{pre}}(B)$ has $\Omega(T/B)$ worst-case regret, while $\mathcal{L}_{\mathrm{post}}$ has bounded regret. The surrogate reveals when the current primary outcome is informative without revealing the optimal action. Appendix A.4 gives the corresponding separation with noisy informative outcomes.

5 Regret under a Primary-Outcome Acquisition Budget

We first establish ASB's regret guarantee under the pathwise hard budget, then give a lower bound under a weaker expected acquisition budget to isolate the budget dependence that remains after this relaxation.

5.1 ASB Regret under the Hard Budget

ASB competes with the best policy in Π, which need not select the best action on every round. To separate this gap from online learning regret, define $\mathrm{App}_T(\Pi) := \min_{\pi \in \Pi} \sum_{t=1}^T [\max_{a \in \mathcal{A}} r^\star(X_t, a) - r^\star\{X_t, \pi(X_t)\}]$ as the cumulative gap between the best action on each round and the best policy in Π. Then $R_T - \mathrm{App}_T(\Pi)$ is regret relative to the best policy in the class.

Theorem 3 (Best-in-class regret)**.** *Fix $\delta \in (0,1)$. In the setting of Section 2, let $|\Pi| = N$ and $B \in \{1,\dots,T\}$. Run* ASB *(Algorithm 1) using the hard-budget rule specified in Appendix B, the fixed learning rate $\eta_t \equiv \eta_\star$ from Appendix C.4, and $\varepsilon_t \equiv \min\{1/2, \sqrt{K \log(12NT/\delta)/T}\}$. Then, with probability at least $1 - \delta$,*

$$R_T - \mathrm{App}_T(\Pi) \le \widetilde{O}\left[\min\left\{T, \sqrt{KT \log(N/\delta)}\left(1 + \sqrt{T/B}\right)\right\}\right]. \tag{14}$$

Up to logarithmic factors, Equation (14) is the usual finite-policy contextual-bandit scale $\sqrt{KT \log N}$ multiplied by $1 + \sqrt{T/B}$. At $B = T$, this is the usual scale with primary-outcome feedback every round; the budget factor grows as B decreases.

The guarantee is end-to-end and does not assume correctness of the fitted bridge, residual score, or surrogate law. Appendix C.6 separately gives a sufficient condition under which the learned post-surrogate refinement reduces the acquisition-dependent contribution to the regret bound.

5.2 Primary-Outcome Scarcity Lower Bound

We next lower-bound regret after relaxing the pathwise hard budget to $\sup_\nu \mathbb{E}_\nu \sum_{t=1}^T Q_t \le B$, where ν indexes environments in the lower-bound family. Every pathwise hard-budget learner satisfies this constraint, so the same lower bound applies to the original pathwise setting.

Theorem 4 (Regret lower bound under a primary-outcome budget)**.** *There is a universal constant $c_{\mathrm{lb}} > 0$ such that, for every horizon T, $N \ge K \ge 2$, and $B \in \{1, \dots, T\}$, there is a contextual-bandit family in the setting of Section 2, with K actions and a policy class Π of size N satisfying $\mathrm{App}_T(\Pi) = 0$, for which every learner $\mathcal{L}$ with $\sup_\nu \mathbb{E}_\nu \sum_{t=1}^T Q_t \le B$ satisfies*

$$\sup_\nu \mathbb{E}_\nu R_T \ge c_{\mathrm{lb}} \min\left\{T, \sqrt{KT \log N / \log K}\sqrt{T/B}\right\}.$$

Unlike the PRE-specific separation in Section 4.2, Theorem 4 applies to every learner, showing that the $\sqrt{T/B}$ dependence remains even without the PRE restriction.

6 Empirical Tests of Acquisition Design

Synthetic experiments isolate the acquisition mechanisms while holding the primary-outcome decision problem fixed. KuaiRec then evaluates the learned post-surrogate refinement in a data-derived contextual benchmark. Within each matched comparison, the policy learner and primary-outcome budget are held fixed. The KuaiRec comparison repeats this design across budgets.

6.1 Synthetic tests of acquisition structure

The synthetic study asks three questions within one family: how regret changes when either acquisition component—decision relevance or residual RMS—is removed; whether acquisition timing relative to the current surrogate changes regret; and whether relocating a fixed average amount of residual information between policy-agreement and policy-disagreement contexts changes regret.

Contexts are divided into a region where the represented policies agree and a region where they disagree. Across all variants, the context distribution, policy class, and $P(Y \mid X, A)$ are fixed, with $\mathbb{E}(Y \mid X, A, S) = \mathbb{E}(Y \mid X, A)$. The surrogate can therefore reveal differences in residual magnitude without revealing which action is better.

All tests use $K = 8$ actions, $N = 256$ policies, horizon $T = 10^5$, and an exact primary-outcome budget of $B = 3{,}000$. Figure 1 summarizes the tests; Appendix G.2 gives the data-generating process.

Acquisition components. The left panel compares ASB with two one-component ablations and two simple references. *Relevance only* keeps decision relevance but removes redistribution after observing S_t; *Residual only* uses the learned residual RMS score without decision relevance; *Uniform* uses neither score; and *Fixed-N* is a budget-matched version of the prefix strategy in Tucker et al. (2023), acquiring the first B primary outcomes. In this family, ASB has substantially lower cumulative regret than each of the four alternatives. The comparison shows why both are useful: decision relevance identifies rounds that distinguish competing policies, while the residual RMS score prioritizes surrogate states by the estimated uncertainty remaining after S_t.

Acquisition timing. The middle panel tests the information-order distinction in Section 4 with PRE and POST oracles using exact residual-RMS quantities around the common fitted bridge. PRE and POST coincide when the surrogate carries no residual-magnitude information; as λ increases, POST attains lower regret. Since $P(Y \mid X, A)$ is unchanged and S_t does not reveal the better action, the gap comes from residual-magnitude information being available before acquisition.

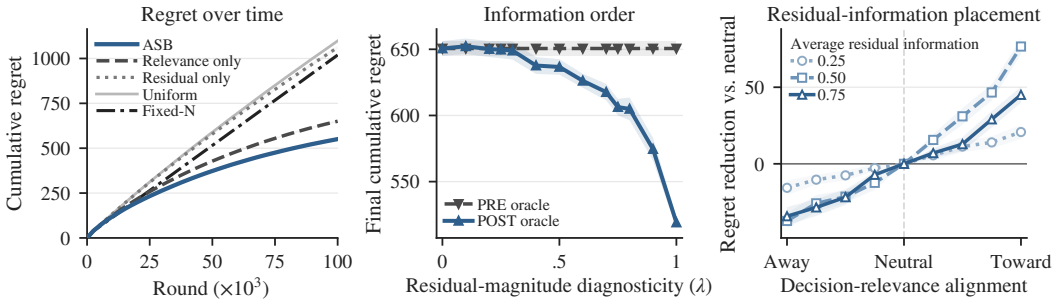


Figure 1: **Synthetic tests of acquisition structure.** Left: cumulative regret for ASB, Relevance only, Residual only, Uniform, and Fixed-N under the same exact primary-outcome budget. Middle: final cumulative regret for matched PRE and POST oracles as λ increases the surrogate's information about residual magnitude, from none at $\lambda = 0$ to exact identification of the high-residual state at $\lambda = 1$. Right: at each fixed average residual-information level, information is shifted from policy-agreement contexts (Away), through equal information in the two regions (Neutral), to policy-disagreement contexts (Toward); positive values mean lower regret than neutral placement. Bands show mean ± 1 SE over 64 matched environments.

Information placement. The right panel moves residual information between the policy-agreement and policy-disagreement regions while holding its average amount fixed within each curve. At all three average information levels, placing more of that information in the disagreement region lowers regret. Thus the same average amount of residual information is more useful when it lies where represented policies disagree. The middle of the three average-information levels allows the largest contrast in residual information between the agreement and disagreement regions and also shows the largest regret change.

6.2 KuaiRec acquisition benchmark

Using KuaiRec's nearly complete matrix subset (Gao et al., 2022), we construct a contextual acquisition benchmark with $K = 5$ actions. Its surrogate–primary relationship comes from observed user–video outcomes. On each round the learner selects one of five candidate videos, observes $S7 = \mathbf{1}\{\text{play duration} > 7 \text{ seconds}\}$ as a free surrogate, and then decides whether to acquire $Y = \min\{\text{watch ratio}, 1\}$ as the primary outcome. Both regret and terminal policy value are evaluated using the watch-ratio primary outcome.

The benchmark has $N = 9$ represented policies and horizon $T = 22{,}417$. We evaluate primary-outcome budgets of $2, 3, 4, 6, 8, 12, 16$, and 24%. At each budget, ASB is compared with the matched *Relevance only* reference, which removes only the learned redistribution after observing S_t. The comparison therefore isolates the contribution of the learned post-surrogate refinement as the primary-outcome budget varies.

The bridge, residual score, and conditional surrogate law are fitted on a disjoint KuaiRec subset and frozen before evaluation; only acquired benchmark primary outcomes enter the online learner. Appendix G.3 gives the full benchmark construction and evaluation protocol.

The clearest separation occurs at 4–8%, where both mean contrasts favor ASB; the differences are small at the lowest budget and again at the upper end of the evaluated range. Cumulative regret summarizes the learning trajectory, whereas terminal policy value evaluates the final mixture. The regret contrast remains positive at budgets where the terminal-value contrast is near zero, so learning-path gains need not persist in terminal value.

At the low end, full-acquisition initialization consumes most of the budget, leaving few later rounds on which post-surrogate refinement can change what is acquired. Toward the high end, both methods acquire many more primary outcomes and the contrasts narrow. Between these limits, enough post-initialization rounds remain for refinement to change which outcomes are acquired, while primary feedback remains limited enough that those choices affect learning.

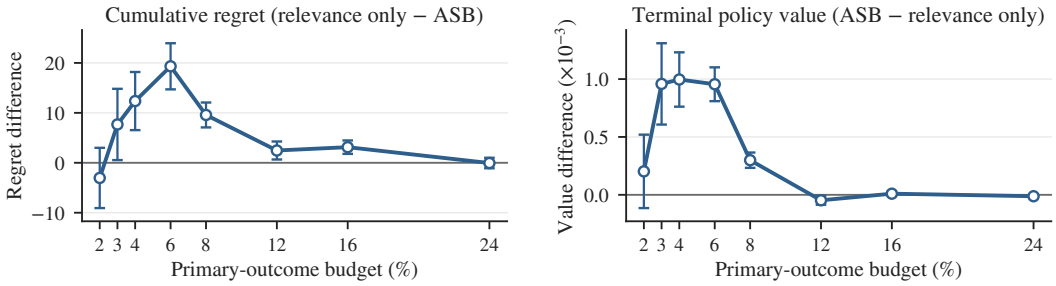


Figure 2: **Post-surrogate refinement on KuaiRec.** Left: paired cumulative-regret difference, Relevance only minus ASB. Right: paired exact terminal policy-value difference, ASB minus Relevance only, evaluated after the online run using the complete benchmark table. Positive values favor ASB in both panels. Error bars show mean ± 1 SE over 32 paired order/action randomizations and 8 acquisition randomizations per order/action stream.

7 Discussion

In this setting, the ordering matters: the surrogate arrives after the action but before the primary outcome is measured. It cannot revise A_t, but it can still affect which primary outcomes are observed and used in later policy updates. This leaves the primary outcome in its original role—defining action value and regret—while giving the surrogate a separate role in measurement. That acquisition question is concrete: would measuring the current outcome distinguish competing policies on this round, and how much uncertainty remains after S_t?

Once primary measurement is selective, three separate questions arise. Selective measurement can bias an uncorrected update; the augmented feedback preserves the primary target without requiring a correct bridge. Once that target is protected, the timing question remains: seeing the current surrogate before measurement can leave the minimum comparison variance unchanged in some settings and change the attainable regret rate in others, depending on whether S_t distinguishes residual uncertainty on rounds that matter to the policy comparison. A separate limitation comes from primary-outcome scarcity itself. The lower bound shows a worst-case $\sqrt{T/B}$ penalty, up to logarithmic factors, even after the pathwise hard budget is relaxed to an expected acquisition budget.

The next difficulty is learning the acquisition rule online: the learner does not know in advance which rounds are worth measuring. The selectively acquired primary outcomes feed both the policy updates and the bridge and residual fits, so evidence collected for action learning also changes which later outcomes are acquired. ASB's worst-case guarantee does not require the fitted bridge, residual score, or surrogate law to be correct; errors in them can hurt how efficiently the budget is used without changing the primary target preserved by the augmented update. The synthetic experiments show why both decision relevance and post-surrogate residual information matter: in this family, removing either one increases regret, and the same average residual information is more useful when it is concentrated where represented policies disagree. KuaiRec then shows how a learned refinement behaves across finite budgets: the separation from relevance-only acquisition is clearest at 4–8%; at the low end, initialization leaves little room for later refinement, while at the high end both methods observe many more primary outcomes and the contrast narrows. Cumulative-regret gains can also persist when terminal-value differences are near zero. Thus, an opportunity to use post-surrogate information is not the same as being able to learn and exploit that opportunity from limited primary feedback.

The present formulation treats measurement as passive: acquiring the current primary outcome reveals it without changing future contexts or outcomes. If measurement itself changes future state, acquisition becomes part of a dynamic decision problem. Within the setting studied here, the information order has a direct consequence: by the time S_t arrives, the action is fixed, but the decision to measure Y_t is still open. What matters about the post-action signal is therefore not only what it predicts, but what it can still change. The surrogate can change which primary outcomes are measured for future learning, while Y_t continues to define action value and regret.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing for clarity and readability, related-work search and discovery, implementation of Python code for experiments, and checking and refinement of proof sketches and mathematical arguments. All AI-assisted material was reviewed by the authors, who take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

Proofs and supporting theoretical details are provided in the appendices. Appendix E gives one fitting specification, and Appendix G records the reported experimental constructions, fixed settings, comparators, randomization, and evaluation protocol. The supplementary code contains the executable refit schedules and low-level runtime details needed to reproduce the reported figures.

REFERENCES

Ruicheng Ao, Hongyu Chen, and David Simchi-Levi. Prediction-guided active experiments. *arXiv preprint arXiv:2411.12036*, 2024. URL https://arxiv.org/abs/2411.12036.

Ruicheng Ao, Hongyu Chen, Siyang Gao, Hanwei Li, and David Simchi-Levi. Best arm identification with LLM judges and limited human audits. *arXiv preprint arXiv:2601.21471*, 2026. URL https://arxiv.org/abs/2601.21471.

Kianoosh Ashouritaklimi, Valentin Kilian, Daolang Huang, Tom Rainforth, and François Caron. Prediction-powered active testing. *arXiv preprint arXiv:2607.08347*, 2026. URL https://arxiv.org/abs/2607.08347.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 19–26. PMLR, 2011. URL https://proceedings.mlr.press/v15/beygelzimer11a.html.

Nicholas Brawand, Nima Leclerc, Anhthy Ngo, Matthew Peterson, Sriram Vishwanath, Laith Alhussein, and Ben Wellner. Active multiple-prediction-powered inference. *arXiv preprint arXiv:2605.08429*, 2026. URL https://arxiv.org/abs/2605.08429.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, ICML '11, pp. 1097–1104, 2011. URL https://arxiv.org/abs/1103.4601.

Yonathan Efroni, Nadav Merlis, Aadirupa Saha, and Shie Mannor. Confidence-budget matching for sequential budgeted learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2937–2947. PMLR, 2021. URL https://proceedings.mlr.press/v139/efroni21a.html.

Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk Van Der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: When do intermediate observations help? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9374–9395. PMLR, 2023. URL https://proceedings.mlr.press/v202/esposito23a.html.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Kyra Gan, Esmaeil Keyvanshokooh, Xueqing Liu, and Susan Murphy. Contextual bandits with budgeted information reveal. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3970–3978. PMLR, 2024. URL https://proceedings.mlr.press/v238/gan24a.html.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, pp. 540–550, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3511808.3557220. URL https://doi.org/10.1145/3511808.3557220.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *Biometrika*, 2026. doi: 10.1093/biomet/asag053. URL https://doi.org/10.1093/biomet/asag053. Advance article.

Zeyu Jia, Jian Qian, Alexander Rakhlin, and Chen-Yu Wei. How does variance shape the regret in contextual bandits? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9861a7c3972ed5d36dda3826d44bb246-Abstract-Conference.html.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):480–509, 2025. doi: 10.1093/jrsssb/qkae099. URL https://doi.org/10.1093/jrsssb/qkae099.

Predrag Klasnja, Shawna Smith, Nicholas J. Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B. Hekler, and Susan A. Murphy. Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019. doi: 10.1093/abm/kay067. URL https://doi.org/10.1093/abm/kay067.

Michael R. Kosorok and Thomas R. Fleming. Using surrogate failure time data to increase cost effectiveness in clinical trials. *Biometrika*, 80(4):823–833, 1993. doi: 10.1093/biomet/80.4.823. URL https://doi.org/10.1093/biomet/80.4.823.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Thomas M. McDonald, Lucas Maystre, Mounia Lalmas, Daniel Russo, and Kamil Ciosek. Impatient bandits: Optimizing recommendations for the long-term without delay. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3580305.3599386. URL https://doi.org/10.1145/3580305.3599386.

Reagan Mozer, Nicole E. Pashley, and Luke Miratrix. Stratified sampling for model-assisted estimation with surrogate outcomes. *arXiv preprint arXiv:2602.12992*, 2026. URL https://arxiv.org/abs/2602.12992.

Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018. doi: 10.1007/s12160-016-9830-8. URL https://doi.org/10.1007/s12160-016-9830-8.

Sunhyun Park and Seong-ho Lee. Surrogate-assisted optimal sampling for risk prediction under measurement constraints. *arXiv preprint arXiv:2606.03477*, 2026. URL https://arxiv.org/abs/2606.03477.

Ross L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989. doi: 10.1002/sim.4780080407. URL https://doi.org/10.1002/sim.4780080407.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0499. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2567c95fd41459a98a73ba893775d22a-Abstract-Conference.html.

Mohammad Shahverdikondori, Amir Mohammad Abouei, Alireza Rezaeimoghadam, and Negar Kiyavash. Pure exploration beyond reward feedback: The role of post-action context. *arXiv preprint arXiv:2502.03061*, 2025. URL https://arxiv.org/abs/2502.03061.

Lei Shi, Waverly Wei, and Jingshen Wang. Using surrogates in covariate-adjusted response-adaptive randomization experiments with delayed outcomes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 71085–71121, 2024. doi: 10.52202/079017-2272. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/833295cb9278a3ba973842a94ea68e3c-Abstract-Conference.html.

Ajay Narayanan Sridhar, Ronak Singh, Mehrdad Mahdavi, and Vijaykrishnan Narayanan. Correlation-aware contextual bandits with surrogate rewards for LLM routing. *arXiv preprint arXiv:2607.09015*, 2026. URL https://arxiv.org/abs/2607.09015.

Aaron D. Tucker, Caleb Biddulph, Claire Wang, and Thorsten Joachims. Bandits with costly reward observations. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2147–2156. PMLR, 2023. URL https://proceedings.mlr.press/v216/tucker23a.html.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48957–48975. PMLR, 2024. URL https://proceedings.mlr.press/v235/valancius24a.html.

Chaoqi Wang, Ziyu Ye, Zhe Feng, Ashwinkumar Badanidiyuru Varadaraja, and Haifeng Xu. Follow-ups also matter: Improving contextual bandits via post-serving contexts. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/29f47df77b7e536ebd0fe5e0cc964a32-Abstract-Conference.html.

Zeyu Xu, Xiaojie Mao, Hao Mei, and Yue Liu. Evaluating surrogates in individual treatment regimes. *arXiv preprint arXiv:2512.00405*, 2025. URL https://arxiv.org/abs/2512.00405.

Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and Markov decision processes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39834–39863. PMLR, 2023. URL https://proceedings.mlr.press/v202/ye23d.html.

Kelly W. Zhang, Thomas Baldwin-McDonald, Kamil Ciosek, Lucas Maystre, and Daniel Russo. Impatient bandits: Optimizing for the long-term without delay. *Journal of Machine Learning Research*, 27(156):1–58, 2026a. URL https://jmlr.org/papers/v27/25-0030.html.

Wenxin Zhang, Aaron Hudson, Maya Petersen, and Mark van der Laan. An online meta-level adaptive design framework with targeted learning inference: Applications to evaluating and utilizing surrogate outcomes in adaptive designs. *Journal of the American Statistical Association*, 2026b. doi: 10.1080/01621459.2026.2657052. URL https://doi.org/10.1080/01621459.2026.2657052. Accepted author version published online.

Tijana Zrnic and Emmanuel Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010. PMLR, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.