WHEN BIASED OFFLINE DATA HELP ONLINE LEARNING: THE ROLE OF ACTION CONTRASTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Historical reward levels can be systematically biased while preserving the action comparisons needed for a new decision problem. However, guarantees based on absolute reward-bias bounds can overlook this benefit when all actions share the same reward-level shift. We isolate this information object in a finite-cell Bernoulli contextual bandit by bounding the diameter of source–target reward biases across actions and leaving their common level unrestricted. With a supplied valid relative-bias bound β, M context cells, K actions, N balanced source observations, and T target rounds, upper and same-model minimax lower bounds share the rate $T\min\{\sqrt{MK/T}, \beta+\sqrt{MK/N}\}$ up to logarithms. For fixed M, K, the minimax regret is $o(\sqrt{T})$ if and only if $N/T \to \infty$ and $\beta\sqrt{T} \to 0$. Unlike absolute-discrepancy models, this class permits a large action-common shift; the familiar minimum-of-two rate is characterized here for that quotient information class. A relative-confidence allocation turns historical contrasts into observable gap bounds and permits any predictable proposal; information-directed sampling is one instantiation, not a condition for validity. When β must be learned, target calibration is charged to regret, and a $T^{-1/2}$ honesty barrier prevents this route from uniformly recovering the strict order gain within the same horizon. Controlled experiments recover the predicted phase boundary, while response-bank experiments expose the finite-sample costs of protection and of learning the contrast certificate.

1 INTRODUCTION

Existing reward-transfer methods such as MIN-UCB measure shift by the absolute difference between the offline and online mean rewards of each action (Cheung & Lyu, 2024). When these differences are large, their guarantees may be no better than those achievable with online data alone. Yet an offline reward estimate can be wrong in level and still be informative for action selection. If every action's reward changes by the same amount (a *common reward-level shift*), each historical mean is biased, but every pairwise advantage is preserved. For example, after a platform-wide change, the baseline response rate of every message in an email campaign may shift while the best message within a customer segment remains the same; by contrast, a smaller message-specific drift can reverse that choice. Such historical data pose a dilemma: discarding them wastes decision-relevant information, whereas treating their levels as current rewards risks false certainty. The statistical question is therefore not whether the source reproduces target reward levels, but what it preserves about *action contrasts* and how much target interaction remains necessary.

Related guarantees use different notions of shift. MIN-UCB uses coordinatewise bounds on absolute reward discrepancies (Cheung & Lyu, 2024); HybElimUCB-RA transfers offline Bradley–Terry comparisons to online rewards under pairwise preference-bias bounds (He et al., 2025); and Hybrid LinUCB intersects online and hybrid ellipsoids under a Euclidean parameter-discrepancy bound (Zhou et al., 2026). Direction-aware transfer instead separates informative and contaminated parameter directions (Han et al., 2026). Policy-class warm starting already uses task-relative source similarity (Zhang et al., 2019), while nonparametric contextual-bandit transfer assumes shared reward functions under covariate shift (Cai et al., 2024).

Table 1 separates single-action reward-bias control from relative action-bias control. Hybrid LinUCB controls reward levels through its parameter bound; HybElimUCB-RA already uses pairwise information, but observes offline comparisons rather than rewards. In our setting, the learner observes

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.

Table 1: Bias information used by the closest comparators. For reward feedback, $b_a = \nu_a - \mu_a$ denotes the bias of action a in a fixed context. Our bound controls differences between these biases. BT denotes Bradley–Terry preferences.

Method	Offline → online feedback	Bias quantity controlled
MIN-UCB	reward → reward	single-action bias: $\lvert b_a \rvert \le V_a$
HybElimUCB-RA	preferences (BT) → reward	pairwise preference discrepancy
Hybrid LinUCB	linear reward → linear reward	Euclidean parameter discrepancy
This paper	reward → reward	relative action bias: $\max_a b_a - \min_a b_a \le \beta$

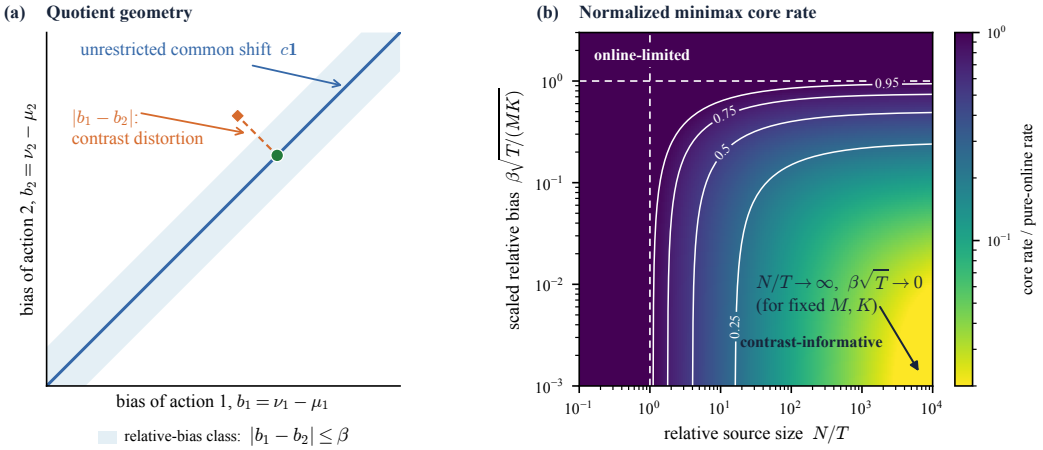


Figure 1: Common shifts and relative bias. (a) The band $|b_1 - b_2| \le \beta$ surrounds the common-shift line $c\mathbf{1}$. The orange diamond represents a bias vector outside the band; the green circle is its orthogonal projection onto that line. The dashed contrast component has length $|b_1 - b_2|/\sqrt{2}$. (b) The color map shows the core regret rate normalized by $\sqrt{MKT}$, namely $\min\{1, \beta\sqrt{T/(MK)} + \sqrt{T/N}\}$, which decreases toward the lower right. For fixed M, K, reaching that corner asymptotically requires $N/T \to \infty$ and $\beta\sqrt{T} \to 0$. Panel (b) visualizes the theoretical rate, not an empirical estimate.

absolute source rewards, but only the source-to-target changes in their pairwise differences are constrained; that is, the bias is restricted only modulo action-common shifts. We call this a *quotient* information class and make three contributions.

A minimax characterization for quotient information. We ask when this class changes the minimax order of online learning. In the finite-cell model of Section 2, with M context cells, K actions, N source observations and T target rounds, the answer is governed by two competing uncertainties: $\sqrt{MK/T}$ from target interaction and $\beta + \sqrt{MK/N}$ from historical contrasts, where β is a supplied valid bound on action-dependent bias. Treating action-common bias as a nuisance direction, we give upper and same-model minimax lower bounds that isolate source sampling error from relative bias and match these terms up to logarithms and constants (Sections 2 and 4). Consequently, for fixed M, K, the minimax regret improves from the pure-online $\Theta(\sqrt{T})$ order to $o(\sqrt{T})$ exactly when $N/T \to \infty$ and $\beta\sqrt{T} \to 0$ (Fig. 1b). This is a statement about the information available to every learner in the model class, not a pointwise necessity claim or a claim about one numerical implementation of information-directed sampling (IDS).

A safe modular realization. Our sequential rule, relative-confidence allocation, has three steps: source intervals filter out actions that cannot be target-optimal; source and target intervals bound the remaining action gaps; and a cumulative confidence constraint keeps a predictable proposal when it is safe to do so and otherwise injects the exploration needed to restore the constraint (Section 3). We instantiate the proposal with IDS (Russo & Van Roy, 2018; Chen, 2026); validity does not require its

posterior to be correct. The allocation provides modularity in the choice of proposal and a certificate for the executed trajectory, not a better worst-case rate than every source-filtered optimistic rule.

The price and empirical role of contrast information. The order-improvement criterion treats β as supplied information. When β is unknown, our independent target-pilot protocol certifies an upper envelope for it, charges every pilot observation to the horizon, and reuses pilot rewards without double counting. Even with unlimited source data, a uniformly valid envelope cannot shrink past $T^{-1/2}$ with high probability at a zero-bias instance. This honesty barrier means that contrast quality must be externally certified or learned at an explicit target-interaction cost (Section 5). We test the predicted boundary, common-shift invariance, and protection cost in controlled environments and in fixed response banks of recorded feedback (Section 6); one bank consists of recorded responses from a randomized email campaign and instantiates the example above.

2 The Relative-Bias Model

The model fixes what the learner may use—a biased source sample, online target feedback and a relative-bias upper bound—for any finite contextual decision problem with binary feedback; in the running campaign example, a cell is a prespecified customer segment, an action is a message, and a Bernoulli reward records whether the customer responds. Contexts X_t arrive independently and uniformly over M cells, each with $K \geq 2$ actions. Let $\mathcal{F}_{t-1}$ contain the source sample and all observations and learner randomization revealed through round $t-1$, and let $\mathcal{G}_t$ denote the information available after the learner observes X_t and randomizes a proposal, but before the final action draw. The learner chooses a $\mathcal{G}_t$-measurable distribution p_t, draws $A_t \mid \mathcal{G}_t \sim p_t$ using fresh randomness, and observes only the reward Y_t of this action, with

$$Y_t \mid \mathcal{G}_t, A_t \sim \mathrm{Bernoulli}(\mu_{X_t,A_t}). \tag{1}$$

Here $\nu_{x,a}, \mu_{x,a} \in [0,1]$ are the fixed source and target means. The source sample contains N observations, where N is a nonnegative multiple of MK. Each cell–action pair has $N/(MK)$ independent observations. The source sample, contexts, learner random seeds, and target reward streams are mutually independent before adaptive selection. The learner interacts with the target environment for T rounds, where $T \geq MK$. Its cumulative regret is $R_T = \sum_{t \leq T}(\max_a \mu_{X_t,a} - \mu_{X_t,A_t})$. Source acquisition lies outside this round budget, whereas any new target calibration rounds are charged to it.

Write $b_{x,a} = \nu_{x,a} - \mu_{x,a}$ and define the oscillation seminorm $\|b_x\|_{\mathrm{osc}} = \max_a b_{x,a} - \min_a b_{x,a}$. The learner receives a valid relative-bias bound $\beta \geq 0$:

$$\|b_x\|_{\mathrm{osc}} = \max_a(\nu_{x,a} - \mu_{x,a}) - \min_a(\nu_{x,a} - \mu_{x,a}) \leq \beta \quad \text{for every } x. \tag{2}$$

Condition (2) is equivalent to requiring that, for any two actions, the source reward difference and the target reward difference be within β of each other. It allows an unknown common shift in each cell, so source contrasts remain informative even when absolute reward levels change.

Let $\mathfrak{R}_{M,K,T,N}(\beta)$ denote the minimax expected regret: the infimum, over all adaptive learners obeying the information structure above, of the worst-case expected regret over all source and target means satisfying (2), with the stated source sample and supplied bound.

Relation to absolute-bias classes. If an absolute-bias model assumes $\|b_x\|_\infty \leq V$, then $\|b_x\|_{\mathrm{osc}} \leq 2V$. The converse fails: no absolute bound that vanishes with β is implied by (2). For any $c \in [-1/2, 1/2]$, setting every target mean to $(1-c)/2$ and every source mean to $(1+c)/2$ gives a valid Bernoulli instance with $b_x = c\mathbf{1}$, $\|b_x\|_{\mathrm{osc}} = 0$, and $\|b_x\|_\infty = |c|$. The present assumption is therefore not a reparameterized small absolute shift; it removes exactly the action-common nuisance direction illustrated in Fig. 1. In coordinates, fix a reference action a_0 and set $q_{x,a} = b_{x,a} - b_{x,a_0}$. Then

$$\max_a |q_{x,a}| \leq \|b_x\|_{\mathrm{osc}} \leq 2 \max_a |q_{x,a}|.$$

Thus the relative-bias bound is, up to constants, an ℓ_∞ ball on $K-1$ quotient coordinates. By contrast, a positive-definite ellipsoid on the original K coordinates cannot have the common-shift direction as its null space.

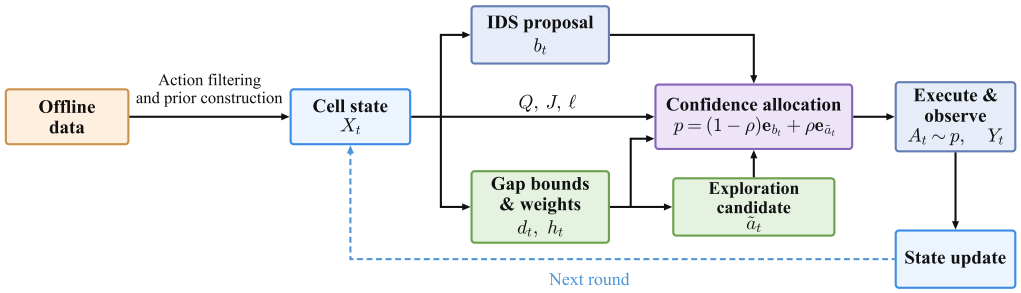


Figure 2: **Overview of Relative IDS.** Offline rewards may be biased in level yet preserve action contrasts; given a valid upper bound β on how far source and target contrasts may disagree (2), Relative IDS uses source contrasts for target decisions. Source filtering fixes each cell's active set; offline data inform its prior. In cell X_t, information-directed sampling (IDS) (Russo & Van Roy, 2018) draws a preferred action b_t; gap bounds d_t and weights h_t (3) select the exploration candidate $\tilde{a}_t$. The learner draws A_t from $p = (1-\rho)\mathbf{e}_{b_t} + \rho\,\mathbf{e}_{\tilde{a}_t}$ with the smallest ρ satisfying the prefix constraint (4), keeping b_t when $\rho = 0$ is feasible. The reward Y_t observed after executing A_t updates the current cell's counts and posterior; Q, J accumulate expected costs and ℓ advances. The IDS distribution is refreshed only periodically (Section E.4). Because validity rests on the confidence intervals, source filter and prefix constraint, the regret bound of Theorem 1 holds for any predictable proposal and matches the minimax rate up to logarithms and the failure term. Corollary 1 identifies when the minimax order improves over online-only learning. The final Q values across cells enter the executed-trajectory regret bound (10). Section 5 studies paid calibration of β; Section 6 tests these predictions.

3 Learning from Historical Contrasts

Relative-confidence allocation turns the source sample and β from Section 2 into decisions: it bounds action gaps and adds guarded exploration to any predictable proposal, here from IDS (Fig. 2).

Action filtering and observed gap bounds. Fix a cell and suppress the index x. Let $I(s, n; z) = \{p : n\,\mathrm{kl}(s/n, p) \le z\}$ be a Bernoulli confidence interval (Garivier & Cappé, 2011), with $I(0, 0; z) = [0, 1]$. Source intervals use $z_s = \log(4MK/\delta)$ and target intervals use $z_o = \log(4MKT/\delta)$. Their endpoints are L^s_a, U^s_a and $L_{t,a}, U_{t,a}$, respectively. The fixed source-active set

$$\mathcal{A} = \{a : U^s_a + \beta \ge \max_b L^s_b\}$$

retains a target-optimal action on the common confidence event. For $a \in \mathcal{A}$, define

$$d_{t,a} = \min\left\{1, \max\left(0, \max_{b\in\mathcal{A}\setminus\{a\}} \min\{U^s_b - L^s_a + \beta, U_{t,b} - L_{t,a}\}\right)\right\},$$
$$h_{t,a} = 1/\max(1, n_{t,a}), \qquad C = 2z_o. \tag{3}$$

Here $n_{t,a}$ counts the previous target observations of a in the cell. The gap bound uses the tighter of the source and target contrasts. For a singleton active set, set $d_{t,a} = 0$.

A minimal-exploration allocation. Let $b_t \in \mathcal{A}$ be a preferred action from a predictable proposal, let the exploration candidate $\tilde{a}_t$ minimize $d^2_{t,a}/h_{t,a}$ over $\mathcal{A}$, and let $\mathbf{e}_a \in \mathbb{R}^{\mathcal{A}}$ be the unit vector of action a. Initialize $Q = J = 0$ and all target counts to zero. At the cell's local round ℓ, choose the smallest $\rho \in [0, 1]$ such that $p = (1-\rho)\mathbf{e}_{b_t} + \rho\,\mathbf{e}_{\tilde{a}_t}$ satisfies

$$(Q + p^\top d_t)^2 \le C\ell(J + p^\top h_t). \tag{4}$$

The action with maximal target upper endpoint satisfies $d^2_{t,a} \le Ch_{t,a}$, so $\rho = 1$ is feasible on every history. We use $\rho = 0$ whenever it is feasible and otherwise find the first feasible point by scalar root search. We then draw $A_t \sim p$, update $Q \leftarrow Q + p^\top d_t$ and $J \leftarrow J + p^\top h_t$, and update that action's reward statistics. We call this cumulative allocation, which enforces (4), the *guard*. The guarantee in Theorem 1 uses the prefix constraint, the actual pull counts and source filtering, not the proposal, and Q enters the regret bound (10). A source-filtered rule that always maximizes the target

upper endpoint can attain the same two worst-case caps in Theorem 1, so the cumulative allocation is not claimed to improve that rule's asymptotic order. The experiments therefore separate empirical decision quality from the additional protection and certification supplied by the guard.

An information-directed proposal. Our IDS proposal uses a source-conditioned posterior that combines a common shift, action-specific residuals and a source-independent component. For the current cell, let $B = \arg\max_{a\in\mathcal{A}} \mu_a$ under this numerical posterior, and define

$$\Delta_{t,a}^{\mathrm{post}} = \mathbb{E}_t[\mu_B - \mu_a], \qquad g_{t,a} = I_t(B; Y_a), \qquad u_t \in \arg\min_{u\in\Delta(\mathcal{A})} \frac{(u^\top \Delta_t^{\mathrm{post}})^2}{u^\top g_t}. \tag{5}$$

Draw b_t from u_t and apply the guard; the guarantee covers the numerical proposal in Section E.

4 When Offline Information Improves Learning

Given a valid upper bound β, how much can source data help? Theorem 1 bounds minimax regret, Corollary 1 identifies order gains over online-only learning, and Section 4.1 bounds trajectory regret.

Theorem 1 (Minimax characterization with a supplied contrast certificate)**.** *For $0 < \delta < 1$, the allocation rule in Section 3 satisfies*

$$\begin{aligned} \mathbb{E}R_T &\le \min\{U_o, U_s\} + \delta T, \\ U_o &= \sqrt{2\log(4MKT/\delta)\, TMK(2+\log T)}, \\ U_s &= T\min\left\{1, 2\beta + 4\sqrt{\frac{MK\log(4MK/\delta)}{2N}}\right\}. \end{aligned} \tag{6}$$

In the same model, the minimax regret obeys the statistical lower bound

$$\mathfrak{R}_{M,K,T,N}(\beta) \ge \frac{T}{40}\max\left\{\sqrt{\frac{MK}{N+T}}, \min\left(\beta, \sqrt{\frac{MK}{T}}\right)\right\}. \tag{7}$$

For $N = 0$, omit U_s and interpret $\sqrt{MK/N} = \infty$ below. These bounds share the core rate

$$\Phi(M,K,T,N,\beta) = T\min\left\{\sqrt{MK/T},\ \beta + \sqrt{MK/N}\right\}, \tag{8}$$

up to logarithms, universal constants, and the displayed failure term.

In the historical term $\beta + \sqrt{MK/N}$, more source data reduce the sampling error $\sqrt{MK/N}$, whereas β reflects how well the source preserves target contrasts. The lower bound separates these two roles through two families of instances: a family with shared source and target means, which is limited by $N + T$ and gives the $\sqrt{MK/(N+T)}$ term, and a family with identical source laws but different target-optimal actions, which gives the $\min(\beta, \sqrt{MK/T})$ term; Section B gives the complete argument. An unrestricted level shift is not fatal because regret compares actions within a cell and is invariant to common shifts. Even at $\beta = 0$, however, an order gain requires the source sampling error to fall below the online uncertainty. These results do not make the allocation uniquely minimax, and they do not turn an estimated β into free information (Section 5).

Corollary 1 (Order improvement in the contrast-quotient class)**.** *For fixed M, K and sequences $N = N(T)$, $\beta = \beta(T)$,*

$$\boxed{\mathfrak{R}_{M,K,T,N}(\beta) = o(\sqrt{T}) \quad\Longleftrightarrow\quad N/T \longrightarrow \infty \ \text{ and } \ \beta\sqrt{T} \longrightarrow 0.} \tag{9}$$

Pure-online minimax regret on the same target class is $\Theta(\sqrt{T})$. Thus the displayed condition gives a strict worst-case order improvement, not merely a smaller upper-bound constant.

A concrete separation from pure-online learning. With $N \asymp T^2$, $\beta \le 1/T$, and $\delta = (1+N+T)^{-2}$, the allocation rule achieves $O(\sqrt{\log T})$ regret, against a pure-online minimax lower bound of $\Omega(\sqrt{T})$. More generally, the iff statement concerns the optimal achievable order: necessity follows from the lower bound, and sufficiency follows from the bound $T[\beta + \sqrt{2MK\log(2K)/N}]$ of the greedy rule based on source empirical means. The condition distinguishes an order improvement from a finite-budget gain; it also explains why increasing source size is most valuable when the remaining relative bias is below the online uncertainty scale.

Comparison with MIN-UCB. Let V bound every single-action absolute bias $|\nu_{x,a} - \mu_{x,a}|$. Under balanced sampling, the cellwise MIN-UCB regret upper bound has the same core form as (8), with V in place of β, up to constants and logarithms (Cheung & Lyu, 2024, Theorem 4.9). An absolute bound V always permits a valid choice $\beta \le 2V$, so the corresponding core rate is no worse in order; this does not imply pointwise dominance over MIN-UCB. A common reward-level shift can leave V nonvanishing while $\beta = 0$. MIN-UCB's bound may then reduce to its online term—MIN-UCB retains an online fallback, so a linear source-bias term does not imply linear total regret—whereas our class admits $o(\sqrt{T})$ regret under the conditions of Corollary 1. Beyond worst-case rates, with fixed positive gaps, enough source data can also make our filter retain only optimal actions: on the common confidence event, every retained gap is at most $2\beta + 4r_s$, where $r_s = \sqrt{MK \log(4MK/\delta)/(2N)}$, so if this bound is below the smallest positive gap, choosing $\delta = 1/T$ gives expected regret at most one.

4.1 A Certificate from the Executed Trajectory

Let Q_x be the final value of the accumulator Q in cell x. For a risk level η assigned to action randomization, a bounded martingale argument gives

$$R_T \le \min\left\{T, \sum_x Q_x + \sqrt{\tfrac{T}{2}\log(1/\eta)}\right\} \tag{10}$$

with probability at least $1 - \delta - \eta$. Observed counts, successes, and executed action probabilities determine the right-hand side. Whereas Theorem 1 characterizes what any learner can achieve over the model class, (10) certifies the particular trajectory that was executed.

5 Paying for Relative-Bias Information

Sections 3 and 4 take a valid upper bound β as given. We now consider obtaining such a bound from target rewards, with the pilot regret and calibration risk included in the guarantee. Our protocol fixes a pilot size c and collects c target observations of each action within each context cell. Contexts continue to arrive under the original law; the algorithm does not create a balanced prefix of contexts. A cell becomes active only after its own pilot is complete.

When a cell activates, the learner forms KL intervals with threshold $\log(4MK/\delta_c)$ for each action, from the source sample and from the pilot rewards. The two samples are independent, and with probability at least $1 - \delta_c$ all intervals hold simultaneously, so the cell's bias diameter is at most

$$\beta_x^U = \min\{2, \max_a(U_a^s - L_a^c) - \min_a(L_a^s - U_a^c)\},$$

where $[L_a^c, U_a^c]$ is the pilot interval of action a and the superscript c denotes calibration (Section C). The learner then runs the relative-confidence rule in this cell with β_x^U in place of β. It first replays the pilot rewards once to initialize the counts and posterior, and it sets the accumulators to $Q = J = 0$; the local clock ℓ counts the pilot observations. Reusing these rewards improves the target intervals but does not make the pilot free: the guarantee below includes its rounds and the calibration risk δ_c.

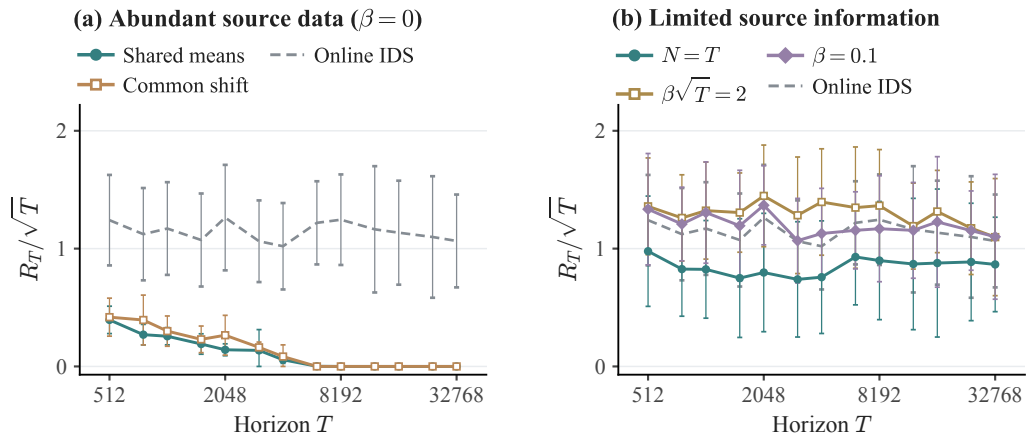
Let B_T be the number of target rounds spent on pilots, including unfinished pilots in rarely visited cells. On a joint event of probability at least $1 - \delta_c - \delta - \eta$, the resulting procedure satisfies

$$R_T \le \min\left\{T, B_T + \sum_x Q_x + \sqrt{\frac{T}{2}\log\frac{1}{\eta}}\right\}. \tag{11}$$

Every quantity on the right-hand side is observed or declared. The proof in Section D handles initialization, random activation times and pilot reuse. A smaller pilot yields a broader bias envelope but leaves more target rounds for learning; Section 6 tunes this trade-off on a development suite.

Estimating a sharp envelope is also statistically different from receiving one. For two arms with source means $1/2$, every uniformly valid observed upper bound U on the true bias diameter, computed from at most T target observations, obeys at the zero-bias target instance with both means $1/2$,

$$\Pr\{U \ge 1/(8\sqrt{T})\} \ge 3/4 - \delta_c.$$



(c) Method comparison at $T = 32768$

	Shared means	Common shift	$N = T$	Boundary bound	Fixed bound
Relative IDS	**0.00**	**0.00**	**0.87**	**1.10**	**1.10**
Greedy + budget	0.00	0.00	0.94	1.53	1.53
MIN-UCB	0.00	1.42	1.24	1.50	1.47
Directional MIN-UCB	0.00	1.44	1.45	1.51	1.48
Anchor-TS (source)	0.77	1.19	1.02	1.18	1.24
Source IDS	0.00	0.00	0.74	1.52	1.52
Source greedy	0.00	0.00	0.94	1.53	1.53
Source ε-TS (1/K)	0.00	0.00	0.87	1.55	1.55
Online IDS	1.06	1.06	1.06	1.06	1.06
Online ε-TS (0.75)	1.23	1.23	1.23	1.23	1.23
Anchor-TS (online)	1.33	1.33	1.33	1.33	1.33
Exp-KL-MS	1.35	1.35	1.35	1.35	1.35
LinMED	1.12	1.12	1.12	1.12	1.12

Figure 3: Source information and regret scaling ($R_T/\sqrt{T}$, lower is better). (a,b) Relative IDS (colored) and Online IDS (dashed gray); each T uses a separate instance family. Error bars are approximate 95% simultaneous t intervals over 845 method–setting means from sixteen paired seeds. (c) Means of all 13 methods at $T = 32768$; bold identifies Relative IDS, and methods without source data repeat across regimes. Boundary and fixed bounds are $\beta\sqrt{T} = 2$ and $\beta = 0.1$.

This result, proved in Proposition 2, is a lower bound on honest certification, not on the regret of a learner that receives β from an external source. It shows that the source-size condition alone cannot remove the need for target information to learn an unknown diameter, so the supplied-certificate condition of Corollary 1 cannot simply be replaced by unlimited source data. In particular, target-only calibration using at most the same T deployment rounds cannot uniformly produce the $o(T^{-1/2})$ certificate required by Corollary 1. The strict order improvement is therefore an oracle-information statement unless external data or stronger structure identify the contrast quality. Related adaptation impossibilities arise for fixed-confidence best-arm identification with biased offline data (Yang et al., 2025); here the obstruction concerns an honest horizon-scale upper envelope.

Table 2: Cumulative regret at $T = 12000$ with supplied bias information (lower is better): Relative IDS and all 19 external baseline configurations. Each source-budget column averages 15 settings; the last column averages all 45. Entries are means ± standard errors over 32 seeds, with settings averaged within each seed. Bold marks the lowest mean in each column, not statistical significance; paired intervals are in Table 8 and internal variants in Table 6.

	Source observations per cell–action			
Method	128	2,048	32,768	All 45 settings
Online controls				
Online IDS	125.44 ± 4.90	125.44 ± 4.90	125.44 ± 4.90	125.44 ± 4.90
Online Beta-TS	154.77 ± 4.17	154.77 ± 4.17	154.77 ± 4.17	154.77 ± 4.17
Online ϵ-TS $(1/K)$	152.91 ± 11.03	152.91 ± 11.03	152.91 ± 11.03	152.91 ± 11.03
Online ϵ-TS (0.75)	143.18 ± 5.93	143.18 ± 5.93	143.18 ± 5.93	143.18 ± 5.93
Exp-KL-MS	203.46 ± 3.98	203.46 ± 3.98	203.46 ± 3.98	203.46 ± 3.98
LinMED	205.77 ± 8.75	205.77 ± 8.75	205.77 ± 8.75	205.77 ± 8.75
Anchor-TS, empirical ($\sigma = 0.5$)	146.95 ± 12.05	146.95 ± 12.05	146.95 ± 12.05	146.95 ± 12.05
Anchor-TS, empirical ($\sigma = 1$)	162.17 ± 6.24	162.17 ± 6.24	162.17 ± 6.24	162.17 ± 6.24
Anchor-TS, posterior ($\sigma = 1$)	162.81 ± 6.03	162.81 ± 6.03	162.81 ± 6.03	162.81 ± 6.03
Direct source reuse				
Source IDS	152.46 ± 3.17	268.67 ± 1.18	247.18 ± 0.75	222.77 ± 1.32
Source TS	180.53 ± 2.66	277.06 ± 1.23	247.70 ± 0.80	235.10 ± 1.28
Source ϵ-TS $(1/K)$	158.58 ± 2.59	270.24 ± 1.28	247.24 ± 0.75	225.35 ± 1.19
Source ϵ-TS (0.75)	175.57 ± 3.00	275.20 ± 1.26	247.59 ± 0.79	232.78 ± 1.37
Source greedy	342.32 ± 6.55	260.31 ± 1.75	247.76 ± 1.15	283.46 ± 2.30
Bias-aware baselines				
MIN-UCB	233.76 ± 2.50	171.60 ± 2.07	139.85 ± 1.56	181.73 ± 1.88
Directional MIN-UCB	837.34 ± 0.88	645.05 ± 0.94	615.28 ± 1.50	699.22 ± 0.76
Anchor-TS, empirical ($\sigma = 0.5$)	113.36 ± 3.86	107.23 ± 4.13	104.44 ± 4.30	108.34 ± 4.00
Anchor-TS, empirical ($\sigma = 1$)	177.85 ± 3.44	165.28 ± 3.15	157.93 ± 3.03	167.02 ± 3.08
Anchor-TS, posterior ($\sigma = 1$)	175.28 ± 3.55	163.04 ± 2.76	154.59 ± 2.54	164.30 ± 2.86
Proposed method				
Relative IDS	**90.69 ± 2.27**	**50.56 ± 1.93**	**38.35 ± 1.38**	**59.87 ± 1.26**

6 EXPERIMENTS

We test the information boundary of Corollary 1 and evaluate transfer with supplied or calibrated bias bounds (Sections 4 and 5). Controls include Thompson sampling (TS) variants, LinMED (Balagopalan & Jun, 2025), Exp-KL-MS (Qin et al., 2026) and Anchor-TS (Li et al., 2026), each under its stated information contract. We use controlled instances and fixed response banks of recorded feedback; resampling a bank is replay, not deployment. Calibration rounds count toward regret.

Information boundary. Figure 3 shows the boundary predicted by Corollary 1: abundant source data with preserved contrasts reduce normalized regret toward zero, including under a common $+0.1$ shift, whereas limited or uninformative source data leave regret on the online scale. Panel (c) compares all 13 methods at $T = 32768$; Sections H.1 and K give the setup and full results.

Method comparison. Relative IDS receives the relative-bias bound and a common-shift scale for its numerical prior; each bias-aware baseline receives a valid bias bound in the form it requires. Relative IDS has lower mean regret than each of the 19 external configurations at all three source budgets and overall (Table 2). Its overall mean, 59.87, is 44.7% below that of the strongest external configuration (108.34), and all 19 overall paired intervals lie above zero (Table 8). These results average five source mechanisms; Tables 6 and 7 report each mechanism and the internal variants.

Transfer under source shifts. Under the $+0.1$ common shift (Fig. 4a), Relative IDS has mean regret 37.65, versus 125.44 for Online IDS and 233.87 for MIN-UCB; at $\beta = 0.08$ (panel b), its mean is 60.74. Averaged over five mechanisms and three gaps, increasing source observations per

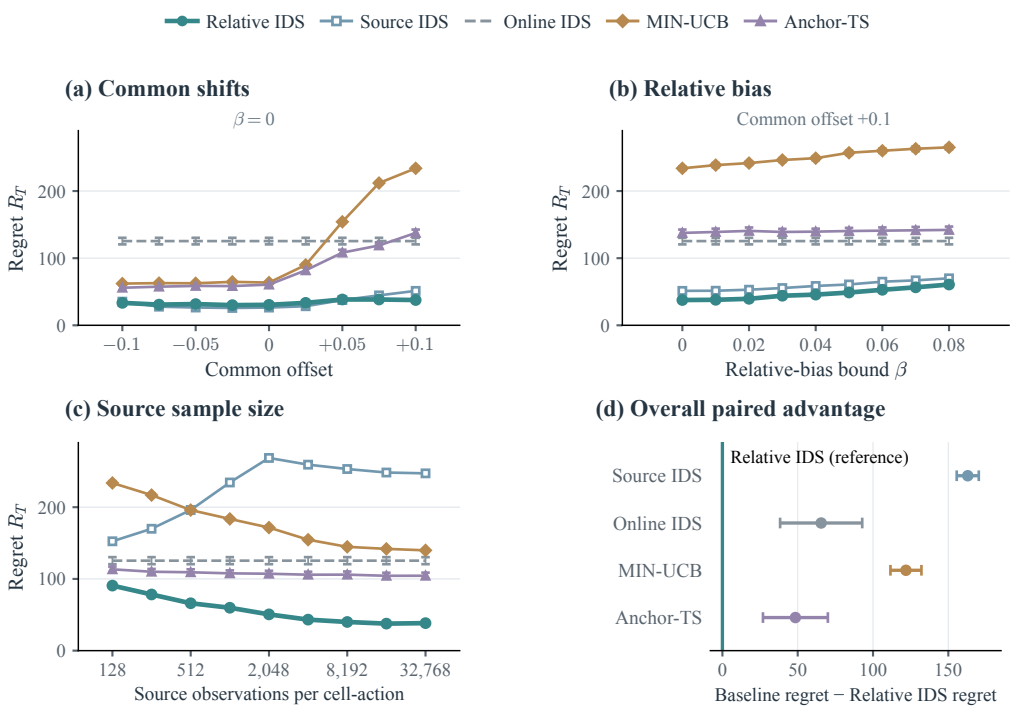


Figure 4: Transfer with supplied bias information ($T = 12000$). (a) Common offsets ($\beta = 0$). (b) A common $+0.1$ offset with uniform action residuals in $[-\beta/2, \beta/2]$. Panels (a,b) average three gaps and three budgets; (c) averages all five source mechanisms and three gaps per budget. In (a–c), error bars show one Monte Carlo standard error over 32 seeds, averaging settings within each seed. (d) Baseline minus Relative IDS regret over the 45 settings in Table 2, with approximate simultaneous 95% paired-t intervals for the 28,800-comparison family (Section H.5); positive values favor Relative IDS. Anchor-TS uses source-assisted empirical means with $\sigma = 0.5$.

cell–action from 128 to 32,768 lowers its mean regret from 90.69 to 38.35 (panel c). All four paired intervals in panel (d) lie above zero.

Transfer after target calibration. A development suite selects $c = 8$ calibration observations per cell–action for Relative IDS; a $c = 128$ variant tests a larger budget. Validation uses 32 new paired seeds, five mechanisms and five response banks, with OBD's campaigns averaged within each seed. Protocols, calibration costs and guard comparisons appear in Sections F, G and L. Independent-reference evaluations account for policy-selection uncertainty (Sections I and N).

7 Discussion and Conclusion

Biased offline rewards help online learning through the action contrasts they preserve, not through the accuracy of their common level. We have characterized the minimax order for the resulting quotient information class, separating source sampling error from relative bias while permitting an unrestricted action-common shift. The familiar minimum of online and source-assisted uncertainty also appears in earlier hybrid-bandit models; the distinction here lies in proving which source discrepancies matter for an absolute-reward decision problem. Relative-confidence allocation converts this information into safe decisions and an executed-trajectory certificate without tying validity to a particular proposal. When the contrast certificate is not supplied, target calibration must establish it and pay for the required interaction. The strict order improvement, however, assumes an external certificate. The response-bank results come from replay rather than deployment, and the matched-pilot experiments do not show that the cumulative guard improves mean regret. Extending the quotient geometry and the certification lower bound beyond independent finite cells is the natural next problem, and unbalanced source allocation and structured cross-cell sharing are the clearest next tests.

Reproducibility statement

The supplement specifies the source-conditioned prior, allocation rule, calibration protocol, comparison methods, data roles and risk budgets. The accompanying implementation records code and input hashes, selected actions, observed rewards, calibration costs and constraint diagnostics. Development choices are frozen before independent-seed validation. Simulation uncertainty is distinguished from uncertainty about a source population represented by a finite feedback bank.

Ethics statement

The experiments reuse released marketing, treatment and recommendation records. No new treatment, advertisement or recommendation is delivered. Response-bank sampling evaluates the stated conditional learning problem; it is not presented as a new clinical or online intervention. Dataset provenance, missing outcomes and action definitions are specified in the experimental supplement.

AI use statement

Generative AI tools helped with language editing and checking the mathematical proofs. The proofs are given in the appendix, and the supplementary ZIP archive contains Lean code that verifies selected deterministic claims. The authors verified all AI-generated content themselves.

References

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/022.pdf.

Kapilan Balagopalan and Kwang-Sung Jun. Minimum empirical divergence for sub-Gaussian linear bandits. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1585–1593, 2025. URL https://proceedings.mlr.press/v258/balagopalan25a.html.

Changxiao Cai, T. Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207–232, 2024. doi: 10.1214/23-AOS2341.

Keru Chen. Information-directed offline-to-online reinforcement learning. *arXiv preprint arXiv:2605.29405*, 2026. URL https://arxiv.org/abs/2605.29405.

Wang Chi Cheung and Lixing Lyu. Leveraging (Biased) information: Multi-armed bandits with offline data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8286–8309, 2024. URL https://proceedings.mlr.press/v235/cheung24a.html.

Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, Théophane Gregoir, and Thibaud Rahier. A large scale benchmark for individual treatment effect prediction and uplift modeling. *arXiv preprint arXiv:2111.10106*, 2021. URL https://arxiv.org/abs/2111.10106.

Dylan J. Foster, Yanjun Han, Jian Qian, and Alexander Rakhlin. Online estimation via offline estimation: An information-theoretic framework. In *Advances in Neural Information Processing Systems*, volume 37, pp. 42840–42898, 2024. doi: 10.52202/079017-1357. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4b6908c6695009dd84846e273d6c2607-Abstract-Conference.html.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 3953–3957, 2022. doi: 10.1145/3511808.3557624. URL https://kuairand.com/.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 359–376, 2011. URL https://proceedings.mlr.press/v19/garivier11a.html.

Zean Han, Ruihan Lin, Zezhen Ding, and Jiheng Zhang. Direction-aware offline-to-online learning in linear contextual bandits. arXiv preprint arXiv:2604.24016, 2026. URL https://arxiv.org/abs/2604.24016.

Qijia He, Minghan Wang, Xutong Liu, Zhiyong Wang, and Fang Kong. Learning across the gap: Hybrid multi-armed bandits with heterogeneous offline and online data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/1d3d77002cdb7fc73fec66e4dd2d6dbe-Abstract-Conference.html.

Kevin Hillstrom. The MineThatData e-mail analytics and data mining challenge. MineThatData blog, 2008. URL https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html. Randomized three-arm email experiment, 64000 customers.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012. doi: 10.1214/ECP.v17-2079. URL https://www.cs.columbia.edu/~djhsu/papers/quadratic-ecp.pdf.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401. URL https://tor-lattimore.com/downloads/book/book.pdf.

Bochao Li, Yao Fu, Wei Chen, and Fang Kong. Sample-mean anchored Thompson sampling for offline-to-online learning with distribution shift. *arXiv preprint arXiv:2605.10289*, 2026. URL https://arxiv.org/abs/2605.10289.

Hao Qin, Kwang-Sung Jun, and Chicheng Zhang. Achieving adaptivity and optimality for multi-armed bandits using Exponential-Kullback Leibler Maillard Sampling. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=IuVkRmecVp.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018. doi: 10.1287/opre.2017.1663. URL https://doi.org/10.1287/opre.2017.1663.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/33e75ff09dd601bbe69f351039152189-Abstract-round2.html.

Peter Sandercock, Maciej Niewada, and Anna Czlonkowska. International stroke trial database (version 2). Dataset, University of Edinburgh, Department of Clinical Neurosciences, 2011. URL https://doi.org/10.7488/ds/104. Corrected data release.

Sushant Vijayan, Arun Suggala, Karthikeyan Shanmugam, and Soumyabrata Pal. Regret minimization in linear bandits with offline data via extended D-optimal exploration. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=4WcK8gKgCi.

Le Yang, Vincent Y. F. Tan, and Wang Chi Cheung. Best arm identification with possibly biased offline data. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4715–4730, 2025. URL https://proceedings.mlr.press/v286/yang25b.html.

Chicheng Zhang, Alekh Agarwal, Hal Daumé III, John Langford, and Sahand Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7335–7344, 2019. URL https://proceedings.mlr.press/v97/zhang19b.html.

Kongchang Zhou, Zheshun Wu, Wei Chen, and Fang Kong. Regret analysis of hybrid linear bandits with biased offline data. Manuscript submitted to ICLR, 2026. URL https://openreview.net/forum?id=NA4seLpkPd.

APPENDIX CONTENTS

A The Scalar Allocation Step **14**

B Relative-Confidence Allocation for Bernoulli Rewards **14**
- B.1 A source-relative confidence rule 14
- B.2 Matching statistical models for the upper and lower bounds 15

C Observed Regret and Bias Calibration **17**
- C.1 A trajectory certificate from observed counts 17

D Causal Paid Calibration and Warm Starts **18**

E Source-Conditioned IDS: Complete Specification **19**
- E.1 Source-conditioned prior 19
- E.2 The information-directed proposal 20
- E.3 One complete round 20
- E.4 Periodic proposal refreshes preserve the guarantees 21
- E.5 Computation and regret across refresh intervals 21

F Paid-Calibration Experiment Protocol **22**

G Complete Paid-Calibration Validation **24**

H Experimental Design and Complete Comparisons **29**
- H.1 Source information and the rate boundary 29
- H.2 Strong baselines under explicit information contracts 29
- H.3 Controlled Bernoulli instances 29
- H.4 Baselines and information supplied 30
- H.5 Complete-grid performance with supplied information 30
- H.6 Randomized email data 34
- H.7 Reproduction 34

I Independent Randomized-User Policy Evaluation **34**

J Joint Confidence Bounds for Policy Contrasts **35**

K Testing the Bernoulli Gain Boundary across Horizons **36**

L A Source-Filtered Target KL-UCB Control **44**

M Anchor-TS, OOPE and Offline-Oracle Comparisons **44**
- M.1 A common model and fixed parameters 44
- M.2 Implementation contracts 45
- M.3 Paired extensions and uncertainty 45
- M.4 Main empirical observations 46
- M.5 Aggregate comparisons and learning curves 46
- M.6 Complete setting means 57

N Real-Feedback Benchmarks Across Applications **64**
- N.1 Data, actions and population 64
- N.2 Feedback access and charged calibration 65
- N.3 Held-out reference evaluation 67
- N.4 All Methods and Source Budgets 68

A The Scalar Allocation Step

Throughout this appendix, notation is local to one cell. We show that the smallest feasible mixing weight exists and can be found by bisection. By the prefix constraint at earlier rounds, the accumulators satisfy $Q^2 \le C(\ell-1)J$ before local round ℓ. Write $(d_b, h_b) = (d_{t,b_t}, h_{t,b_t})$ and $(d_e, h_e) = (d_{t,\tilde{a}_t}, h_{t,\tilde{a}_t})$ for the costs of the preferred action and the exploration candidate, where $d_e^2 \le Ch_e$. With $u = d_e - d_b$ and $v = h_e - h_b$, the feasibility polynomial is

$$F(\rho) = (Q + d_b + \rho u)^2 - C\ell(J + h_b + \rho v).$$

Since F is a quadratic in ρ with nonnegative leading coefficient u^2, it is continuous and convex. At $\rho = 1$, Cauchy–Schwarz gives

$$(Q + d_e)^2 \le C(\sqrt{(\ell-1)J} + \sqrt{h_e})^2 \le C\ell(J + h_e),$$

so $F(1) \le 0$. If $F(0) \le 0$, use $\rho = 0$. Otherwise the set $\{\rho \in [0,1] : F(\rho) \le 0\}$ is a closed interval $[\rho_*, 1]$ with $\rho_* > 0$, and bisection finds its left endpoint. Sampling from the resulting distribution and updating only the chosen action's reward statistics preserves the filtration of Section 2. Numerically, the implementation checks the original constraint up to a roundoff tolerance. The guarantees use the exact constraint $F(\rho) \le 0$, so a positive numerical residual must be recorded rather than identified with exact feasibility; such a residual does not change which rewards are observable.

The proof that an exploration action with $d_e^2 \le Ch_e$ exists, the source filtering argument, both same-model minimax bounds and the strict-order witness are supplied in Section B. The observed-regret and independent-calibration arguments are in Section C.

B Relative-Confidence Allocation for Bernoulli Rewards

We prove the allocation upper bound and the minimax lower bound in the same Bernoulli model, then compare their dependence on the source budget, target horizon, and relative bias.

B.1 A source-relative confidence rule

Fix one cell, with target and source means $\mu_a, \nu_a \in [0,1]$. The source contains n_a^s independent Bernoulli observations of each arm a, independent of future target feedback. The learner is given a valid relative-bias bound

$$\max_a(\nu_a - \mu_a) - \min_a(\nu_a - \mu_a) \le \beta. \tag{12}$$

The unknown common shift need not be small. Conditional on the selected arm and past feedback, each target reward is Bernoulli with its fixed mean. The horizon is at most T and the risk budget is δ.

Write $\mathrm{kl}(p,q)$ for binary relative entropy. For s successes in $n > 0$ trials, use the endpoints of

$$I(s,n;z) = \{p \in [0,1] : n\,\mathrm{kl}(s/n, p) \le z\}, \qquad I(0,0;z) = [0,1].$$

Source intervals $[L_a^s, U_a^s]$ use $z_s = \log(4K/\delta)$; target intervals $[L_{t,a}, U_{t,a}]$ use $z_o = \log(4KT/\delta)$. Counts, successes, and declared risk budgets are the only inputs to these intervals. Source filtering is optional:

$$A = \{a : U_a^s + \beta \ge \max_b L_b^s\}. \tag{13}$$

The unfiltered version uses $A = [K]$. All preferences and exploration choices are restricted to A. For $b \ne a$, define

$$C_t(b,a) = \min\{U_b^s - L_a^s + \beta, U_{t,b} - L_{t,a}\}, \qquad C_t(a,a) = 0,$$

and let

$$d_{t,a} = \min\{1, \max\{0, \max_{b \in A} C_t(b,a)\}\}, \quad h_{t,a} = \frac{1}{\max\{1, n_{t,a}\}}, \quad C = \max\{1, 2z_o\}. \tag{14}$$

Counts $n_{t,a}$ exclude the current reward. A preferred arm can be chosen by any predictable rule. Choose an exploration arm in A with $d_{t,a}^2 \le Ch_{t,a}$, using the smallest such ratio in our implementation. Mix it with the preferred arm using the smallest exploration probability that satisfies

$$Q_t^2 \le CtJ_t, \qquad Q_t = \sum_{s \le t} p_s^\top d_s, \quad J_t = \sum_{s \le t} p_s^\top h_s. \tag{15}$$

The scalar search is the one in Section A. Only the action actually executed supplies a reward update.

Proposition 1 (Binary confidence and feasibility). *Under* (12)*, with probability at least* $1-\delta$*, the target-optimal arm remains in* A *and, in every round, each* $d_{t,a}$ *bounds the target gap of arm* a*. The set of exploration arms* $\{a \in A : d_{t,a}^2 \le Ch_{t,a}\}$ *is nonempty for every realized history, including histories outside that confidence event, and the prefix constraint is feasible.*

Proof. For a fixed count n, Bernoulli Chernoff bounds give $\Pr(n\,\mathrm{kl}(\widehat{p},p) > z) \le 2e^{-z}$. We apply union bounds over the K source coordinates and over the KT possible target arm–count pairs, allocating risk $\delta/2$ to each group. Adaptive selection preserves the per-arm Bernoulli sequence indexed by its sample count; it does not give the learner future feedback. All stated intervals therefore cover their means on one event.

Let a^* be target-optimal. The relative-bias condition gives $\nu_b - \nu_{a^*} \le \beta$ for every b, hence $L_b^s \le U_{a^*}^s + \beta$ on this event. Thus $a^* \in A$. Both entries in the minimum defining $C_t(b,a)$ bound $\mu_b - \mu_a$, so taking the maximum over a set containing a^* gives the claimed gap bound.

For feasibility, let b maximize target $U_{t,b}$ over A. The target part of the construction gives $d_{t,b} \le U_{t,b} - L_{t,b}$. Pinsker's inequality places the KL interval inside its Hoeffding outer interval; when $n_{t,b} > 0$, its width is at most $\sqrt{2z_o/n_{t,b}}$. For zero count its width is one. Thus $d_{t,b}^2 \le Ch_{t,b}$ on every history, without assuming coverage. If the previous prefix obeys its constraint, the pure exploration choice ($\rho = 1$) on an arm with $d^2 \le Ch$ satisfies

$$(Q_{t-1}+d)^2 \le C\big(\sqrt{(t-1)J_{t-1}} + \sqrt{h}\big)^2 \le Ct(J_{t-1}+h).$$

Continuity in the mixing probability completes the induction. □

B.2 Matching statistical models for the upper and lower bounds

Now let contexts be independent and uniform on M cells, with $K \ge 2$ arms and $T \ge MK$. Source observations are balanced: $n_a^s = N/(MK)$. Source and target means lie in $[0,1]$, and (12) holds separately in each cell. The cellwise algorithms use risk δ/M and maximum horizon T. Let $\mathfrak{R}^{\mathrm{bin}}_{M,K,T,N}(\beta)$ be the minimax expected regret with exactly this information, including the known valid β.

Theorem 2 (Same-model binary bounds). *The filtered rule* (13)–(15) *satisfies*

$$\begin{aligned}
&\mathbb{E}R_T \le \min\{U_o, U_s\} + \delta T,\\
&U_o = \sqrt{C_M T\, MK(2+\log T)}, \quad C_M = \max\{1, 2\log(4MKT/\delta)\},\\
&U_s = T\min\left\{1, 2\beta + 4\sqrt{\frac{MK\log(4MK/\delta)}{2N}}\right\}.
\end{aligned} \tag{16}$$

For $N = 0$*, omit* U_s*. The minimax regret over all source-assisted algorithms obeys the lower bound*

$$\mathfrak{R}^{\mathrm{bin}}_{M,K,T,N}(\beta) \ge \frac{T}{40}\max\left\{\sqrt{\frac{MK}{N+T}}, \min\left\{\beta, \sqrt{\frac{MK}{T}}\right\}\right\}. \tag{17}$$

Thus the bounds share the core rate $T\min\{\sqrt{MK/T}, \beta + \sqrt{MK/N}\}$*, up to explicit logarithms, universal constants, and the displayed failure term. Note that the source term* U_s *is an algorithmic upper bound, not a lower confidence endpoint for an individual run.*

Upper bound. Let $\mathcal{E}$ be the common confidence event. Pathwise,

$$R_T \le \sum_{t=1}^{T} d_{t,A_t} + T\mathbf{1}\{\mathcal{E}^c\}.$$

Indeed, on $\mathcal{E}$ each realized target gap is at most d_{t,A_t}; on its complement, $R_T \le T \le \sum_t d_{t,A_t} + T$. Include the proposal draw in the pre-action sigma-field $\mathcal{G}_t$. The final distribution p_t and cost vector d_t are then $\mathcal{G}_t$-measurable, and $A_t \mid \mathcal{G}_t \sim p_t$. Hence the unconditional tower property gives

$$\mathbb{E}\sum_{t=1}^{T} d_{t,A_t} = \mathbb{E}\sum_{t=1}^{T} p_t^\top d_t = \mathbb{E}\sum_x Q_x.$$

Since $\Pr(\mathcal{E}^c) \le \delta$, it remains to bound this last term. Let T_x be the realized number of target visits to cell x. The final cellwise prefix constraints hold pathwise as $Q_x^2 \le C_M T_x J_x$. Hence Cauchy–Schwarz gives, on every realized history,

$$\sum_x Q_x \le \sqrt{C_M} \sum_x \sqrt{T_x J_x} \le \sqrt{C_M T \sum_x J_x}.$$

Taking expectations and then applying Jensen's inequality yields

$$\mathbb{E} \sum_x Q_x \le \sqrt{C_M T}\, \mathbb{E} \sqrt{\sum_x J_x} \le \sqrt{C_M T\, \mathbb{E} \sum_x J_x}.$$

For the remaining expectation, the vector h_t is also $\mathcal{G}_t$-measurable. The same unconditional tower property gives

$$\mathbb{E} \sum_x J_x = \mathbb{E} \sum_{t=1}^{T} p_t^\top h_t = \mathbb{E} \sum_{t=1}^{T} \frac{1}{\max\{1, n_{t,X_t,A_t}\}} \le MK(2 + \log T).$$

The last inequality is pathwise after grouping realized pulls by cell–action and applying the harmonic-sum bound. This tower-property calculation is not conditioned on the common confidence event; that event is used only to replace target gaps by the observable costs $d_{t,a}$. This proves the bound $U_o + \delta T$; the argument does not assume that the proposal is a calibrated posterior.

For the source term put $r_s = \sqrt{MK \log(4MK/\delta)/(2N)}$. The source interval widths obey $U_a^s - L_a^s \le 2r_s$. If $a \in A$, then $\max_b L_b^s \le U_a^s + \beta$, whence

$$\max_{b \in A}(U_b^s - L_a^s + \beta) \le (U_a^s - L_a^s) + \max_b (U_b^s - L_b^s) + 2\beta \le 4r_s + 2\beta.$$

Every selected action lies in A, so the accumulated cost $\sum_x Q_x$ is also at most U_s on every history. Its expectation is bounded by both U_s and U_o; taking their minimum and adding the common bad-event contribution proves (16). □

Lower bound. We use the bandit divergence decomposition (Lattimore & Szepesvári, 2020, Lemma 15.1), including the likelihood contribution of the independent source sample. Draw the best-arm index I_x of each cell independently and uniformly from $[K]$. In the first family both source and target means in cell x are $1/2 + \epsilon \mathbf{1}\{a = I_x\}$, so relative bias is zero. Condition on the context sequence and other cells' indices. In cell x, compare the instance P_i, whose best arm is i, with the instance P_0, in which all means in the cell are $1/2$. Let n_x be the number of visits and $N_{x,i}$ the number of pulls of arm i. For a fixed history, the same policy uses the same conditional action kernel in both instances, and other-cell reward laws coincide; these terms therefore contribute zero conditional KL, even though the distributions of the action trajectories can differ. The source contributes $N/(MK)$ times the single-observation divergence, and each target pull of arm i contributes one such divergence. Since $\mathrm{kl}(1/2, 1/2 + \epsilon) = -\frac{1}{2}\log(1 - 4\epsilon^2) \le 8\epsilon^2$ for $\epsilon \le 1/8$, the adaptive chain rule yields

$$\mathrm{KL}(P_0, P_i) \le 8\epsilon^2 \left(\frac{N}{MK} + \mathbb{E}_0 N_{x,i} \right).$$

Since $0 \le N_{x,i} \le n_x$, Pinsker and averaging over i give

$$\frac{1}{K} \sum_i \mathbb{E}_i N_{x,i} \le \frac{n_x}{K} + 2\epsilon n_x \sqrt{\frac{N}{MK} + \frac{n_x}{K}}.$$

Here $\sum_i \mathbb{E}_0 N_{x,i} = n_x$ and concavity of the square root handle adaptively allocated pulls. The average regret in this cell is therefore at least

$$\epsilon n_x \left(1 - \frac{1}{K} - 2\epsilon \sqrt{\frac{N}{MK} + \frac{n_x}{K}} \right).$$

Write $\lambda = T/M$. The visit count n_x is binomial with mean λ and second moment at most $\lambda^2 + \lambda$. Define its size-weighted law by $\Pr(\widetilde{n}_x = j) = j \Pr(n_x = j)/\lambda$. Then $\mathbb{E}\widetilde{n}_x = \mathbb{E}n_x^2/\lambda \le \lambda + 1$, so Jensen under this law gives

$$\mathbb{E}\left[n_x \sqrt{\frac{N}{MK} + \frac{n_x}{K}}\right] \le \lambda \sqrt{\frac{N}{MK} + \frac{\lambda + 1}{K}}.$$

After summing cells, choose $\epsilon = \frac{1}{8}\sqrt{MK/(N+T+M)}$. The bracket is at least $1/4$ since $K \ge 2$. Thus the Bayes risk is at least $\frac{T}{32}\sqrt{MK/(N+T+M)}$, which is at least $\frac{T}{40}\sqrt{MK/(N+T)}$ because $T \ge MK \ge 2M$.

For the second family keep every source mean equal to $1/2$ but retain the same target alternatives. All source laws now coincide and the KL bound has no $N/(MK)$ term. The relative-bias diameter is ϵ, so the instances satisfy (12) whenever $\epsilon \le \beta$. Choose $\epsilon = \frac{1}{8}\min\{\beta, \sqrt{MK/(T+M)}\}$. The bracket is again at least $1/4$, giving $\frac{T}{32}\min\{\beta, \sqrt{MK/(T+M)}\}$. Since $T \ge 2M$, this is at least $\frac{T}{40}\min\{\beta, \sqrt{MK/T}\}$. Both finite priors are supported on the stated model, with all means in $[1/4, 3/4]$. Each Bayes lower bound is a minimax lower bound; taking their maximum, not their sum, proves (17). □

Comparing the two displayed rate expressions. Put $a = \sqrt{MK/T}$, $b = \sqrt{MK/N}$ and $\ell = \max\{\sqrt{MK/(N+T)}, \min(\beta, a)\}$, with $b = \infty$ for $N = 0$. The first term in ℓ is between $\min(a,b)/\sqrt{2}$ and $\min(a,b)$. Also $\max\{\min(a,b), \min(a,\beta)\} = \min\{a, \max(b,\beta)\}$, and $\max(b,\beta) \le b + \beta \le 2\max(b,\beta)$. Consequently

$$\ell \le \min\{a, b+\beta\} \le 2\sqrt{2}\,\ell.$$

Thus the lower bound is at least the stated core rate divided by $80\sqrt{2}$; the upper bound differs from that core rate by universal constants, its displayed logarithms and failure term. The $N = 0$ case is interpreted directly as $\ell = a$ and the core rate Ta, without multiplying zero by an infinite quantity.

For fixed M, K, the optimal-order strict-gain condition is again $N/T \to \infty$ and $\beta = o(T^{-1/2})$. Necessity follows from the lower bound: divided by $\sqrt{T}$, its two terms force $N/T \to \infty$ and $\beta\sqrt{T} \to 0$, respectively. For sufficiency, the greedy rule based on source empirical means has expected regret at most $T[\beta + \sqrt{2MK\log(2K)/N}]$: its target gap is at most β plus twice the maximum source empirical error, and the Hoeffding log-sum-exp bound controls that expected maximum. In detail, for $Z = \max_a |\widehat{\nu}_a - \nu_a|$ and $n = N/(MK) > 0$, $\mathbb{E}e^{\lambda Z} \le 2Ke^{\lambda^2/(8n)}$ for $\lambda > 0$. Jensen's inequality and minimization over λ give $\mathbb{E}Z \le \sqrt{\log(2K)/(2n)}$, yielding the displayed bound. Selecting between this rule and cellwise MOSS (Audibert & Bubeck, 2009) using their known upper bounds proves attainability. This witness does not assert that the confidence-budget implementation removes every logarithm. For that implementation, choosing $\delta = (1+N+T)^{-2}$, $N \asymp T^2$, and $\beta \le 1/T$ directly yields $O(\sqrt{\log T})$ regret, against pure-online $\Omega(\sqrt{T})$ on the same fixed-M, K class.

C Observed Regret and Bias Calibration

C.1 A trajectory certificate from observed counts

Write $R_T = \sum_{t \le T}(\mu_{X_t,*} - \mu_{X_t,A_t})$ for the realized pseudo-regret, and choose $\delta_a \in (0,1)$ separately from the confidence-interval risk. This is the action-randomization risk denoted by η in the main text. For the allocation rule, the following upper bound is computed entirely from observed quantities:

$$R_T \le \min\left\{T, \sum_x Q_x + \sqrt{\frac{T}{2}\log\frac{1}{\delta_a}}\right\} \quad \text{with probability at least } 1 - \delta - \delta_a. \tag{18}$$

Counts, successes, action probabilities, and the stated risk budgets determine Q. To verify this statement, let $G_t = \sum_a p_t(a)(\mu_{X_t,*} - \mu_{X_t,a})$ be the conditional expected gap given the pre-action information $\mathcal{G}_t$, which includes the randomized proposal. The differences $\mu_{X_t,*} - \mu_{X_t,A_t} - G_t$ are

martingale differences with conditional range length at most one. Conditional Hoeffding bounds and the exponential supermartingale give $R_T - \sum_t G_t \le \sqrt{(T/2)\log(1/\delta_a)}$ with probability at least $1-\delta_a$. On the common confidence event, $\sum_t G_t \le \sum_x Q_x$. A union bound, not independence of the events, proves (18). The minimax lower bound has a different meaning: it limits worst-case performance over a model class and is not a lower endpoint for this run.

If a valid β is not supplied, separate target calibration observations can give a conservative replacement. For each cell and action, construct simultaneous source intervals $[L_a^s, U_a^s]$ and target-calibration intervals $[L_a^c, U_a^c]$. Then

$$\beta_x^U = \min\Big\{2, \max_a(U_a^s - L_a^c) - \min_a(L_a^s - U_a^c)\Big\} \tag{19}$$

bounds the cell's bias diameter on their joint confidence event. For fixed sample counts and M cells, KL intervals with threshold $\log(4MK/\delta_c)$ for all source and calibration groups give that event probability at least $1-\delta_c$. Here $[L_a^s, U_a^s]$ uses this calibration threshold; the learning rule recomputes its source intervals with threshold $\log(4MK/\delta)$. A group without observations keeps the interval $[0,1]$, which can make the bound uninformative. If the target observations are collected sequentially, a union bound over arm–count pairs or a valid confidence sequence must replace this fixed-count calculation.

For a pilot of newly collected target observations, charge its rounds to the online budget and add at most one unit of regret per pilot observation to the remaining-horizon bound. Add its confidence risk to the total error probability as well. The following two-instance argument quantifies the target information required to calibrate an unknown diameter.

Proposition 2 (Unknown relative bias requires target information)**.** *Consider two arms, both with source mean $1/2$, and an arbitrary number of independent source observations. A procedure may collect at most $T \ge 1$ adaptively chosen Bernoulli target rewards. The target means are unknown; write $\beta(P)$ for the relative-bias diameter of instance P. Suppose an observed statistic U is a uniformly valid upper confidence bound at a fixed level $0 < \delta < 1$, so $\Pr_P\{U \ge \beta(P)\} \ge 1-\delta$ for every target instance in this model. Under the zero-bias instance P_0 with both target means $1/2$,*

$$\Pr_0\left\{U \ge \frac{1}{8\sqrt{T}}\right\} \ge \frac{3}{4} - \delta.$$

In particular, for fixed $\delta < 3/4$, no such uniformly honest bound is $o_P(T^{-1/2})$ even at this zero-bias instance.

Proof. Let $\epsilon = 1/(8\sqrt{T})$ and let P_1 have target means $(1/2+\epsilon, 1/2)$. All source laws are identical under P_0 and P_1, regardless of the source sample count. The true bias diameters are 0 under P_0 and ϵ under P_1. Include all procedure randomness in the observable transcript. Since the same procedure chooses actions from past observations in both instances, the adaptive KL chain rule gives

$$\mathrm{KL}(P_0, P_1) = \mathbb{E}_0 N_1 \,\mathrm{kl}(1/2, 1/2+\epsilon) \le 8T\epsilon^2 = \tfrac{1}{8}.$$

Here $N_1 \le T$ counts first-arm observations and $\mathrm{kl}(1/2, 1/2+\epsilon) \le 8\epsilon^2$ for $\epsilon \le 1/8$. Pinsker's inequality bounds transcript total variation by $1/4$. Uniform coverage at P_1 implies $\Pr_1\{U \ge \epsilon\} \ge 1-\delta$, hence the same event has P_0 probability at least $3/4 - \delta$. Its probability does not vanish with T, proving the last claim. This result concerns learning an unknown true diameter over a class containing both instances, not a setting in which a zero diameter is supplied as valid external information. □

D Causal Paid Calibration and Warm Starts

Fix $c \ge 1$. In each context cell, round-robin sampling obtains exactly c Bernoulli rewards per action before activation. The context process determines when these rewards arrive, not their fixed-arm sampling law. The source sample is independent of target reward streams. A union bound over the MK source and MK pilot intervals, each with two-sided failure probability at most $2\exp[-\log(4MK/\delta_c)]$, bounds the total failure probability by δ_c. Differences of their endpoints give (19). The bound is computed from past observed rewards only. Its use does not require conditioning all later confidence statements on a realized future estimator or action sequence.

The confidence event for the learning intervals covers every source coordinate and every target arm–count pair up to the total horizon. In particular, the target event includes pilot counts. This event has failure probability at most δ whether or not the envelope supplied to the rule is random. We intersect it with the calibration event; no independence between the two events is used. On the intersection, source filtering retains a target-optimal arm and the source/target gap minimum bounds all active gaps.

Before a cell's first post-pilot decision its total local visit index is $\ell = cK + 1$ and $Q = J = 0$, hence $Q^2 \le C(\ell - 1)J$. The scalar allocation induction therefore starts at that index without alteration. Replaying each observed reward once initializes counts and the numerical proposal; it does not accumulate fictitious regret certificates. The proposal may depend on the pilot and its fitted envelope because the confidence guarantee accepts every predictable proposal.

Let I_t indicate whether round t is a pilot round, which is known before its reward is observed, and let G_t be the conditional expected target gap given the pre-action information $\mathcal{G}_t$. On the joint event,

$$\sum_t G_t \le \sum_t I_t + \sum_x Q_x = B_T + \sum_x Q_x.$$

Pilot gaps lie in $[0, 1]$, including those from cells that never activate, which is why replacing B_T by zero would be invalid. The difference between realized pseudo-regret and $\sum_t G_t$ is a martingale sum of conditional range at most one per round. Hoeffding's exponential bound and a union bound prove (11).

Implementation contract. The policy interface accepts only source success/count matrices and successive triples (X_t, A_t, Y_t). It has no argument for target means, unselected rewards, a future pilot envelope, or an evaluator's optimal action. Each context cell has its own pilot history and learner state. Tests verify that a cell cannot activate early, that the pilot is replayed exactly once, that updates reject unselected actions, and that every cumulative allocation inequality holds. The ground-truth means are used only by the separate evaluator to compute regret and diagnose envelope coverage.

E Source-Conditioned IDS: Complete Specification

Relative IDS is the method identified as `filtered_width_ids` in the code and as *Width IDS + budget* in the archived experiment records. We use Relative IDS for this method throughout the paper; Width IDS denotes the IDS proposal used alone, without the cumulative guard or the source filter. The algorithm combines the confidence allocation of Section 3 with the following numerical decision model, maintaining an independent state for each context cell.

E.1 Source-conditioned prior

For each action, let (s_a^s, n_a^s) be its source success and observation counts. The target mean grid is $\mathcal{G} = \{(i + 1/2)/257 : 0 \le i < 257\}$. A declared scale b_c defines 33 equally spaced common shifts in $[-b_c, b_c]$, or only zero when $b_c = 0$. This scale is a decision-model input; the confidence guarantee uses only the relative-bias bound β. Section H gives b_c for the controlled and Hillstrom suites; in Section K, $b_c = 0.1$ in the common-shift regime and $b_c = 0$ otherwise. In the paid-calibration and real-feedback experiments, b_c for a cell is its largest estimated absolute-bias bound, $\max_a \widehat{B}_{x,a}$ (Sections F and N.2).

The residual widths are the distinct elements of $\mathcal{W} = \{0, \beta/4, \beta\}$. At width $w > 0$, the source factor is

$$L_{w,a}(m, c) = \frac{1}{w} \int_{[m+c-w/2, m+c+w/2] \cap [0,1]} v^{s_a^s} (1 - v)^{n_a^s - s_a^s} \, dv.$$

For $w = 0$, use the Bernoulli likelihood evaluated at $m + c$, set to zero when $m + c \notin [0, 1]$. The integral is evaluated with incomplete beta functions. For a fixed width and shift, normalize $L_{w,a}$ over $m \in \mathcal{G}$ to obtain $q_{0,(w,c),a}(m)$. The initial shift probabilities within that width are proportional to $\prod_a \sum_{m \in \mathcal{G}} L_{w,a}(m, c)$. Each width receives total mass $1/(2|\mathcal{W}|)$, so widths have equal initial mass after conditioning on the source; their masses are not set by comparing the source evidence across widths. The remaining mass $1/2$ is assigned to a source-independent component with uniform mean-grid probabilities for each action.

Let h index a width–shift pair or the independent component, with weight $\omega_{t,h}$ and conditional mean probabilities $q_{t,h,a}$. After the executed action a produces reward $y \in \{0, 1\}$, update

$$Z_{t,h} = \sum_{m \in \mathcal{G}} q_{t,h,a}(m) m^y (1-m)^{1-y},$$

$$\omega_{t+1,h} = \frac{\omega_{t,h} Z_{t,h}}{\sum_j \omega_{t,j} Z_{t,j}}, \qquad q_{t+1,h,a}(m) = \frac{q_{t,h,a}(m) m^y (1-m)^{1-y}}{Z_{t,h}}.$$

Other actions retain their conditional probabilities. Their marginals can change through the updated shared component weights. Updates use log probabilities for the component weights. No unobserved action reward enters any update.

E.2 The information-directed proposal

Draw 256 joint mean vectors by first drawing a component h and then drawing each action mean conditionally on that same component. Independent marginal draws would lose the common-shift dependence. For $B = \arg\max_{a \in \mathcal{A}} \mu_a$ with a fixed tie rule, let $v_b = \Pr_t(B = b)$, $m_a = \mathbb{E}_t \mu_a$ and $m_{a|b} = \mathbb{E}_t[\mu_a \mid B = b]$. The information gain of action a is

$$g_{t,a} = \sum_{b : v_b > 0} v_b \, \mathrm{kl}(m_{a|b}, m_a).$$

This follows by expanding the mutual information between the best-action index and a Bernoulli reward. Empirical averages over the joint draws estimate this quantity and the posterior gap in (5).

The implementation enumerates single actions and action pairs, minimizing the scalar information ratio on every pair. This is the two-action IDS construction of Russo & Van Roy (2018). If all estimated information values are zero, it chooses an action of minimum estimated posterior gap; under this objective, an action with zero estimated gap and zero information needs no further exploration. No artificial positive information floor is added. To save computation, the proposal distribution is recomputed only before visits $1, 1 + r, 1 + 2r, \ldots$ to its cell, for a fixed integer $r \geq 1$. At the other visits, a fresh preferred action is drawn from the most recent distribution, and every reward still updates the posterior. Except for the comparison in Section E.5, all reported experiments use $r = 8$, the implementation default, which predates that comparison. Section E.4 shows that the guarantees hold for every r, and Section E.5 compares $r = 8$ with $r \in \{1, 4, 16\}$ and explains why we keep it.

E.3 One complete round

1. Observe the context and load the state maintained for its cell. Compute target confidence endpoints and the costs $d_{t,a}, h_{t,a}$ using only previous rewards. The source-active set is fixed at initialization.
2. Refresh the IDS distribution when due and draw its preferred action b_t on the active set. Choose the exploration candidate $\tilde{a}_t$ minimizing $d_{t,a}^2 / h_{t,a}$.
3. Find the smallest $\rho \in [0, 1]$ satisfying (4) for $p = (1 - \rho)\mathbf{e}_{b_t} + \rho \, \mathbf{e}_{\tilde{a}_t}$, and draw the executed action $A_t \sim p$.
4. Accumulate the expected costs in Q, J, observe the executed action's reward, and update that action's count, success sum and posterior. Increment the cell's local clock.

The random draw of the proposal can be included in the pre-action filtration. All costs and the final action distribution are then predictable, so Proposition 1 and Theorem 2 apply directly. The frequentist guarantee therefore applies to the finite-grid and Monte Carlo proposal under a valid declared relative-bias bound, independently of whether its numerical prior is correctly specified.

The state occupies $O(HKG)$ numbers for H component states and $G = 257$ mean nodes. An executed posterior update costs $O(HG)$; an IDS refresh also draws joint samples and searches $O(K^2)$ action pairs. The scalar confidence search uses a fixed numerical tolerance. The mean grid and the rule for constructing the width set are fixed; b_c is set as described above, and $r = 8$ except in Section E.5.

E.4 Periodic proposal refreshes preserve the guarantees

Fix an integer refresh interval $r \geq 1$ before target interaction; all reported experiments except those in Section E.5 use $r = 8$. In each cell, the proposal distribution on the fixed active set is recomputed before local visits $1, 1 + r, 1 + 2r, \ldots$ from the source sample, the target feedback observed so far and learner randomness. At the other visits, the learner draws a new preferred action from the most recent distribution. The posterior is updated after every reward, and the target intervals, the costs d_t, h_t and the allocation are recomputed in every round. Under the assumptions of Theorem 1, the expected-regret bound (6) and the trajectory certificate (10) then hold for every r, with the same constants.

Proof. The distribution in use at local visit ℓ was computed before visit $\tau(\ell) = 1 + r\lfloor(\ell - 1)/r\rfloor \leq \ell$, so it depends only on the source sample, earlier target feedback and learner randomness. Include that randomness and the current preferred-action draw in $\mathcal{G}_t$. Then b_t, d_t, h_t, $\tilde{a}_t$ and the final distribution p_t are $\mathcal{G}_t$-measurable, and the final action is drawn with fresh randomness, so $A_t \mid \mathcal{G}_t \sim p_t$ and

$$\mathbb{E}[d_{t,A_t} \mid \mathcal{G}_t] = p_t^\top d_t, \qquad \mathbb{E}[h_{t,A_t} \mid \mathcal{G}_t] = p_t^\top h_t.$$

Apart from $b_t \in \mathcal{A}$, this predictability is the only property of the proposal used in Proposition 1, in the upper-bound proof of Theorem 2 and in Section C. Coverage and feasibility use the actual counts and current intervals. The two identities give $\mathbb{E}R_T \leq \mathbb{E}\sum_x Q_x + \delta T$ and $\mathbb{E}\sum_x J_x \leq MK(2 + \log T)$, and the pathwise prefix constraints then yield U_o. Because p_t is supported on the fixed active set, $\sum_x Q_x \leq U_s$ when $N > 0$. For the certificate, the realized gaps minus their $\mathcal{G}_t$-conditional means remain martingale differences with range at most one, and on the confidence event those conditional means sum to at most $\sum_x Q_x$. Each step therefore carries over unchanged, and none involves r. □

The argument uses only these two properties, so it applies to any refresh schedule decided from past information, from recomputing in every round ($r = 1$) to never recomputing after a cell's first visit. The theory therefore gives no reason to prefer one value of r over another: the bound and the form of the certificate are identical for all of them, whereas the realized actions, the value of $\sum_x Q_x$ and the regret can differ. For realized visit counts T_x, the schedule performs $\sum_x \lceil T_x/r \rceil \leq M + T/r$ proposal computations, compared with T at $r = 1$. How r affects regret is an empirical question, examined next.

E.5 Computation and regret across refresh intervals

This comparison was run after $r = 8$ had been fixed; it checks that default rather than selecting it. We compare $r \in \{1, 4, 8, 16\}$ on the five Bernoulli source regimes of Section K at $T = 2048$, with $M = K = 4$ and the same sixteen seeds $129000, \ldots, 129015$. All other settings are unchanged, including the 256 posterior draws per refresh. For each regime and seed, the four values of r share the source sample, the context sequence, the uniform draws that generate potential rewards and the learner seed, giving 320 paired trajectories. At $r = 8$, the five mean regrets agree, to the reported precision, with the $T = 2048$ results of Section K. In every run, we checked that each reward updated the posterior, that refreshes occurred exactly at the scheduled visits and that all 655,360 prefix constraints held.

Relative to refreshing in every round, $r = 8$ reduces proposal computations from 2048 to 257.69 per trajectory on average (87.4%) and process CPU time by 59–75% across regimes. Its paired mean-regret differences from $r = 1$ range from -0.945 to 0.867, and all five paired 95% intervals contain zero. Their upper limits, 1.2 to 8.6 (8–19% of the $r = 1$ means), bound the regret increase that these data leave plausible; the intervals do not establish equivalence.

The same data compare $r = 8$ with its neighbors. Refreshing twice as often, at $r = 4$, takes 28% more CPU time than $r = 8$ on average, and its paired differences from $r = 8$ range from -3.49 to 2.17, with every 95% interval containing zero; the extra computation brought no detectable improvement. Refreshing half as often, at $r = 16$, saves a further 15% of CPU time but raises mean regret in three regimes and lowers it in two. In the shared-means regime, the increase over $r = 8$ is 1.78, with paired 95% interval $[0.66, 2.90]$; this is the only one of the fifteen paired comparisons with $r = 8$ whose interval excludes zero ($p = 0.004$ unadjusted, 0.06 after a Holm correction over all fifteen). With $N = T$, the increase is larger, 6.04, but its interval $[-0.56, 12.64]$ contains zero.

Table 3: Refresh-interval comparison at $T = 2048$. Entries are mean regret $\pm$ one standard error over sixteen paired seeds; lower is better. The last row gives mean process CPU seconds per trajectory, including initialization, averaged with equal weight over the five regimes. Runs used four single-threaded workers on one machine, and the order of the four values of r was randomized within each regime–seed pair.

Source regime	$r = 1$	$r = 4$	$r = 8$	$r = 16$
Abundant, shared means	6.34 ± 0.59	5.70 ± 0.33	6.39 ± 0.41	8.18 ± 0.42
Abundant, common shift	12.94 ± 1.70	14.16 ± 1.58	11.99 ± 1.39	12.48 ± 1.42
Source size $N = T$	36.86 ± 3.94	36.01 ± 4.00	36.07 ± 4.12	42.11 ± 4.59
Boundary bound, $\beta\sqrt{T} = 2$	65.23 ± 3.45	62.02 ± 3.90	65.51 ± 3.54	63.48 ± 3.61
Fixed bound, $\beta = 0.1$	61.03 ± 2.96	64.05 ± 2.65	61.90 ± 2.76	58.85 ± 3.19
CPU seconds	1.136	0.480	0.374	0.317

We therefore keep $r = 8$. In this suite, refreshing more often brought no detectable gain, whereas refreshing half as often bought a modest further saving at the price of the only nominally significant increase in regret. This is a post hoc check of one default, not an optimality result. No other values of r were tested, and the comparison covers one horizon, one gap size, sixteen seeds per regime and one machine; it does not cover the response banks of Section 6. With a larger r, each cell also uses its first proposal, computed from the source data alone, for its first r visits, so infrequent refreshes may cost more with short horizons or large gaps, particularly when the source is uninformative. Because the guarantees hold for every r (Section E.4), r can be adapted to a computational budget without affecting validity.

F Paid-Calibration Experiment Protocol

Information and feedback. The learner receives source successes and counts, the horizon, a context identifier at each round, and the reward of its chosen action. It never receives target means, unchosen rewards or the future context sequence. The evaluator maintains those quantities separately. Source and target draws use independent random-number streams, and all methods in a paired replicate share the same source statistics and potential feedback. Sharing potential feedback reduces comparison noise; it does not expose unchosen feedback to any policy. Environment streams are seeded by the replicate and the first four bytes of the environment name. The two common-shift names both start with `comm`, so they share target context and feedback streams and are not independent replications. Other controlled environments use different streams despite having the same target means. Replaying all 32 seed pairs also showed that Anchor-TS's bias-corrected hybrid draw did not change any post-pilot action from the choice given by the maximum of its online empirical mean and online Gaussian draw. With shared target and policy randomness, the trajectories therefore coincide. This is a property of these runs, not a general common-shift invariance of Anchor-TS.

Environments. We use fixed response banks from Hillstrom, Criteo, IST, KuaiRand and OBD (all, men and women campaigns). Conditional on each bank, a cell–action reward is Bernoulli with its observed response frequency; contexts follow the fixed empirical cell distribution. The experiment thus resamples real recorded outcomes; it is neither a prospective deployment nor a counterfactual estimate for the original population. The source sample has 640 draws per supported cell–action pair. Its binomial sufficient statistics are equivalent to independent resampling from the corresponding binary source bank. The target horizon is 8,000 rounds, except for OBD, whose larger action sets use 64,000. Section N details the raw-data construction and cell semantics.

Five additional four-cell, four-action environments isolate the source mechanism. They share the target means below, with one row per cell:

$$\begin{pmatrix} .30 & .34 & .38 & .42 \\ .45 & .39 & .41 & .37 \\ .29 & .32 & .35 & .38 \\ .43 & .40 & .37 & .34 \end{pmatrix}.$$

The matched source equals this matrix; common shifts add $+.25$ or $-.20$. The residual source adds .20 and the action vector $(.03, -.02, .01, -.02)$ to each row. The misleading source reverses each target row. All means stay in $[0, 1]$. These settings distinguish a harmless common shift from action-order contamination.

Development and frozen validation. Development uses eight seeds (151000–151007) on Hillstrom, Criteo, IST and the five controlled environments; OBD and KuaiRand do not participate in pilot selection. Each paid method, including each comparator, selects its own pilot count from $c \in \{4, 8, 16, 32, 64, 128\}$ by the same rule. The criterion is the mean, across environments, of the ratio of the method's mean regret to that of Online IDS; among configurations within 1% of the minimum, the smallest pilot is chosen. The resulting counts are 8 for Relative IDS, Width IDS, Relative greedy and MIN-UCB; 4 for Relative TS and Width TS; and 16 for empirical-mean Anchor-TS and Width greedy. Methods without calibration start acting immediately. Relative IDS with 128 pilot observations per cell–action is retained as a cost ablation. Validation fixes these choices and uses 32 new seeds (152000–152031) in all twelve environments, comprising 6,144 learning trajectories.

What is charged. Pilots are round-robin within each actually arriving context. Every pilot reward consumes one target round and contributes its actual regret. A cell activates after cK observed rewards; unfinished pilots still contribute cost. Completed pilot rewards initialize the online statistics exactly once. The calibration failure budget is $\delta_c = .01$; the subsequent confidence budget is $.04$, divided across supported cells. At activation, the calibration intervals give β_x^U and the absolute-bias envelope $\widehat{B}_{x,a} = \max\{|L^s_{x,a} - U^c_{x,a}|, |U^s_{x,a} - L^c_{x,a}|\}$. Pure-online and direct-source controls do not pay an artificial calibration tax.

Comparators and ablations. Online IDS, LinMED, Exp-KL-MS and Online Beta-TS test target-only learning. Source IDS, Source TS and Source greedy test direct reuse under the same source observations. MIN-UCB and empirical-mean Anchor-TS receive the same estimated absolute-bias envelope; their pilots are independently tuned. Relative TS and Relative greedy keep the guard but replace the IDS proposal. Width variants remove both the cumulative confidence allocation and the source filter. At $c = 8$, the source filter of Relative IDS kept every action in all 1,247 activated cells of the 384 validation replicates (12 environments $\times$ 32 seeds), so the matched Relative IDS–Width IDS pair isolates the guard in these runs; separately tuned TS and greedy variants serve as performance comparisons. The separate shared-model experiment in Section M includes OOPE and E2D.Off under their required finite-class interfaces.

Numerical verification. The scalar search evaluates the prefix constraint using exactly the dot products used by the state update. Bisection remains on the feasible side, seeking a machine-precision interior margin when the endpoint allows one. Assertion tolerances scale with the magnitude of the squared-prefix constraint. Long-prefix stress tests, feedback-interface tests and independent checks of the true bias against the estimated envelope accompany the code. All validation trajectories were rerun after boundary-rounding errors in the scalar search were corrected by the interior margin described above; the earlier numerical failures remain documented in the research provenance. The distinction between exact feasibility and numerical residuals is stated in Section A.

Uncertainty and scope. Tables report replicate means and standard errors; paired differences use the shared replicate identifier. For OBD macro averages, the three campaigns are first averaged within each seed. These uncertainties describe the declared resampling experiment conditional on each fixed response bank. They do not quantify population identification error, and a frozen-bank regret comparison is not an observed online treatment effect.

For Fig. 5, each bank's Online IDS mean is a fixed normalization denominator; its estimation uncertainty is not propagated into the displayed intervals. The Bonferroni family contains ten Relative IDS total-regret means (five banks, two pilot sizes) and five paired Relative IDS minus Width IDS differences at $c = 8$. Development selects $c = 8$ for Relative IDS; $c = 128$ is a prespecified cost comparison. Calibration is completed separately in each context. The post-pilot component sums regret over the rounds remaining after these pilots, rather than comparing methods at a common post-pilot horizon. On Hillstrom, raising the pilot from $c = 8$ to $c = 128$ adds 49.38 to pilot regret and lowers post-pilot regret by 23.01, so mean total regret rises by 26.37; on the OBD macro average

the corresponding changes are 224.75, 186.70 and 38.06. Mean total regret also rises on Criteo but falls on IST and KuaiRand.

G Complete Paid-Calibration Validation

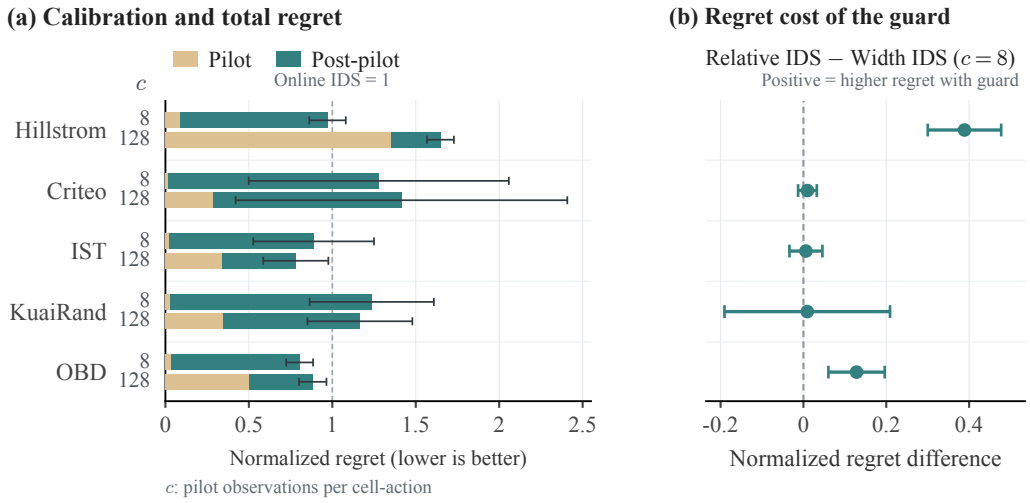


Figure 5: Target cost of calibrated transfer. Each bank's regret is normalized by its fixed Online IDS mean (dashed line in (a)). (a) Relative IDS pilot and post-pilot regret by pilot count c; black intervals cover the total. (b) Paired Relative IDS minus Width IDS regret at $c = 8$, where Width IDS omits the cumulative guard: positive values indicate a guard cost. Bonferroni-adjusted t intervals give approximate joint 95% coverage of ten totals and five differences over 32 held-out seeds.

Table 4: Regret on the response banks: means ± standard errors over 32 paired seeds, with the lowest mean in each column in bold. All target pilot rounds are charged to regret. Table 5 lists all methods and environments.

Method	Hillstrom	Criteo	IST	KuaiRand	OBD
Target only					
Online IDS	39.0 ± 3.2	8.9 ± 1.3	36.8 ± 6.4	70.6 ± 7.1	484.8 ± 3.8
LinMED	89.3 ± 9.4	28.4 ± 4.5	50.5 ± 7.4	**58.1 ± 9.4**	431.4 ± 14.4
Direct source reuse					
Source IDS	**2.4 ± 0.6**	**6.0 ± 1.4**	**13.1 ± 2.3**	171.4 ± 1.2	**310.6 ± 17.1**
Target-calibrated transfer					
MIN-UCB ($c = 8$)	132.5 ± 3.2	30.3 ± 0.8	59.3 ± 2.0	143.5 ± 1.9	888.1 ± 0.3
Width IDS ($c = 8$)	22.7 ± 1.8	11.2 ± 2.2	32.4 ± 4.4	86.7 ± 8.1	327.7 ± 15.0
Relative IDS ($c = 8$)	37.8 ± 1.3	11.3 ± 2.2	32.7 ± 4.2	87.3 ± 8.3	389.9 ± 12.2
Relative IDS ($c = 128$)	64.2 ± 1.0	12.5 ± 2.8	28.7 ± 2.3	82.3 ± 7.0	427.9 ± 12.5

Guard and pilot costs. At $c = 8$, the mean regret of Relative IDS minus that of Width IDS, which isolates the guard in these runs, is 15.15 on Hillstrom and 62.18 on the OBD macro average, with simultaneous intervals $[11.70, 18.60]$ and $[29.22, 95.15]$; on Criteo, IST and KuaiRand it is at most 0.65 and the intervals include zero. In the controlled environments it ranges from 39.04 to 49.65, except under the misleading source, where it is -14.93. Among response banks other than OBD, Relative IDS's pilot regret is at most 3.29. On Hillstrom, Criteo and IST, the source and target best actions coincide, favoring direct source reuse. On KuaiRand the best actions agree on only about 0.575% of the context weight, and Relative IDS has mean regret 87.33 versus 171.38 for Source IDS.

Table 5 retains all frozen configurations and all environments. Pilot/arm is zero for methods that do not calibrate; the reported pilot rounds include unfinished cell pilots. OBD macro rows summarize three campaigns within each seed, and campaign-specific rows follow separately.

Table 5: Paid-calibration results by environment. Regret is reported as mean ± standard error; pilot rounds are included in regret. Pilot/arm is the pilot count c per cell–action; the `obd` rows are macro averages over the three OBD campaigns, which follow as `obd all`, `obd men` and `obd women`.

Environment	Method	Pilot/arm	Regret	Pilot rounds
hillstrom	Anchor-TS	16	60.7 ± 4.0	144
hillstrom	Online Beta-TS	0	49.6 ± 2.2	0
hillstrom	Exp-KL-MS	0	64.6 ± 2.8	0
hillstrom	Source TS	0	8.6 ± 1.2	0
hillstrom	LinMED	0	89.3 ± 9.4	0
hillstrom	MIN-UCB	8	132.5 ± 3.2	72
hillstrom	Online IDS	0	39.0 ± 3.2	0
hillstrom	Relative greedy	8	41.8 ± 1.5	72
hillstrom	Relative IDS	8	37.8 ± 1.3	72
hillstrom	Relative TS	4	37.4 ± 1.3	36
hillstrom	Source greedy	0	6.2 ± 3.0	0
hillstrom	Source IDS	0	2.4 ± 0.6	0
hillstrom	Width greedy	16	20.7 ± 7.0	144
hillstrom	Width IDS	8	22.7 ± 1.8	72
hillstrom	Width TS	4	33.4 ± 1.6	36
hillstrom	Relative IDS	128	64.2 ± 1.0	1152
criteo	Anchor-TS	16	21.4 ± 2.0	64
criteo	Online Beta-TS	0	18.4 ± 2.1	0
criteo	Exp-KL-MS	0	22.2 ± 2.1	0
criteo	Source TS	0	10.4 ± 1.7	0
criteo	LinMED	0	28.4 ± 4.5	0
criteo	MIN-UCB	8	30.3 ± 0.8	32
criteo	Online IDS	0	8.9 ± 1.3	0
criteo	Relative greedy	8	22.2 ± 4.8	32
criteo	Relative IDS	8	11.3 ± 2.2	32
criteo	Relative TS	4	20.1 ± 2.7	16
criteo	Source greedy	0	9.4 ± 3.8	0
criteo	Source IDS	0	6.0 ± 1.4	0
criteo	Width greedy	16	12.4 ± 4.5	64
criteo	Width IDS	8	11.2 ± 2.2	32
criteo	Width TS	4	20.1 ± 2.7	16
criteo	Relative IDS	128	12.5 ± 2.8	512
ist	Anchor-TS	16	39.3 ± 3.4	128
ist	Online Beta-TS	0	42.4 ± 3.5	0
ist	Exp-KL-MS	0	50.4 ± 1.9	0
ist	Source TS	0	29.4 ± 3.0	0
ist	LinMED	0	50.5 ± 7.4	0
ist	MIN-UCB	8	59.3 ± 2.0	64
ist	Online IDS	0	36.8 ± 6.4	0
ist	Relative greedy	8	44.6 ± 6.7	64
ist	Relative IDS	8	32.7 ± 4.2	64
ist	Relative TS	4	33.1 ± 2.5	32
ist	Source greedy	0	26.8 ± 7.7	0
ist	Source IDS	0	13.1 ± 2.3	0
ist	Width greedy	16	60.9 ± 10.8	128
ist	Width IDS	8	32.4 ± 4.4	64
ist	Width TS	4	33.0 ± 2.5	32
ist	Relative IDS	128	28.7 ± 2.3	1024

Continued on next page

Table 5: (continued)

Environment	Method	Pilot/arm	Regret	Pilot rounds
kuairand	Anchor-TS	16	96.7 ± 4.1	216
kuairand	Online Beta-TS	0	86.0 ± 6.1	0
kuairand	Exp-KL-MS	0	105.5 ± 4.9	0
kuairand	Source TS	0	172.6 ± 1.1	0
kuairand	LinMED	0	58.1 ± 9.4	0
kuairand	MIN-UCB	8	143.5 ± 1.9	128
kuairand	Online IDS	0	70.6 ± 7.1	0
kuairand	Relative greedy	8	106.6 ± 10.6	128
kuairand	Relative IDS	8	87.3 ± 8.3	128
kuairand	Relative TS	4	93.4 ± 5.7	64
kuairand	Source greedy	0	232.4 ± 4.0	0
kuairand	Source IDS	0	171.4 ± 1.2	0
kuairand	Width greedy	16	133.8 ± 10.4	216
kuairand	Width IDS	8	86.7 ± 8.1	128
kuairand	Width TS	4	92.3 ± 5.4	64
kuairand	Relative IDS	128	82.3 ± 7.0	1112
obd	Anchor-TS	16	861.5 ± 0.7	1859
obd	Online Beta-TS	0	584.4 ± 4.1	0
obd	Exp-KL-MS	0	584.6 ± 3.2	0
obd	Source TS	0	355.9 ± 14.0	0
obd	LinMED	0	431.4 ± 14.4	0
obd	MIN-UCB	8	888.1 ± 0.3	1006
obd	Online IDS	0	484.8 ± 3.8	0
obd	Relative greedy	8	441.1 ± 16.2	1006
obd	Relative IDS	8	389.9 ± 12.2	1006
obd	Relative TS	4	449.4 ± 12.2	574
obd	Source greedy	0	815.3 ± 24.9	0
obd	Source IDS	0	310.6 ± 17.1	0
obd	Width greedy	16	398.5 ± 20.4	1859
obd	Width IDS	8	327.7 ± 15.0	1006
obd	Width TS	4	402.6 ± 14.3	574
obd	Relative IDS	128	427.9 ± 12.5	12473
obd all	Anchor-TS	16	1561.2 ± 1.8	2875
obd all	Online Beta-TS	0	1033.7 ± 10.0	0
obd all	Exp-KL-MS	0	1021.5 ± 7.5	0
obd all	Source TS	0	474.8 ± 44.4	0
obd all	LinMED	0	800.5 ± 36.8	0
obd all	MIN-UCB	8	1612.1 ± 0.8	1595
obd all	Online IDS	0	848.5 ± 9.0	0
obd all	Relative greedy	8	865.3 ± 45.4	1595
obd all	Relative IDS	8	727.5 ± 35.5	1595
obd all	Relative TS	4	701.5 ± 37.3	949
obd all	Source greedy	0	1438.4 ± 55.6	0
obd all	Source IDS	0	547.8 ± 48.6	0
obd all	Width greedy	16	784.5 ± 66.7	2875
obd all	Width IDS	8	577.0 ± 48.9	1595
obd all	Width TS	4	574.9 ± 45.2	949
obd all	Relative IDS	128	790.7 ± 29.5	20795
obd men	Anchor-TS	16	525.9 ± 0.7	1230
obd men	Online Beta-TS	0	343.7 ± 5.6	0
obd men	Exp-KL-MS	0	359.0 ± 4.6	0
obd men	Source TS	0	302.0 ± 11.2	0
obd men	LinMED	0	270.8 ± 16.0	0
obd men	MIN-UCB	8	543.4 ± 0.3	686

Continued on next page

Table 5: (continued)

Environment	Method	Pilot/arm	Regret	Pilot rounds
obd men	Online IDS	0	292.6 ± 5.7	0
obd men	Relative greedy	8	252.8 ± 16.0	686
obd men	Relative IDS	8	245.1 ± 14.8	686
obd men	Relative TS	4	324.4 ± 11.7	405
obd men	Source greedy	0	585.3 ± 33.7	0
obd men	Source IDS	0	192.2 ± 9.1	0
obd men	Width greedy	16	246.3 ± 16.9	1230
obd men	Width IDS	8	215.6 ± 14.8	686
obd men	Width TS	4	316.2 ± 11.5	405
obd men	Relative IDS	128	250.5 ± 11.1	8846
obd women	Anchor-TS	16	497.2 ± 0.5	1472
obd women	Online Beta-TS	0	375.7 ± 5.9	0
obd women	Exp-KL-MS	0	373.2 ± 4.4	0
obd women	Source TS	0	290.9 ± 9.6	0
obd women	LinMED	0	223.0 ± 16.0	0
obd women	MIN-UCB	8	508.7 ± 0.2	736
obd women	Online IDS	0	313.4 ± 7.3	0
obd women	Relative greedy	8	205.1 ± 18.5	736
obd women	Relative IDS	8	197.1 ± 12.3	736
obd women	Relative TS	4	322.2 ± 10.8	368
obd women	Source greedy	0	422.1 ± 31.9	0
obd women	Source IDS	0	191.6 ± 10.4	0
obd women	Width greedy	16	164.5 ± 21.5	1472
obd women	Width IDS	8	190.5 ± 13.3	736
obd women	Width TS	4	316.8 ± 11.2	368
obd women	Relative IDS	128	242.7 ± 14.8	7777
matched	Anchor-TS	16	167.7 ± 6.3	256
matched	Online Beta-TS	0	151.2 ± 7.8	0
matched	Exp-KL-MS	0	205.7 ± 6.3	0
matched	Source TS	0	46.6 ± 4.8	0
matched	LinMED	0	184.8 ± 12.0	0
matched	MIN-UCB	8	242.5 ± 4.1	128
matched	Online IDS	0	135.4 ± 8.2	0
matched	Relative greedy	8	122.2 ± 6.2	128
matched	Relative IDS	8	114.7 ± 5.4	128
matched	Relative TS	4	109.9 ± 4.8	64
matched	Source greedy	0	21.2 ± 8.6	0
matched	Source IDS	0	19.5 ± 4.9	0
matched	Width greedy	16	97.2 ± 14.7	256
matched	Width IDS	8	68.3 ± 5.7	128
matched	Width TS	4	96.9 ± 5.0	64
matched	Relative IDS	128	152.2 ± 4.2	2048
common positive	Anchor-TS	16	152.0 ± 7.8	256
common positive	Online Beta-TS	0	160.6 ± 6.8	0
common positive	Exp-KL-MS	0	185.7 ± 5.6	0
common positive	Source TS	0	248.4 ± 1.7	0
common positive	LinMED	0	198.8 ± 12.0	0
common positive	MIN-UCB	8	244.0 ± 4.1	128
common positive	Online IDS	0	136.2 ± 8.6	0
common positive	Relative greedy	8	122.3 ± 5.9	128
common positive	Relative IDS	8	109.4 ± 4.5	128
common positive	Relative TS	4	103.7 ± 4.5	64
common positive	Source greedy	0	20.5 ± 5.8	0
common positive	Source IDS	0	235.5 ± 2.3	0

Continued on next page

Table 5: (continued)

Environment	Method	Pilot/arm	Regret	Pilot rounds
common positive	Width greedy	16	88.0 ± 9.1	256
common positive	Width IDS	8	70.3 ± 7.0	128
common positive	Width TS	4	92.3 ± 4.8	64
common positive	Relative IDS	128	149.1 ± 3.6	2048
common negative	Anchor-TS	16	152.0 ± 7.8	256
common negative	Online Beta-TS	0	160.6 ± 6.8	0
common negative	Exp-KL-MS	0	185.7 ± 5.6	0
common negative	Source TS	0	14.0 ± 4.4	0
common negative	LinMED	0	198.8 ± 12.0	0
common negative	MIN-UCB	8	245.0 ± 4.0	128
common negative	Online IDS	0	136.2 ± 8.6	0
common negative	Relative greedy	8	120.6 ± 6.2	128
common negative	Relative IDS	8	114.1 ± 5.5	128
common negative	Relative TS	4	99.9 ± 3.8	64
common negative	Source greedy	0	19.2 ± 6.4	0
common negative	Source IDS	0	16.8 ± 6.1	0
common negative	Width greedy	16	89.6 ± 10.8	256
common negative	Width IDS	8	71.3 ± 7.2	128
common negative	Width TS	4	88.1 ± 4.2	64
common negative	Relative IDS	128	149.3 ± 3.4	2048
residual	Anchor-TS	16	156.1 ± 7.8	256
residual	Online Beta-TS	0	157.1 ± 8.1	0
residual	Exp-KL-MS	0	206.8 ± 5.4	0
residual	Source TS	0	223.5 ± 2.4	0
residual	LinMED	0	181.5 ± 12.1	0
residual	MIN-UCB	8	240.9 ± 3.2	128
residual	Online IDS	0	133.7 ± 9.2	0
residual	Relative greedy	8	129.3 ± 5.6	128
residual	Relative IDS	8	124.7 ± 4.8	128
residual	Relative TS	4	114.5 ± 5.1	64
residual	Source greedy	0	58.8 ± 8.8	0
residual	Source IDS	0	206.5 ± 2.7	0
residual	Width greedy	16	92.0 ± 13.8	256
residual	Width IDS	8	75.1 ± 4.6	128
residual	Width TS	4	104.3 ± 5.5	64
residual	Relative IDS	128	155.1 ± 4.8	2048
misleading	Anchor-TS	16	145.5 ± 6.7	256
misleading	Online Beta-TS	0	154.6 ± 5.9	0
misleading	Exp-KL-MS	0	198.8 ± 6.5	0
misleading	Source TS	0	559.3 ± 6.6	0
misleading	LinMED	0	159.9 ± 12.1	0
misleading	MIN-UCB	8	239.7 ± 3.2	128
misleading	Online IDS	0	132.1 ± 9.5	0
misleading	Relative greedy	8	223.8 ± 11.5	128
misleading	Relative IDS	8	177.2 ± 7.3	128
misleading	Relative TS	4	225.1 ± 9.7	64
misleading	Source greedy	0	720.2 ± 8.8	0
misleading	Source IDS	0	560.0 ± 9.5	0
misleading	Width greedy	16	367.2 ± 27.3	256
misleading	Width IDS	8	192.1 ± 10.9	128
misleading	Width TS	4	225.6 ± 9.9	64
misleading	Relative IDS	128	198.1 ± 6.6	2048

H Experimental Design and Complete Comparisons

H.1 Source information and the rate boundary

The horizon experiment uses $M = K = 4$ and $T \in \{512, 2048, 8192, 32768\}$, with target gap $2/\sqrt{T}$. Two regimes, shared means and a common $+0.1$ source shift, have $N = T^2/(MK)$ and $\beta = 0$. The remaining three regimes either use $N = T$ or set all source means to $1/2$ while supplying a valid relative-bias bound for which $\beta\sqrt{T}$ does not vanish (the boundary-bound regime, $\beta\sqrt{T} = 2$, and the fixed-bound regime, $\beta = 0.1$). The full source contract and complete tables for all thirteen methods are in Section K.

With shared means, Relative IDS reaches zero observed regret from $T = 8192$ onward, whereas Online IDS mean regret grows to 192.78 at $T = 32768$; the advantage persists under a common shift. The direct-source methods Source greedy and Source IDS also reach zero observed regret with abundant aligned data, so the experiment tests the *information boundary*, not unique algorithmic dominance. A source-filtered KL-UCB control, run with separate seeds, recovers the boundary without IDS (Section L). When $N = T$ or the source does not identify the best action, normalized regret stays on the online scale. These finite-horizon tests probe the mechanisms; worst-case necessity follows from the lower bound.

H.2 Strong baselines under explicit information contracts

Comparators include Online IDS, LinMED (Balagopalan & Jun, 2025), Exp-KL-MS (Qin et al., 2026), Thompson sampling (TS), MIN-UCB and Anchor-TS (Li et al., 2026). Offline-Online Phased Elimination (OOPE) (Vijayan et al., 2026) and E2D.Off, the Estimation to Decisions meta-algorithm with offline oracles based on Oracle-Efficient Online Estimation (OEOE) (Foster et al., 2024), are evaluated only in a compatible shared-model suite, with finite-class TS controls given E2D.Off's model information.

At 2,048 source observations per cell–action in the shared-model suite, Relative IDS has mean regret 20.44, below Online IDS (135.61) and empirical-mean Anchor-TS (55.88), whereas Source IDS attains 14.20. These results support the benefit of transfer in this aligned setting, but do not establish an advantage over direct source reuse. The rest of this appendix and Section M give implementation details, all 25 implementations of the shared-model suite, source budgets and paired uncertainty estimates; configurations are fixed across test settings.

H.3 Controlled Bernoulli instances

The target mean of action a in cell $x \in \{0, 1, 2, 3\}$ is $0.35 + 0.02x + \Delta\mathbf{1}\{a = a_x^*\}$, where the best action a_x^* is drawn independently and uniformly in each cell. We cross gaps $\Delta \in \{0.02, 0.08, 0.2\}$ and source counts per cell–action $n^s \in \{128, 2048, 32768\}$ with five source mechanisms:

1. Shared means: $\nu_{x,a} = \mu_{x,a}$ and $\beta = 0$.
2. Positive common shift: $\nu_{x,a} = \mu_{x,a} + 0.1$ and $\beta = 0$.
3. Negative common shift: $\nu_{x,a} = \mu_{x,a} - 0.1$ and $\beta = 0$.
4. Bounded residuals: $\nu_{x,a} = \mu_{x,a} + 0.1 + e_{x,a}$, with independent $e_{x,a} \sim \mathrm{Unif}[-0.04, 0.04]$ and supplied $\beta = 0.08$.
5. Misleading but valid source ("changed rankings" in Section M): add $0.1 - 0.15$ to the target-best action and $0.1 + 0.15$ to every other action, with supplied $\beta = 0.3$.

All means lie in $[1/4, 3/4]$. Contexts are drawn independently and uniformly from $M = 4$ cells, with $K = 4$ actions and $T = 12000$ target rounds. Source counts are balanced, so the total budget is $N = 16n^s$. The declared common-shift scale is zero in the shared-mean regime and 0.1 otherwise; it is an additional input to the decision model and is not inferred from unobserved target means.

The 45 settings each use seeds 125000–125031. Methods share context and per-round, per-action reward random numbers within a setting and seed, but update from their executed actions only. Reusing a seed across settings does not yield a new independent observation. Method selection and numerical parameters were fixed before these seeds were run, although earlier development

results were available at that time. Setting-wise paired-t intervals use Bonferroni correction over $45\binom{25}{2}$ method pairs within settings, the enlarged family of Section M, which contains all original pairs. We do not reduce this correction when displaying fewer methods. These are approximate simulation intervals, not distribution-free confidence guarantees. The complete paired comparison rows accompany the reproducibility materials.

H.4 Baselines and information supplied

Online IDS and Source IDS use independent Beta$(1, 1)$ priors on the action means. Source IDS includes all source successes and failures as prior updates. Both use the same 256-draw IDS calculation and eight-visit refresh schedule as Relative IDS. Online Beta-TS and Source TS use the corresponding priors without the information-ratio objective.

For each action independently, the ϵ-TS controls use a posterior sample with probability ϵ and the posterior mean otherwise. We report $\epsilon = 1/K$ and 0.75, for both online and source-assisted versions. LinMED (Balagopalan & Jun, 2025) uses Bernoulli noise scale 0.5, $\alpha_{\text{emp}} = 0.9$ and $\alpha_{\text{opt}} = 0.01$. Its empirical weight α_{emp} was selected from $\{0.5, 0.75, 0.9, 0.98\}$ on separate shared-mean training seeds and is not reselected for each test regime or for the email dataset. Exp-KL-MS (Qin et al., 2026) uses its Bernoulli form.

MIN-UCB receives a valid coordinatewise source-bias envelope. Directional MIN-UCB uses the declared ellipsoid that contains the same common-shift and residual box, with the published one-hot layered confidence construction (Han et al., 2026). It is the known-certificate variant, not a learned-certificate result with free knowledge of the bias. Relative IDS receives β and its numerical shift scale. Online-only controls receive neither source observations nor source structure. Source greedy uses only the source sample means.

An objective ablation uses the same width-mixture posterior, active set and confidence budget as Relative IDS, replacing IDS by $\epsilon = 1/K$ sampling. Its point estimates and all comparisons are reported alongside the baselines. The ablation isolates the decision objective; it is not a second proposed algorithm.

The shared-model suite adds Anchor-TS, OOPE and OEOE–E2D.Off; Section M specifies the realizable class, source contract, local horizons and implementation variants. Anchor-TS is also run in every world of the biased-source and Hillstrom suites. Updated setting-wise intervals retain all original methods and account for the enlarged comparison family.

Scope of the comparison. HybElimUCB-RA requires offline pairwise comparison outcomes, which these per-action reward datasets do not supply. For one-hot actions, Hybrid LinUCB's global Euclidean certificate implies the coordinatewise bounds used by MIN-UCB; we include the latter as the more permissive absolute-bias control. OOPE and E2D.Off are compared only within their compatible shared-model suite. Source misspecification outside a method's model is not counted as an in-model improvement. The theoretical information requirements are summarized in Table 1. Section M gives the complete setting-level tables, including the original controls, the objective ablation and the additional applicable baselines.

H.5 Complete-grid performance with supplied information

Table 2 compares Relative IDS with all 19 external baseline configurations of the supplied-information suite in Section H.3, and Table 6 compares it with its three internal variants. Together the two tables retain all 23 current implementations and every setting. Both are post hoc aggregations of the archived, corrected runs, not a new validation batch. Each source-budget column weights five mechanisms and three gaps equally; the overall column weights all 45 settings equally. We first average settings within each seed and then compute means and standard errors over the 32 seeds 125000–125031; because settings share seeds, the seeds are the independent replicates. Online controls do not use source data, so their budget columns are identical. Each method receives the information in Section H.4; no target rounds are spent on calibration.

For a comparator and seed s, let Z_s be the comparator's setting-averaged regret minus that of Relative IDS. We report $\bar{Z} \pm t_{31,1-0.05/(2F)}\,\mathrm{SE}(Z)$ with $F = (5+1)(3+1)(3+1)\binom{25}{2} = 28800$: each of the three factors (mechanism, gap and source budget) is either fixed at one level or averaged over its

levels, giving 96 margins, and each margin has $\binom{25}{2}$ configuration pairs. The family keeps the original 25 configuration slots, although the corrected archive has 23 configurations and 24288 evaluated pairs across these margins. It thus covers the displayed overall comparisons and every other fixed or averaged margin, and it is larger than the setting-wise family of Section H.3. The intervals remain approximate simulation intervals.

The overall mean of Relative IDS is 59.87, versus 108.34 for the strongest external configuration, source-assisted empirical Anchor-TS with $\sigma = 0.5$: a 44.7% reduction. For all 19 external configurations, the interval in Table 8 lies above zero. The intervals for the three internal variants contain zero; in particular, the overall means of the two IDS priors, 59.87 and 59.90, do not resolve a difference. These are equal-weight grid averages, not per-mechanism results. By mechanism, Relative IDS has the lowest mean among the methods of Table 2 under the positive common shift and bounded residuals. Table 7 gives all five mechanisms; there, direct source reuse outside its shared-mean model illustrates misspecification behavior, not a violation of its in-model guarantees. Under changed rankings, Relative IDS has mean regret 137.04, compared with 930.87 for Source IDS and 125.44 for Online IDS; the difference from Online IDS, 11.60, has simultaneous interval $[-18.44, 41.64]$.

As a descriptive scale check, we divide each setting's regret by its fixed Online IDS mean before equal-weight averaging. Relative IDS remains lowest, at 0.458 versus 0.860 for the strongest external configuration. This normalization is not used for the paired intervals.

Internal variants. The three variants change the proposal objective or the source-conditioned prior within our relative-confidence framework. They are compared with Relative IDS as a group, separately from the external baselines, and are not selected by source budget. The structure prior gives equal initial weight to its components: exact shared means, an independent target model, a common-shift model when the supplied shift scale is positive, and a residual-bias model when $\beta > 0$. The last two use the finite grid of Section E; executed target rewards update all components and their weights. The IDS variant with a structure prior shares the IDS objective and the relative-confidence allocation with Relative IDS; its numerical prior construction differs. At 32,768 source observations per cell–action, its mean regret is 35.40 versus 38.35 for Relative IDS. The paired difference is -2.95, with approximate simultaneous interval $[-13.31, 7.41]$ in the same family. Most of the observed difference comes from the changed-rankings regime (125.93 versus 137.49), and these runs do not establish a general advantage for that prior.

Parameter scans. Figure 4 evaluates nine common offsets from -0.1 to 0.1 in increments of 0.025, nine relative-bias bounds from 0 to 0.08 in increments of 0.01, and nine source counts per cell–action $n \in \{128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768\}$. The offset and bias scans average the three gaps and three source budgets used in the tables; the source-count scan averages all five mechanisms and three gaps. All scans use the same 32 seeds and fixed algorithm parameters. Panel (d) reports the 45-setting aggregate in the tables.

Table 6: Relative IDS and its three internal variants under the same supplied information; Relative IDS uses the IDS proposal with the width-mixture prior of Section E. Columns and seed-level aggregation match Table 2. Entries are means ± standard errors; bold marks the lowest mean in each column of this table.

Method	128	2,048	32,768	All 45 settings
ϵ proposal, width prior	91.91 ± 2.48	53.47 ± 1.80	39.63 ± 1.41	61.67 ± 1.57
ϵ proposal, structure prior	100.45 ± 3.14	58.80 ± 2.25	45.82 ± 1.85	68.36 ± 1.99
IDS proposal, structure prior	92.47 ± 2.28	51.83 ± 1.89	**35.40 ± 1.45**	59.90 ± 1.47
Relative IDS	**90.69 ± 2.27**	**50.56 ± 1.93**	38.35 ± 1.38	**59.87 ± 1.26**

Table 7: Regret of all 23 implementations by source mechanism, averaged over the three gaps and three source budgets within each of 32 seeds. Entries are means ± standard errors; bold marks the lowest mean in each column.

Method	Shared means	Common $+0.1$	Common -0.1	Bounded residuals	Changed rankings
Online controls					
Online IDS	125.44 ± 4.90	125.44 ± 4.90	125.44 ± 4.90	125.44 ± 4.90	$\mathbf{125.44 \pm 4.90}$
Online Beta-TS	154.77 ± 4.17	154.77 ± 4.17	154.77 ± 4.17	154.77 ± 4.17	154.77 ± 4.17
Online ϵ-TS $(1/K)$	152.91 ± 11.03	152.91 ± 11.03	152.91 ± 11.03	152.91 ± 11.03	152.91 ± 11.03
Online ϵ-TS (0.75)	143.18 ± 5.93	143.18 ± 5.93	143.18 ± 5.93	143.18 ± 5.93	143.18 ± 5.93
Exp-KL-MS	203.46 ± 3.98	203.46 ± 3.98	203.46 ± 3.98	203.46 ± 3.98	203.46 ± 3.98
LinMED	205.77 ± 8.75	205.77 ± 8.75	205.77 ± 8.75	205.77 ± 8.75	205.77 ± 8.75
Anchor-TS, empirical ($\sigma = 0.5$)	146.95 ± 12.05	146.95 ± 12.05	146.95 ± 12.05	146.95 ± 12.05	146.95 ± 12.05
Anchor-TS, empirical ($\sigma = 1$)	162.17 ± 6.24	162.17 ± 6.24	162.17 ± 6.24	162.17 ± 6.24	162.17 ± 6.24
Anchor-TS, posterior ($\sigma = 1$)	162.81 ± 6.03	162.81 ± 6.03	162.81 ± 6.03	162.81 ± 6.03	162.81 ± 6.03
Direct source reuse					
Source IDS	$\mathbf{26.40 \pm 1.40}$	51.30 ± 1.17	35.28 ± 3.51	70.01 ± 2.92	930.87 ± 2.87
Source TS	38.92 ± 1.30	71.69 ± 1.40	$\mathbf{26.64 \pm 2.24}$	90.16 ± 3.47	948.09 ± 1.90
Source ϵ-TS $(1/K)$	27.69 ± 1.84	53.19 ± 1.03	31.95 ± 3.10	71.69 ± 3.11	942.24 ± 1.99
Source ϵ-TS (0.75)	34.96 ± 1.25	69.27 ± 1.41	28.39 ± 2.78	86.40 ± 3.41	944.91 ± 2.06
Source greedy	54.47 ± 4.77	44.40 ± 4.58	43.87 ± 4.31	76.67 ± 6.31	1197.90 ± 2.10
Bias-aware baselines					
MIN-UCB	63.66 ± 1.23	233.87 ± 2.75	62.14 ± 1.18	265.09 ± 3.16	283.91 ± 2.76
Directional MIN-UCB	456.34 ± 0.92	760.27 ± 1.56	704.56 ± 2.18	749.60 ± 1.32	825.34 ± 0.98
Anchor-TS, empirical ($\sigma = 0.5$)	60.60 ± 3.87	137.63 ± 5.01	55.95 ± 4.00	142.10 ± 4.97	145.44 ± 4.68
Anchor-TS, empirical ($\sigma = 1$)	68.21 ± 2.23	212.67 ± 4.19	65.46 ± 2.55	232.56 ± 3.78	256.20 ± 3.95
Anchor-TS, posterior ($\sigma = 1$)	63.73 ± 2.71	212.23 ± 3.54	63.66 ± 2.90	228.38 ± 3.06	253.51 ± 3.85
Relative-confidence methods and internal variants					
ϵ proposal, width prior	30.80 ± 1.04	39.02 ± 0.90	32.81 ± 1.87	59.72 ± 2.52	145.99 ± 5.45
ϵ proposal, structure prior	33.19 ± 1.30	40.41 ± 1.39	34.25 ± 2.59	60.22 ± 2.51	173.73 ± 6.62
IDS proposal, structure prior	30.06 ± 1.02	39.18 ± 1.27	34.08 ± 1.98	$\mathbf{58.71 \pm 2.49}$	137.46 ± 4.48
Relative IDS	30.50 ± 1.13	$\mathbf{37.65 \pm 1.38}$	33.41 ± 1.72	60.74 ± 2.27	137.04 ± 3.64

Table 8: Overall mean regret of each comparator minus that of Relative IDS, averaged over all 45 settings; positive differences favor Relative IDS. Intervals are approximate paired 95% intervals, simultaneous over the 28,800-comparison family of Section H.5.

Comparator	Excess mean regret	Simultaneous 95% interval
Online controls		
Online IDS	65.58	[38.32, 92.83]
Online Beta-TS	94.90	[70.70, 119.09]
Online ϵ-TS $(1/K)$	93.04	[28.70, 157.38]
Online ϵ-TS (0.75)	83.31	[49.61, 117.02]
Exp-KL-MS	143.59	[119.57, 167.61]
LinMED	145.90	[93.47, 198.33]
Anchor-TS, empirical ($\sigma = 0.5$)	87.08	[17.04, 157.13]
Anchor-TS, empirical ($\sigma = 1$)	102.30	[68.36, 136.24]
Anchor-TS, posterior ($\sigma = 1$)	102.94	[70.49, 135.39]
Direct source reuse		
Source IDS	162.90	[155.58, 170.23]
Source TS	175.23	[169.14, 181.32]
Source ϵ-TS $(1/K)$	165.48	[158.34, 172.63]
Source ϵ-TS (0.75)	172.92	[165.82, 180.02]
Source greedy	223.60	[210.56, 236.64]
Bias-aware baselines		
MIN-UCB	121.87	[111.57, 132.16]
Directional MIN-UCB	639.35	[631.54, 647.17]
Anchor-TS, empirical ($\sigma = 0.5$)	48.48	[26.93, 70.03]
Anchor-TS, empirical ($\sigma = 1$)	107.15	[91.32, 122.98]
Anchor-TS, posterior ($\sigma = 1$)	104.43	[91.22, 117.65]
Internal variants		
ϵ proposal, width prior	1.80	$[-5.96, 9.56]$
ϵ proposal, structure prior	8.49	$[-2.56, 19.55]$
IDS proposal, structure prior	0.03	$[-5.84, 5.91]$

H.6 RANDOMIZED EMAIL DATA

The Hillstrom dataset contains 64000 customers, each randomly assigned to one of three campaign arms (Hillstrom, 2008). The context is twice the historical men's-merchandise indicator plus the historical women's-merchandise indicator, which takes four values; both indicators precede treatment. Visits, conversions and post-treatment spending are excluded from the context.

A reward-blind split with seed 120001 assigns 6395 customers to design, 16129 to the source pool, 19219 to the feedback bank and 22257 to reference evaluation. Design frequencies define the context weights. Three cells have positive mass; the fourth remains indexed with zero weight. Source budgets are 32, 128 and 512 customers per nonempty cell–action, that is, 288, 1152 and 4608 actual customers in total. Eight source subsamples each use sixteen paired feedback seeds, 125100–125115, with $T = 8000$. The source subsamples overlap within one pool, so source-draw variation does not measure uncertainty across independent trials.

Rewards during learning are Bernoulli draws with the bank's empirical visit probabilities. The algorithms do not observe those probabilities. For positive bank group counts $n_{x,a}^{\mathrm{bank}}$, the count-only radius is

$$r_{x,a} = \min\left\{1, \sqrt{\frac{\log(2MK/0.01)}{2n_{x,a}^{\mathrm{bank}}}}\right\}, \qquad \beta_x = 2\max_a r_{x,a}.$$

Empty groups use radius one. Under independent same-population randomized sampling, a union bound covers the bank–population differences with probability 0.99. Thus a finite bank is not assumed to have exactly zero bias relative to the source population. The remaining risk 0.04 is divided across the confidence rule's cells. The numerical common-shift scale is $\max_a r_{x,a}$. MIN-UCB receives the coordinate radii and Directional MIN-UCB an ellipsoid containing that box.

For each run, the learned policy is its executed action frequency conditional on context; unvisited cells use a uniform action vector. We average these policies over the declared source and feedback repetitions and evaluate the mixture on reference users. Visits are the learning reward and the primary evaluation outcome; conversion is a secondary outcome, not an assumed calibrated proxy. The point estimates and simultaneous intervals are computed as in Sections I and J. The simultaneous bounds cover all policy contrasts on the fixed partition. Because the reference-user outcomes were consulted during development, however, this evaluation is not a generalization test on untouched data. A training-seed sensitivity check uses a second batch of seeds, 132100–132115, with the same eight source subsamples, feedback bank, method parameters and reference users. In this batch, at 4608 source customers, Relative IDS and Online IDS have bank regrets 10.20 and 36.06, and the reference visit-rate difference is 0.2625 percentage points with joint interval $[-0.2936, 0.8155]$. The two training batches are evaluated on the same all-weight confidence event, with 22257 reference customers in total. For the first batch (seeds 125100–125115), the joint interval is $[0.0626, 0.6443]$ percentage points, and the rectangle interval originally reported is $[0.0205, 0.6834]$. Both interval constructions and training batches are retained in the reported uncertainty analysis.

H.7 REPRODUCTION

The release includes the underlying records, code and input hashes, every seed, source count, executed action and reward, budget diagnostics, and all original comparison rows. Numerical audits reconstruct feedback, regret, confidence costs and posterior updates from the saved observations.

I INDEPENDENT RANDOMIZED-USER POLICY EVALUATION

Estimand and assumptions. Let X denote a cell in a finite partition fixed using independent design covariates. The design also fixes nonnegative weights q_x summing to one. For outcome $j \in \{1, \ldots, J\}$ and treatment $a \in [K]$, let $Y_j(a) \in \{0, 1\}$ be the binary potential outcome and write $m_{x,a,j} = \mathbb{E}[Y_j(a) \mid X = x]$. Evaluation users are independent draws from the evaluation population. We assume consistency, randomized treatment with positive probability in each relevant cell, and no interference between users. We also require treatment assignment to be independent of potential outcomes conditional on the actual evaluation cells; randomization given finer covariates would not alone justify this coarsened comparison if treatment probabilities varied within a cell. These

conditions imply that, conditional on cell–treatment sample counts, each outcome success count is binomial with parameter $m_{x,a,j}$.

For a frozen policy $p(a \mid x)$, the standardized value is

$$V_{q,j}(p) = \sum_{x,a} q_x p(a \mid x) m_{x,a,j}.$$

The weights q describe the design distribution, not an asserted exact population context law. A cell with $q_x = 0$ makes no contribution.

An entirely observed interval. There are G cell–action–outcome coordinates, including empty design cells. Set $\alpha = \delta/(2G)$. For a coordinate with count n and success count s, define

$$L(s,n) = \begin{cases} 0, & s = 0, \\ B^{-1}_{s,n-s+1}(\alpha), & s > 0, \end{cases} \qquad U(s,n) = \begin{cases} 1, & s = n, \\ B^{-1}_{s+1,n-s}(1-\alpha), & s < n, \end{cases}$$

where $B^{-1}_{u,v}$ is the quantile function of a Beta(u, v) distribution. Thus an empty group has interval $[0, 1]$. For any signed weights $c_{x,a}$, put

$$\begin{aligned} \mathcal{L}_j(c) &= \sum_{c_{x,a} \geq 0} c_{x,a} L_{x,a,j} + \sum_{c_{x,a} < 0} c_{x,a} U_{x,a,j}, \\ \mathcal{U}_j(c) &= \sum_{c_{x,a} \geq 0} c_{x,a} U_{x,a,j} + \sum_{c_{x,a} < 0} c_{x,a} L_{x,a,j}. \end{aligned} \tag{20}$$

Proposition 3 (Simultaneous randomized-user comparisons)**.** *Under the assumptions above, with probability at least $1 - \delta$, every outcome and every signed weight vector satisfies*

$$\mathcal{L}_j(c) \leq \sum_{x,a} c_{x,a} m_{x,a,j} \leq \mathcal{U}_j(c).$$

In particular, $c_{x,a} = q_x p(a \mid x)$ gives a policy-value interval, and $c_{x,a} = q_x[p(a \mid x) - p_0(a \mid x)]$ gives a paired comparison. The same event covers all candidate policies and mixtures; reusing the reference users across candidates does not create independent samples.

Proof. Inverting the two binomial tail tests gives each Clopper–Pearson coordinate interval coverage at least $1 - 2\alpha$, conditional on its count. A union bound over the G coordinates gives the rectangle event with conditional probability at least $1 - \delta$. Neither independence between outcomes nor independence between policy comparisons is needed for this union bound. Integrating over the counts preserves the probability statement. On this single event, multiplying a coordinate interval by a nonnegative weight preserves its order, while a negative weight reverses it. Summing gives the bounds $\mathcal{L}_j(c) \leq \sum_{x,a} c_{x,a} m_{x,a,j} \leq \mathcal{U}_j(c)$ for every weight vector simultaneously. Conditioning first on the independent design data, then integrating, also permits the stated random partition and weights. □

The point estimate uses the observed group means. If a coordinate with nonzero weight has zero observations, the code marks the point estimate as unsupported and returns the corresponding range contribution instead of clipping the denominator. Coordinates whose weights cancel exactly need no observations.

Each learning run uses only source and feedback-bank observations. The reference sample was reused when developing later algorithm variants, so these are not fresh blind comparisons of methods chosen before any reference result was seen. Nevertheless, because the rectangle event implies every policy interval deterministically, it covers all policies on the fixed partition, including reference-dependent choices; we do not condition its probability statement on such a choice.

J Joint Confidence Bounds for Policy Contrasts

The rectangle of Section I is simple but can be loose for small policy differences. A mean-error norm bound gives a complementary observed interval. Both cover all weight vectors simultaneously, so their validity does not depend on which comparison is subsequently displayed.

Proposition 4 (Uniform linear policy contrasts)**.** *For one binary outcome, condition on the fixed cell design and reference cell–action assignments. Suppose observations are independent Bernoulli draws with mean μ_j within group j. For each nonempty group j, let $n_j > 0$ be its count and $\widehat{\mu}_j$ its empirical mean, and let d be the number of nonempty groups. For $\eta \in (0,1)$ and $u = \log(1/\eta)$, with probability at least $1-\eta$, simultaneously for every real vector w,*

$$\left|\sum_j w_j(\widehat{\mu}_j - \mu_j)\right| \le \frac{1}{2}\sqrt{\left(d + 2\sqrt{du} + 2u\right)\sum_j \frac{w_j^2}{n_j}}. \tag{21}$$

The vector w may be selected after inspecting the reference outcomes. The dimension d is not reduced to the support of the selected w. The sum in (21) *runs over nonempty groups; the weighted contributions of empty groups are bounded separately using $\mu_j \in [0,1]$, and a nonzero weight on an empty group yields no reported point estimate.*

Proof. If $d = 0$, both sides for the observed coordinates are zero; empty groups retain their range bounds. Assume henceforth that $d \ge 1$. Define $Z_j = 2\sqrt{n_j}(\widehat{\mu}_j - \mu_j)$. Hoeffding's lemma and independence across disjoint user groups give $\mathbb{E}e^{v^\top Z} \le e^{\|v\|_2^2/2}$ for every v. With an independent standard Gaussian vector G and $0 < t < 1$,

$$\mathbb{E}e^{t\|Z\|_2^2/2} = \mathbb{E}_Z\mathbb{E}_G e^{\sqrt{t}G^\top Z} \le \mathbb{E}_G e^{t\|G\|_2^2/2} = (1-t)^{-d/2}.$$

For $0 < \theta < 1/2$, setting $t = 2\theta$ in the previous display and using the elementary inequality $-\log(1-2\theta) - 2\theta \le 2\theta^2/(1-2\theta)$ gives

$$\log \mathbb{E}e^{\theta(\|Z\|_2^2 - d)} \le \frac{d\theta^2}{1-2\theta}.$$

Taking $\theta = \sqrt{u}/(\sqrt{d} + 2\sqrt{u})$ in Chernoff's bound gives a tail probability at most e^{-u} at $\|Z\|_2^2 = d + 2\sqrt{du} + 2u$. This is the identity-matrix case of the sub-Gaussian quadratic-form bound (Hsu et al., 2012). On that single event, Cauchy–Schwarz applied to $\frac{1}{2}\sum_j (w_j/\sqrt{n_j})Z_j$ proves (21) for every w. The assertion about data-selected weights follows because the bound holds for all w on a single event, without conditioning on the selected vector. □

For the joint report we allocate risk 0.025 to a single Clopper–Pearson rectangle over all cell–action coordinates and both outcomes, and $0.025/2$ to Proposition 4 for each outcome. The intersection of the two resulting intervals has total failure probability at most 0.05. Independence between visits and conversions is not required. We do not intersect two intervals that each have failure probability 0.05 while still claiming 0.05 overall. The original rectangle-only report is preserved, and every point estimate is unchanged.

K Testing the Bernoulli Gain Boundary across Horizons

We vary $T \in \{512, 768, 1024, 1536, 2048, 3072, 4096, 6144, 8192, 12288, 16384, 24576, 32768\}$ to examine the benefit condition implied by Theorem 2. Set $M = K = 4$ and $g_T = \frac{1}{2}\sqrt{MK/T} = 2/\sqrt{T}$. Independently in each cell, choose a uniformly random best action; its target mean is $1/2 + g_T/2$, and all other target means are $1/2 - g_T/2$. The instance changes with the horizon, so the experiment probes the gain boundary through horizon-indexed families; the statistical lower bound itself is established in Theorem 2.

Five source regimes. The first two have $N = T^2/(MK)$ and known $\beta = 0$, with source means equal to the target means or to the target means plus 0.1. The third has identical source and target means but only $N = T$. The remaining two again have $N = T^2/(MK)$, but every source mean equals $1/2$, independent of the best action, so the entire source distribution carries no information about which target action is best. They supply $\beta = g_T$ and $\beta = 0.1$, respectively, both valid for the tested horizons. In the first two regimes $N/T \to \infty$ and $\beta\sqrt{T} \to 0$. The third fails the sample-size condition; the last two fail the bias condition.

Source counts are balanced across cell–action pairs. Sampling these sufficient statistics directly from the exact binomial distribution has the same law as summing the specified number of independent binary observations but avoids expanding tens of millions of individual rows. The largest total source count is 67108864. Each algorithm receives only the source statistics, declared bounds, horizon, and its own executed feedback. The budget rules and both MIN-UCB variants receive $\delta = (1+N+T)^{-2}$, divided across cells, so the failure term δT in the budget rule's upper bound does not obscure the asymptotic comparison. LinMED retains its published round-dependent schedule, and the sampling rules retain their fixed parameters.

Comparators and checks. Thirteen methods use sixteen paired seeds at each of the 65 settings, giving 13,520 trajectories. The IDS draws and refresh schedule are unchanged from Section E. Besides Relative IDS, we retain Source greedy and greedy with the budget, Source IDS and Online IDS, selected online and source ϵ-TS, the empirical-mean scale-0.5 source and online Anchor variants, MIN-UCB, directional MIN-UCB, Exp-KL-MS, and selected LinMED. The Anchor controls sample each action once for initialization and use empirical means and Gaussian variances with the actual counts (no added pseudocount), with noise scale 0.5. All parameters are fixed across horizons.

The released traces permit independent reconstruction of source statistics, selected feedback, regret, posterior updates and budget constraints. Tables report all thirteen methods at $T \in \{512, 2048, 8192, 32768\}$.

Results and interpretation. At $T = 512, 2048, 8192, 32768$, respectively, Relative IDS has mean regret $8.94, 6.39, 0, 0$ for shared means and $9.46, 11.99, 0, 0$ for a common shift. Online IDS has $28.09, 57.17, 112.72, 192.78$ in both families. At $T = 32768$, the Relative IDS minus Online IDS difference is -192.78, with approximate paired-t interval $[-268.37, -117.18]$, Bonferroni-adjusted over all $20\binom{13}{2}$ method pairs. Source IDS and Source greedy also have zero observed regret in both abundant-source families. At $T = 8192$ and 32768 in these two families, the source filter of Relative IDS retained only the best action in every cell; since accurate source data identify that action, direct reuse also attains zero observed regret in these runs. These results show how informative source data change learning at the tested horizons. Zero observed regret describes the sixteen runs; the population guarantee is the bound in Theorem 2.

When $N = T$, Relative IDS's regret is $22.12, 36.07, 81.33, 156.70$. With uninformative source observations and $\beta = g_T$, it is $30.74, 65.51, 123.52, 198.68$; with $\beta = 0.1$, it is $30.20, 61.90, 105.78, 199.21$. In all three families, the normalized regret of Relative IDS stays on the scale of online learning, and the method-minus-online intervals at the largest horizon contain zero. In these three families, the source filter retained all four actions, so filtering alone does not resolve the choice of action. Together with the abundant-source cases, these limited-information families separate the roles of source count and information about target ranking.

The lower-bound construction uses source-assisted instances with gaps on the $(N + T)^{-1/2}$ scale. Here the gap is on the larger $T^{-1/2}$ scale, so abundant source data can identify the best action more accurately than in the lower-bound instances. The theorem establishes the worst-case rate; this experiment isolates the information mechanisms behind its two conditions.

Figure 3 summarizes source information and regret scaling. Figures 6 to 8 show these four horizons for six methods. Each source regime has two panels: Relative IDS with Online IDS and LinMED on the left, and Source IDS, Source greedy and Anchor-TS on the right. All panels use the same scale, $R_T/\sqrt{T}$, and average sixteen paired seeds. The plotted means are not offset or smoothed; coincident curves indicate equal observed means. Anchor-TS in these figures is the source-assisted empirical-mean variant with $\sigma = 0.5$. For readability these figures show means only; Table 9 reports all corresponding approximate 95% simultaneous mean intervals. The interval tables cover twenty settings and thirteen methods; the denser main figure covers 65 settings and thirteen methods, with the Bonferroni correction applied to each respective family. The following tables retain all thirteen methods on the original regret scale.

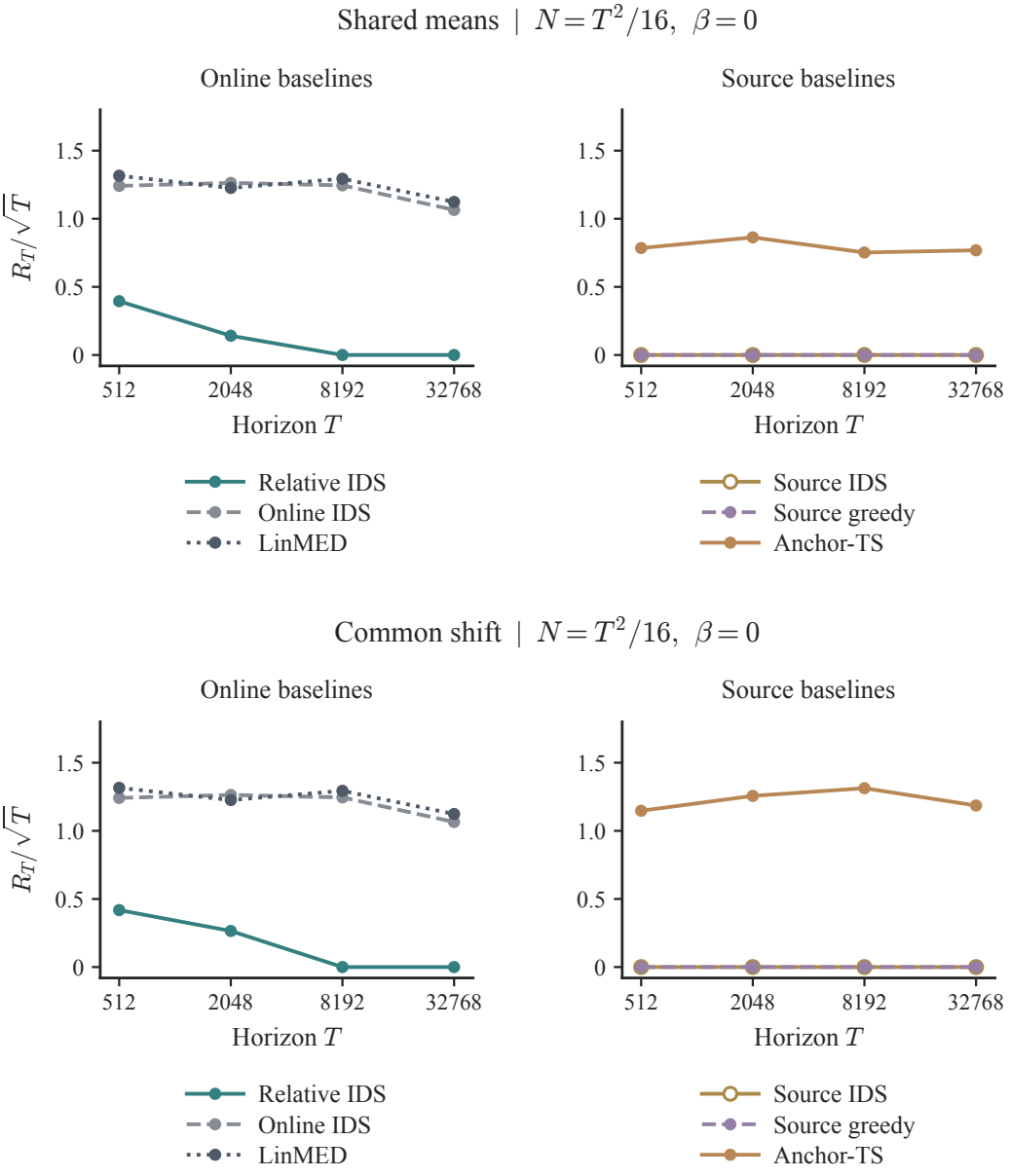


Figure 6: Complete mean trends with abundant, contrast-preserving source data: shared means (top) and a common $+0.1$ source shift (bottom). Both regimes use $N = T^2/16$ and $\beta = 0$. Source IDS and Source greedy overlap at zero. Each horizon uses a different target gap $2/\sqrt{T}$; the points are not prefixes of one learning trajectory. Uncertainty is reported in Table 9; zero observed regret does not rule out rare errors.

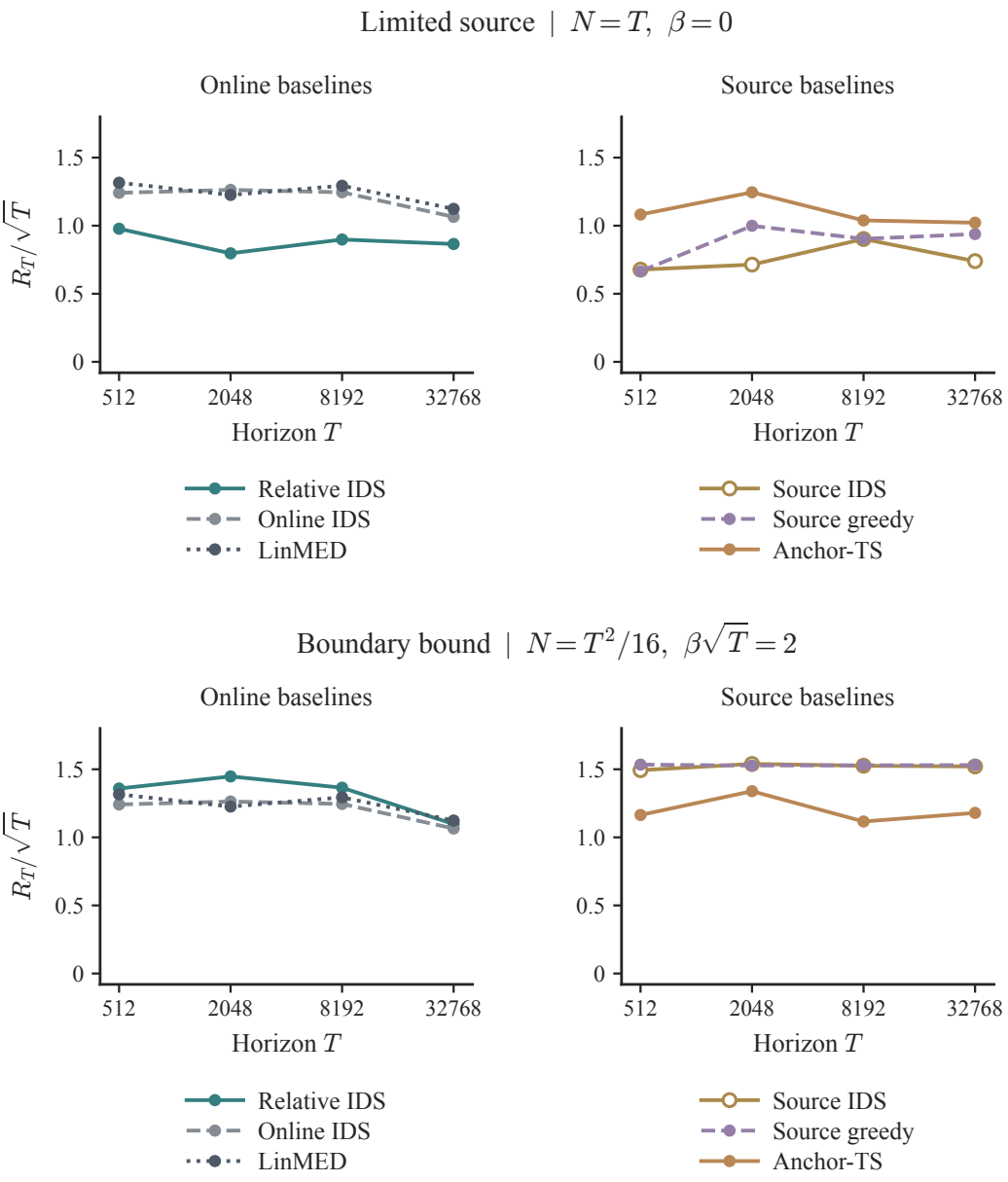


Figure 7: Complete mean trends with limited source information: $N = T$ and $\beta = 0$ (top), and $N = T^2/16$ with the boundary-scale bound $\beta = 2/\sqrt{T}$ (bottom). In the bottom regime every source mean is $1/2$, so the source does not identify the best target action. The horizon grid, target gaps and plotting conventions are the same as in Fig. 6. Corresponding simultaneous intervals appear in Table 9.

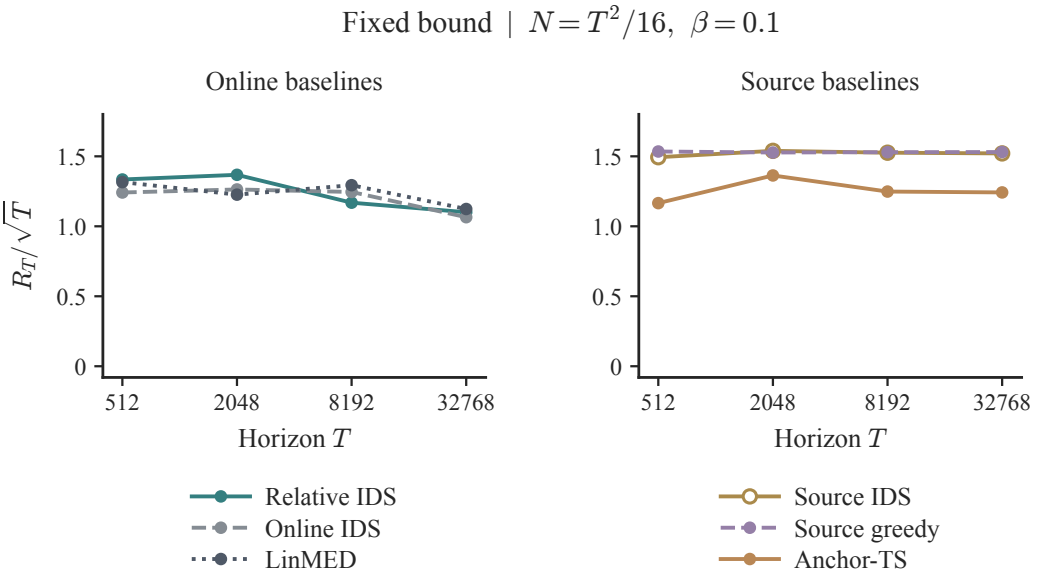


Figure 8: Complete mean trends with the fixed supplied bound $\beta = 0.1$. Source size is $N = T^2/16$ and every source mean is $1/2$. The source law is the same as in the boundary-bound regime; only the supplied bound changes. The horizon grid, target gaps and plotting conventions are the same as in Fig. 6. Corresponding simultaneous intervals appear in Table 9.

Table 9: Simultaneous intervals for the six methods in Figs. 6 to 8. Entries are lower and upper endpoints for mean $R_T/\sqrt{T}$, rounded to three decimals. The approximate 95% simultaneous t construction uses sixteen paired seeds and Bonferroni adjustment over all 260 declared method–setting means. Lower endpoints are clipped at zero. A zero-width interval reflects zero observed sample variance, not the exclusion of rare errors.

Method / horizon	512	2048	8192	32768
Shared means ($\beta = 0$)				
Relative IDS	$[0.292, 0.498]$	$[0.096, 0.186]$	$[0.000, 0.000]$	$[0.000, 0.000]$
Online IDS	$[0.900, 1.583]$	$[0.866, 1.661]$	$[0.904, 1.587]$	$[0.715, 1.415]$
LinMED	$[0.824, 1.807]$	$[0.746, 1.707]$	$[0.905, 1.683]$	$[0.519, 1.728]$
Source IDS	$[0.000, 0.000]$	$[0.000, 0.000]$	$[0.000, 0.000]$	$[0.000, 0.000]$
Source greedy	$[0.000, 0.000]$	$[0.000, 0.000]$	$[0.000, 0.000]$	$[0.000, 0.000]$
Anchor-TS	$[0.397, 1.173]$	$[0.529, 1.198]$	$[0.361, 1.144]$	$[0.361, 1.176]$
Common shift ($\beta = 0$)				
Relative IDS	$[0.275, 0.561]$	$[0.115, 0.415]$	$[0.000, 0.000]$	$[0.000, 0.000]$
Online IDS	$[0.900, 1.583]$	$[0.866, 1.661]$	$[0.904, 1.587]$	$[0.715, 1.415]$
LinMED	$[0.824, 1.807]$	$[0.746, 1.707]$	$[0.905, 1.683]$	$[0.519, 1.728]$
Source IDS	$[0.000, 0.000]$	$[0.000, 0.000]$	$[0.000, 0.000]$	$[0.000, 0.000]$
Source greedy	$[0.000, 0.000]$	$[0.000, 0.000]$	$[0.000, 0.000]$	$[0.000, 0.000]$
Anchor-TS	$[0.768, 1.526]$	$[0.983, 1.530]$	$[1.043, 1.583]$	$[0.876, 1.496]$
Limited source ($N = T$)				
Relative IDS	$[0.562, 1.393]$	$[0.351, 1.243]$	$[0.453, 1.344]$	$[0.510, 1.222]$
Online IDS	$[0.900, 1.583]$	$[0.866, 1.661]$	$[0.904, 1.587]$	$[0.715, 1.415]$
LinMED	$[0.824, 1.807]$	$[0.746, 1.707]$	$[0.905, 1.683]$	$[0.519, 1.728]$
Source IDS	$[0.294, 1.061]$	$[0.276, 1.152]$	$[0.518, 1.286]$	$[0.435, 1.045]$
Source greedy	$[0.224, 1.105]$	$[0.259, 1.740]$	$[0.343, 1.464]$	$[0.273, 1.605]$
Anchor-TS	$[0.802, 1.362]$	$[0.836, 1.655]$	$[0.735, 1.343]$	$[0.637, 1.408]$
Boundary bound ($\beta\sqrt{T} = 2$)				
Relative IDS	$[0.994, 1.724]$	$[1.064, 1.831]$	$[0.942, 1.787]$	$[0.656, 1.539]$
Online IDS	$[0.900, 1.583]$	$[0.866, 1.661]$	$[0.904, 1.587]$	$[0.715, 1.415]$
LinMED	$[0.824, 1.807]$	$[0.746, 1.707]$	$[0.905, 1.683]$	$[0.519, 1.728]$
Source IDS	$[0.896, 2.092]$	$[0.983, 2.095]$	$[1.036, 2.016]$	$[1.027, 2.014]$
Source greedy	$[0.918, 2.151]$	$[1.000, 2.054]$	$[1.055, 2.004]$	$[1.056, 2.005]$
Anchor-TS	$[0.811, 1.517]$	$[0.967, 1.711]$	$[0.823, 1.410]$	$[0.783, 1.576]$
Fixed bound ($\beta = 0.1$)				
Relative IDS	$[0.915, 1.754]$	$[1.069, 1.666]$	$[0.769, 1.568]$	$[0.630, 1.571]$
Online IDS	$[0.900, 1.583]$	$[0.866, 1.661]$	$[0.904, 1.587]$	$[0.715, 1.415]$
LinMED	$[0.824, 1.807]$	$[0.746, 1.707]$	$[0.905, 1.683]$	$[0.519, 1.728]$
Source IDS	$[0.896, 2.092]$	$[0.983, 2.095]$	$[1.036, 2.016]$	$[1.027, 2.014]$
Source greedy	$[0.918, 2.151]$	$[1.000, 2.054]$	$[1.055, 2.004]$	$[1.056, 2.005]$
Anchor-TS	$[0.825, 1.506]$	$[1.131, 1.596]$	$[1.036, 1.460]$	$[0.903, 1.581]$

Table 10: Horizon-indexed Bernoulli comparison: abundant, shared means. Mean cumulative pseudo-regret over 16 paired seeds; all 13 methods are retained. The instance gap is $2/\sqrt{T}$, not fixed as T changes. Lower is better.

Method / horizon	512	2048	8192	32768
Relative IDS	8.938	6.394	0.000	0.000
Greedy + budget	0.000	0.000	0.000	0.000
Source greedy	0.000	0.000	0.000	0.000
Source IDS	0.000	0.000	0.000	0.000
Online IDS	28.091	57.168	112.716	192.778
Online ϵ-TS, .75	28.240	57.229	114.078	223.008
Source ϵ-TS, $1/K$	0.017	0.000	0.000	0.000
Source Anchor, empirical, $\sigma = .5$	17.766	39.073	68.096	139.129
Online Anchor, empirical, $\sigma = .5$	28.268	60.825	107.919	240.140
MIN-UCB	0.000	0.000	0.000	0.000
Directional MIN-UCB	0.000	0.000	0.000	0.000
Exp-KL-MS	32.102	63.471	123.675	243.719
LinMED, selected	29.770	55.489	117.092	203.384

Table 11: Horizon-indexed Bernoulli comparison: abundant, common shift. Mean cumulative pseudo-regret over 16 paired seeds; all 13 methods are retained. The instance gap is $2/\sqrt{T}$, not fixed as T changes. Lower is better.

Method / horizon	512	2048	8192	32768
Relative IDS	9.463	11.990	0.000	0.000
Greedy + budget	0.000	0.000	0.000	0.000
Source greedy	0.000	0.000	0.000	0.000
Source IDS	0.000	0.000	0.000	0.000
Online IDS	28.091	57.168	112.716	192.778
Online ϵ-TS, .75	28.240	57.229	114.078	223.008
Source ϵ-TS, $1/K$	0.077	0.000	0.000	0.000
Source Anchor, empirical, $\sigma = .5$	25.959	56.870	118.824	214.671
Online Anchor, empirical, $\sigma = .5$	28.268	60.825	107.919	240.140
MIN-UCB	0.249	9.797	125.246	257.472
Directional MIN-UCB	0.713	0.713	87.296	261.331
Exp-KL-MS	32.102	63.471	123.675	243.719
LinMED, selected	29.770	55.489	117.092	203.384

Table 12: Horizon-indexed Bernoulli comparison: source size $N = T$. Mean cumulative pseudo-regret over 16 paired seeds; all 13 methods are retained. The instance gap is $2/\sqrt{T}$, not fixed as T changes. Lower is better.

Method / horizon	512	2048	8192	32768
Relative IDS	22.125	36.073	81.327	156.699
Greedy + budget	15.043	45.224	81.763	169.907
Source greedy	15.043	45.224	81.763	169.907
Source IDS	15.324	32.320	81.657	133.932
Online IDS	28.091	57.168	112.716	192.778
Online ϵ-TS, .75	28.240	57.229	114.078	223.008
Source ϵ-TS, $1/K$	18.070	39.015	83.980	158.251
Source Anchor, empirical, $\sigma = .5$	24.484	56.359	94.033	185.009
Online Anchor, empirical, $\sigma = .5$	28.268	60.825	107.919	240.140
MIN-UCB	27.599	54.994	109.526	224.447
Directional MIN-UCB	32.317	66.791	131.541	261.694
Exp-KL-MS	32.102	63.471	123.675	243.719
LinMED, selected	29.770	55.489	117.092	203.384

Table 13: Horizon-indexed Bernoulli comparison: boundary bound ($\beta\sqrt{T} = 2$). Mean cumulative pseudo-regret over 16 paired seeds; all 13 methods are retained. The instance gap is $2/\sqrt{T}$, not fixed as T changes. Lower is better.

Method / horizon	512	2048	8192	32768
Relative IDS	30.743	65.507	123.516	198.683
Greedy + budget	34.720	69.117	138.442	277.069
Source greedy	34.720	69.117	138.442	277.069
Source IDS	33.797	69.653	138.105	275.298
Online IDS	28.091	57.168	112.716	192.778
Online ϵ-TS, .75	28.240	57.229	114.078	223.008
Source ϵ-TS, $1/K$	33.847	69.410	140.375	280.775
Source Anchor, empirical, $\sigma = .5$	26.345	60.596	101.047	213.550
Online Anchor, empirical, $\sigma = .5$	28.268	60.825	107.919	240.140
MIN-UCB	33.173	68.228	138.162	272.201
Directional MIN-UCB	33.693	69.120	137.152	272.666
Exp-KL-MS	32.102	63.471	123.675	243.719
LinMED, selected	29.770	55.489	117.092	203.384

Table 14: Horizon-indexed Bernoulli comparison: fixed bound ($\beta = 0.1$). Mean cumulative pseudo-regret over 16 paired seeds; all 13 methods are retained. The instance gap is $2/\sqrt{T}$, not fixed as T changes. Lower is better.

Method / horizon	512	2048	8192	32768
Relative IDS	30.196	61.902	105.782	199.214
Greedy + budget	34.720	69.117	138.442	277.069
Source greedy	34.720	69.117	138.442	277.069
Source IDS	33.797	69.653	138.105	275.298
Online IDS	28.091	57.168	112.716	192.778
Online ϵ-TS, .75	28.240	57.229	114.078	223.008
Source ϵ-TS, $1/K$	33.847	69.410	140.375	280.775
Source Anchor, empirical, $\sigma = .5$	26.373	61.695	112.974	224.838
Online Anchor, empirical, $\sigma = .5$	28.268	60.825	107.919	240.140
MIN-UCB	33.101	68.739	137.530	266.139
Directional MIN-UCB	33.549	70.487	137.942	268.435
Exp-KL-MS	32.102	63.471	123.675	243.719
LinMED, selected	29.770	55.489	117.092	203.384

L A SOURCE-FILTERED TARGET KL-UCB CONTROL

The main guarantee does not depend on the IDS proposal. To check whether the effect of source information also depends on it, we implemented a simpler optimistic control with a fixed filter. It first constructs

$$\mathcal{A}_x = \{a : U^s_{x,a} + \beta_x \geq \max_b L^s_{x,b}\}$$

from source observations, then runs target KL-UCB restricted to $\mathcal{A}_x$. Source outcomes never enter a target index, and an unchosen target reward is used only by the evaluator. The no-source control uses the same target rule and risk threshold with all actions active.

This control runs on a separately seeded reconstruction of the generator specified in Section K, using the fixed seed block 150000–150015 and common random numbers. The original horizon experiment used seeds 129000–129015 (Section E.5) and a different mapping from seeds to random streams, so the two sets of runs are not paired. We therefore do not pool these observations with the earlier method table or present differences from IDS as paired comparisons. The complete run-level output, seed provenance, and executable checks are included with the project.

Table 15: Source-filtered target KL-UCB as a mechanism control. Entries are mean cumulative pseudo-regret $\pm$ one Monte Carlo standard error over 16 paired reconstruction seeds. Both methods use the fixed target confidence threshold; $|\mathcal{A}|$ is the mean number of source-active actions per cell.

Source regime	T	SF-KL-UCB	Online KL-UCB	$\lvert\mathcal{A}\rvert$
Abundant, shared means	8,192	0.00 ± 0.00	129.04 ± 0.99	1.00
Abundant, common shift	8,192	0.00 ± 0.00	129.04 ± 0.99	1.00
$N = T$, shared means	32,768	252.39 ± 2.33	252.39 ± 2.33	4.00
Uninformative, $\beta = 2/\sqrt{T}$	32,768	257.50 ± 2.09	257.50 ± 2.09	4.00
Uninformative, $\beta = 0.1$	32,768	257.50 ± 2.09	257.50 ± 2.09	4.00

The control reaches zero regret once abundant source data reduce the active set to the true best action. When the source is too small or does not identify the ranking, the filter retains all actions and the complete action–reward path equals that of online KL-UCB under the shared random stream. Thus the source-size and bias boundary is not specific to IDS. The cumulative guard serves other purposes: it admits any proposal and certifies the executed trajectory.

We also reconstructed the five controlled paid-calibration environments of Section F with seed numbers 152000–152031, $T = 8000$, 640 source draws per cell–action, and $c = 8$ charged pilots. The reconstruction maps seeds to random streams differently from Section F, and within a seed all five environments share one context and reward stream. Every observed relative-bias envelope covered the true diameter, but the mean envelope ranged from 1.14 to 1.20, and the resulting filter retained all four actions in every cell. Paid source-filtered KL-UCB and the matched-c pilot-only control were therefore pathwise identical in all 160 paired runs. Because the environments share target means and streams and the filter kept every action, these two methods and no-pilot online KL-UCB each have only 32 distinct trajectories, one per seed, so the 32 seeds are the replicates. Mean regret ($\pm$ standard error over seeds) was 240.60 ± 4.58, compared with 235.64 ± 4.68 for no-pilot online KL-UCB; the paired difference was 4.96 with an unadjusted 95% paired-t interval $[-4.93, 14.84]$. This negative ablation shows that finite-sample certificate coverage alone need not make hard source filtering informative.

M ANCHOR-TS, OOPE AND OFFLINE-ORACLE COMPARISONS

M.1 A COMMON MODEL AND FIXED PARAMETERS

The shared-model suite uses the Bernoulli model in Section H, with $M = K = 4$, gaps $\Delta \in \{.02, .08, .2\}$ and source counts per cell–action $n^s \in \{128, 2048, 32768\}$. Contexts cycle through the four cells, giving each cell exactly 3000 of the 12000 target rounds. This fixed design supplies the same local horizon to every method and permits independent one-hot OOPE instances. It is a comparator experiment, not an additional test of the iid-context minimax theorem. All 25 methods

and implementation settings use the same source observations and potential feedback tables on seeds 131000–131031. Policies update only from their selected actions' rewards.

Parameters were tuned on the earlier, separate shared-model training seeds 123000–123007 and then frozen: online $\epsilon = .75$, source $\epsilon = 1/K$, LinMED $\alpha_{\text{emp}} = .9$, and E2D.Off exploration parameters $\gamma = 1000$ online and 100 with source data. These choices minimize training regret over equally weighted settings; they are not selected again on the new comparison. Relative IDS keeps its 256-draw proposal, eight-visit refresh and confidence budget.

M.2 Implementation contracts

Anchor-TS. The index is the median of an online Gaussian draw, a bias-adjusted hybrid draw and the online mean (Li et al., 2026). Each action is sampled once before Gaussian indices are used. For target count n_a and source count n_a^s, the online and hybrid empirical means divide their reward sums by n_a and $n_a + n_a^s$, respectively; the corresponding Gaussian variances are σ^2/n_a and $\sigma^2/(n_a + n_a^s)$, with no added pseudocount. When a paid pilot has already sampled every action, its observed rewards provide this initialization and are replayed exactly once. We test $\sigma = 1$ and $\sigma = .5$, with and without source observations. The existing posterior variant with a unit pseudocount is retained separately as a sensitivity check; it is not the empirical-mean implementation of Anchor-TS. For shifted synthetic data the supplied absolute envelope is the declared common-shift scale plus $\beta/2$. For Hillstrom it is the same coordinatewise, bank-count calibration radius supplied to MIN-UCB. These are local implementations, not implementations obtained from the original authors.

OOPE. We specialize Algorithm 1 of arXiv v3 to one-hot actions (Vijayan et al., 2026). The extended D-optimal design is solved exactly by water-filling the diagonal source information; no numerical Frank–Wolfe approximation is needed. Distinct source blocks are consumed in each phase. We report the unit noise scale shown in that algorithm, the known .5-sub-Gaussian scale, and the latter scale with the allocation coefficient from OOPE's Eq. (3). These are specified variants of OOPE, not a reproduction of OOPE-FW. Source rewards are used only in the shared-model suite, where the shared linear model holds.

OEOE / E2D.Off. We combine Version Space Averaging with E2D.Off (Foster et al., 2024). In cell x, the predeclared class is

$$\mathcal{F}_x = \{0.35 + 0.02x + v : v \in \{0, .02, .08, .2\}^4\}, \qquad |\mathcal{F}_x| = 256.$$

The squared-loss oracle returns an exact empirical minimizer. Every prefix constraint is retained; the decision LP includes all models, not just surviving models. The source version replays every source prefix through the same oracle before target play. The online version starts without source observations. Both receive the fixed, realizable class, as do finite-class TS controls with and without source data. Other methods operate over unrestricted Bernoulli means rather than receiving this finite class. Thus the comparison distinguishes an oracle/model-information advantage from an allocation advantage. Arbitrary shifted-source pooling is outside this implementation's contract and is not scored as an in-model E2D.Off comparison.

M.3 Paired extensions and uncertainty

Anchor-TS is also evaluated on every original synthetic world (45 settings, 32 seeds) and Hillstrom world (eight source draws, three budgets, sixteen feedback seeds). The saved source data, contexts and potential rewards are reused exactly. Existing methods' actions, feedback and regret are reconstructed from their original records; these are the current IDS comparisons, not the earlier proposal experiments. The extension retains six Anchor configurations, for a total of 23 reported methods per original world, including archived proposal controls in the machine-readable results.

Within each suite, setting-wise paired-t comparisons are corrected using the original, conservative families of settings and method pairs: $9\binom{27}{2}$ for the shared-model suite, $45\binom{25}{2}$ for the shifted synthetic suite, and $24\binom{25}{2}$ for Hillstrom. These families retain the original counts of 27, 25 and 25 configurations, although the current results retain 25, 23 and 23, respectively, after two obsolete configurations were removed from each suite. When aggregating several settings, we first average within each feedback seed, then compute paired intervals across seeds; settings sharing a seed are not

treated as independent replicates. Aggregate comparisons with Relative IDS use Bonferroni factors 26, 120 and 72 for the shared, shifted and Hillstrom suites, respectively, retaining every comparator and aggregate setting in each family. Hillstrom uncertainty is conditional on the fixed, overlapping source subsamples and feedback bank. Plot bands are pointwise, approximate 95% intervals across these seed-level averages; they are not simultaneous bands or population uncertainty.

Audits independently replay the Anchor indices, reconstruct OOPE phase allocations and source consumption, and check E2D.Off's oracle prefixes and LP optima. Executed-action feedback, cumulative regret and our method's prefix budgets are checked against the saved trajectories. The following means summarize the results; removing repeated table rows does not reduce the comparison families above.

M.4 Main empirical observations

Source reuse and source budget. Table 16 summarizes the principal methods. At 2,048 source observations per cell–action in the shared-model suite, Relative IDS has mean regret 20.44, compared with 135.61 for Online IDS and 55.88 for source-assisted empirical-mean Anchor-TS. At the largest Hillstrom source budget, the corresponding means are 10.75, 42.31 and 26.17. These comparisons show that source information can reduce regret relative to online learning in both suites. These suites also favor direct reuse: Source IDS has mean regret 14.20 at that shared-model budget and 1.33 at the largest Hillstrom budget, and its shared-model advantage over Relative IDS, averaged over all settings, is 7.13 with simultaneous interval $[2.89, 11.37]$. The shared-model suite satisfies the equal-means model of Source IDS by construction, and the Hillstrom source pool and feedback bank are random splits of one trial. In the Hillstrom experiment, Relative IDS uses the positive bound β_x of Section H.6 instead of assuming zero bias. Table 18 and the setting-wise tables give all methods and source budgets.

Anchor-TS implementation and scale. In Table 17, the initialized source-assisted empirical implementation and the separate posterior variant at $\sigma = 1$ have mean regrets differing by 0.44–4.48 across the five regime averages. These descriptive averages do not establish equivalence. Scale effects remain setting-dependent: in the separate shared-model suite, the empirical version with $\sigma = .5$ has lower mean regret than $\sigma = 1$ at source budgets 128 and 2,048, but higher mean regret at 32,768 (52.30 versus 34.56). We retain both scales rather than selecting one separately for each test setting.

Common reward-level shifts. In the original synthetic suite, source-assisted empirical Anchor-TS with $\sigma = .5$ has mean regret 60.60 under shared means, 137.63 under a positive common shift and 55.95 under a negative common shift. The corresponding Relative IDS means are 30.50, 37.65 and 33.41. Both shifts preserve action contrasts, yet the finite-sample response depends on the method and the sign of the shift. These comparisons illustrate the relevance of preserved contrasts; they do not isolate the bias-bound definition from other algorithmic differences.

M.5 Aggregate comparisons and learning curves

The learning curves group methods by comparison purpose. Relative IDS is repeated as the common reference. All axes start at zero; vertical scales differ between panels. Lines connect recorded means without smoothing; shading uses the pointwise interval construction defined above. In the legends, Anchor-TS denotes the source-assisted empirical-mean implementation with $\sigma = .5$; its online variant uses the same recursion and scale. Source ϵ-TS uses $\epsilon = 1/K$. The OOPE curve uses the coefficient from OOPE's Eq. (3) and $\sigma = .5$; E2D.Off uses the source-assisted implementation. Complete implementation variants remain in the tables.

Table 16: Regret by source budget: shared means ($T = 12000$, counts per cell–action) and Hillstrom ($T = 8000$, customer counts). N/A marks methods whose model assumptions are incompatible with that suite. Anchor-TS (empirical) uses initialized empirical means and $\sigma = .5$; OOPE (one-hot) uses $\sigma = .5$. E2D.Off and finite-class TS receive an additional realizable class. Additional configurations appear in Table 18.

	Shared means			Hillstrom		
Method	128	2048	32768	288	1152	4608
Online controls						
Online IDS	135.61	135.61	135.61	42.31	42.31	42.31
Online ϵ-TS, .75	149.00	149.00	149.00	47.13	47.13	47.13
Anchor-TS, online	160.51	160.51	160.51	50.93	50.93	50.93
Exp-KL-MS	197.11	197.11	197.11	70.09	70.09	70.09
LinMED	211.70	211.70	211.70	88.07	88.07	88.07
Direct source reuse						
Source IDS	53.75	14.20	0.00	26.10	14.02	1.33
Source ϵ-TS, $1/K$	58.21	19.04	0.00	30.37	18.74	3.08
Source TS	81.05	27.89	0.03	42.02	29.55	8.09
Source greedy	90.00	18.13	0.00	91.81	42.05	0.85
Bias-aware transfer						
MIN-UCB	155.81	22.95	0.00	133.27	116.19	76.86
Directional MIN-UCB	815.91	570.31	0.00	264.22	263.99	242.40
Anchor-TS (empirical)	83.12	55.88	52.30	38.80	31.68	26.17
Shared-model controls						
OOPE (one-hot)	900.00	877.79	233.38	N/A	N/A	N/A
OEOE / E2D.Off	469.77	205.28	86.09	N/A	N/A	N/A
Finite-class TS, source	97.71	33.88	0.00	N/A	N/A	N/A
Relative-confidence allocation						
Relative IDS	67.12	20.44	1.78	29.15	20.87	10.75

Table 17: Anchor-TS sensitivity across the five original shift regimes. Mean regret over all nine gap–budget combinations and 32 paired seeds. Empirical-mean scale choices and the separate posterior sensitivity are grouped explicitly. These are descriptive means; close values do not establish statistical equivalence.

Method	Shared	Shift +	Shift -	Residual	Rankings
Reference methods					
Online IDS	125.44	125.44	125.44	125.44	125.44
Source IDS	26.40	51.30	35.28	70.01	930.87
Source ϵ-TS, $1/K$	27.69	53.19	31.95	71.69	942.24
Relative IDS	30.50	37.65	33.41	60.74	137.04
Initialized empirical-mean Anchor-TS					
Anchor-TS online, empirical, $\sigma = 1$	162.17	162.17	162.17	162.17	162.17
Anchor-TS online, empirical, $\sigma = .5$	146.95	146.95	146.95	146.95	146.95
Anchor-TS source, empirical, $\sigma = 1$	68.21	212.67	65.46	232.56	256.20
Anchor-TS source, empirical, $\sigma = .5$	60.60	137.63	55.95	142.10	145.44
Separate posterior sensitivity					
Anchor-TS online, posterior, $\sigma = 1$	162.81	162.81	162.81	162.81	162.81
Anchor-TS source, posterior, $\sigma = 1$	63.73	212.23	63.66	228.38	253.51

Table 18: Additional configurations beyond Table 16, using the same source budgets and horizons. N/A means the configuration was not included in that suite; shared-model methods require assumptions unavailable in Hillstrom. The bottom block contains the remaining Hillstrom proposal controls. The Eq. (3) labels refer to OOPE's equation.

	Shared means			Hillstrom		
Method	128	2048	32768	288	1152	4608
Additional online controls						
Online Beta-TS	161.47	161.47	161.47	53.69	53.69	53.69
Anchor-TS online, posterior, $\sigma = 1$	151.90	151.90	151.90	66.34	66.34	66.34
Anchor-TS online, empirical, $\sigma = 1$	162.06	162.06	162.06	71.76	71.76	71.76
E2D.Off, online	530.22	530.22	530.22	N/A	N/A	N/A
Finite-class TS, online	172.97	172.97	172.97	N/A	N/A	N/A
Additional source-assisted Anchor-TS configurations						
Anchor-TS source, posterior, $\sigma = 1$	106.70	55.57	34.47	62.33	59.11	51.76
Anchor-TS source, empirical, $\sigma = 1$	112.58	59.53	34.56	65.39	61.94	50.37
Additional OOPE configurations						
OOPE, $\sigma = 1$	900.00	885.24	546.46	N/A	N/A	N/A
OOPE Eq.(3), $\sigma = .5$	895.13	770.71	175.23	N/A	N/A	N/A
Additional Hillstrom controls						
Online ϵ-TS, $1/K$	N/A	N/A	N/A	47.72	47.72	47.72
Source ϵ-TS, .75	N/A	N/A	N/A	40.32	26.35	5.64
Ablation: ϵ instead of IDS	N/A	N/A	N/A	35.14	24.84	11.71

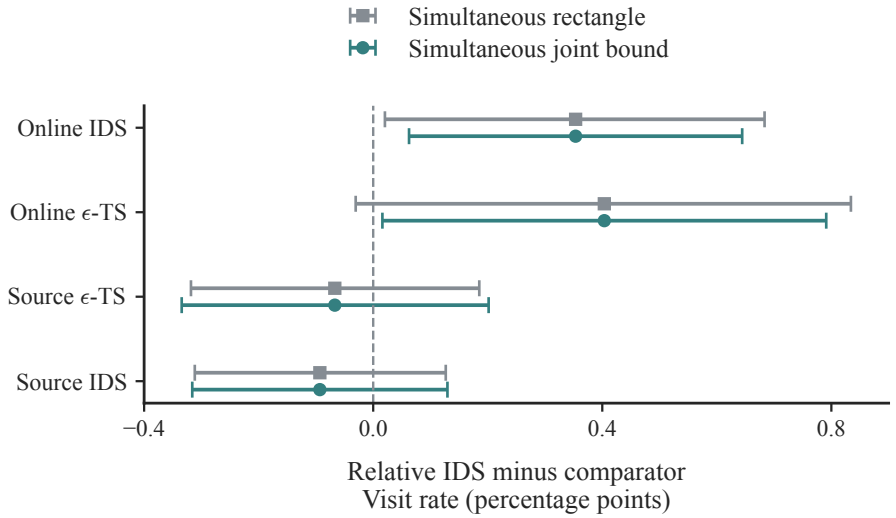


Figure 9: Hillstrom visit-rate differences with 4608 source customers and 22257 reference users. Positive values favor Relative IDS. Bars show the simultaneous rectangle and joint confidence bounds for the learned policy mixtures. Online and source ϵ-TS use $\epsilon = .75$ and $1/K$, respectively.

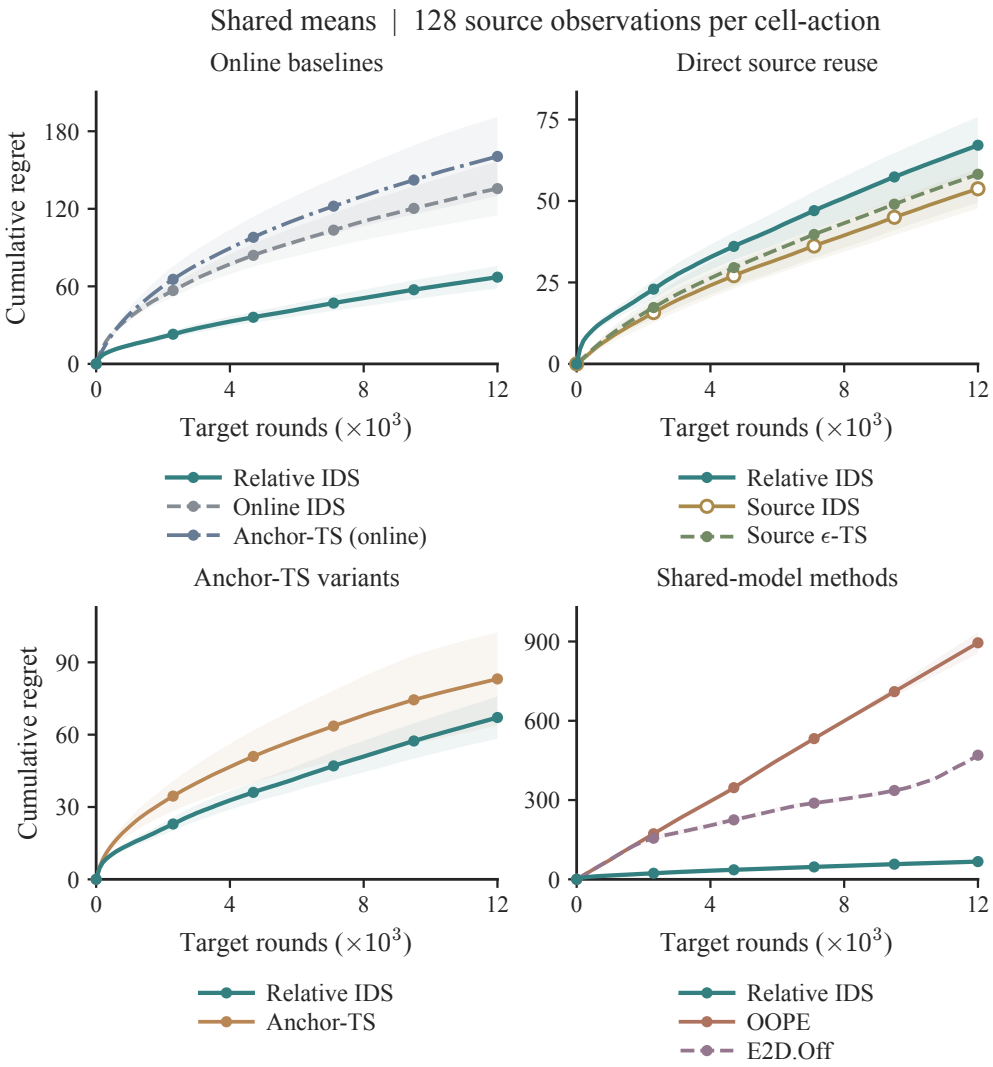


Figure 10: Learning curves in the shared-model suite with 128 source observations per cell–action. Each curve averages three gaps within each of 32 paired seeds. Panels separate online controls (including online Anchor-TS), direct reuse, source-assisted Anchor-TS and shared-model methods, with Relative IDS repeated for comparison. Shading shows pointwise approximate 95% intervals; vertical scales differ.

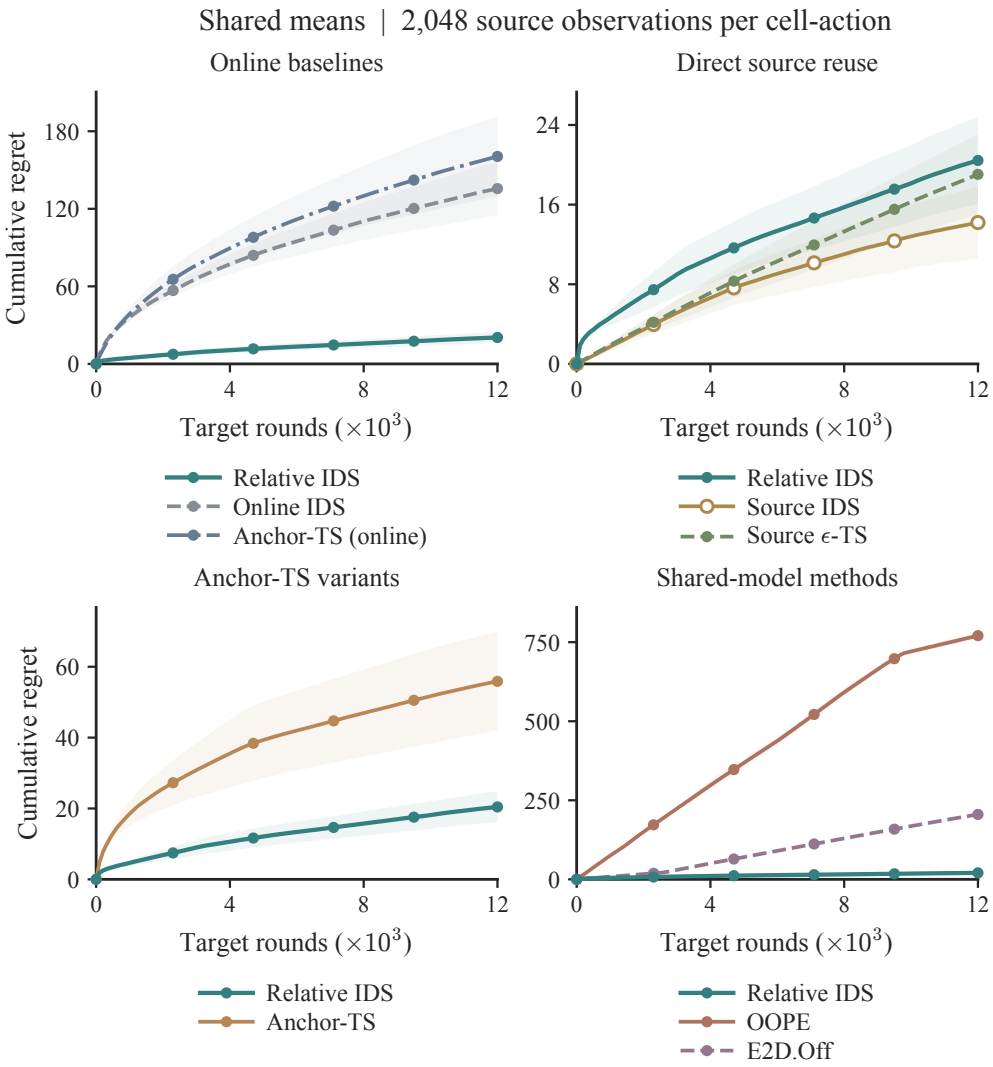


Figure 11: Learning curves in the shared-model suite with 2,048 source observations per cell–action. Each curve averages three gaps within each of 32 paired seeds. Panels separate online controls (including online Anchor-TS), direct reuse, source-assisted Anchor-TS and shared-model methods, with Relative IDS repeated for comparison. Shading shows pointwise approximate 95% intervals; vertical scales differ.

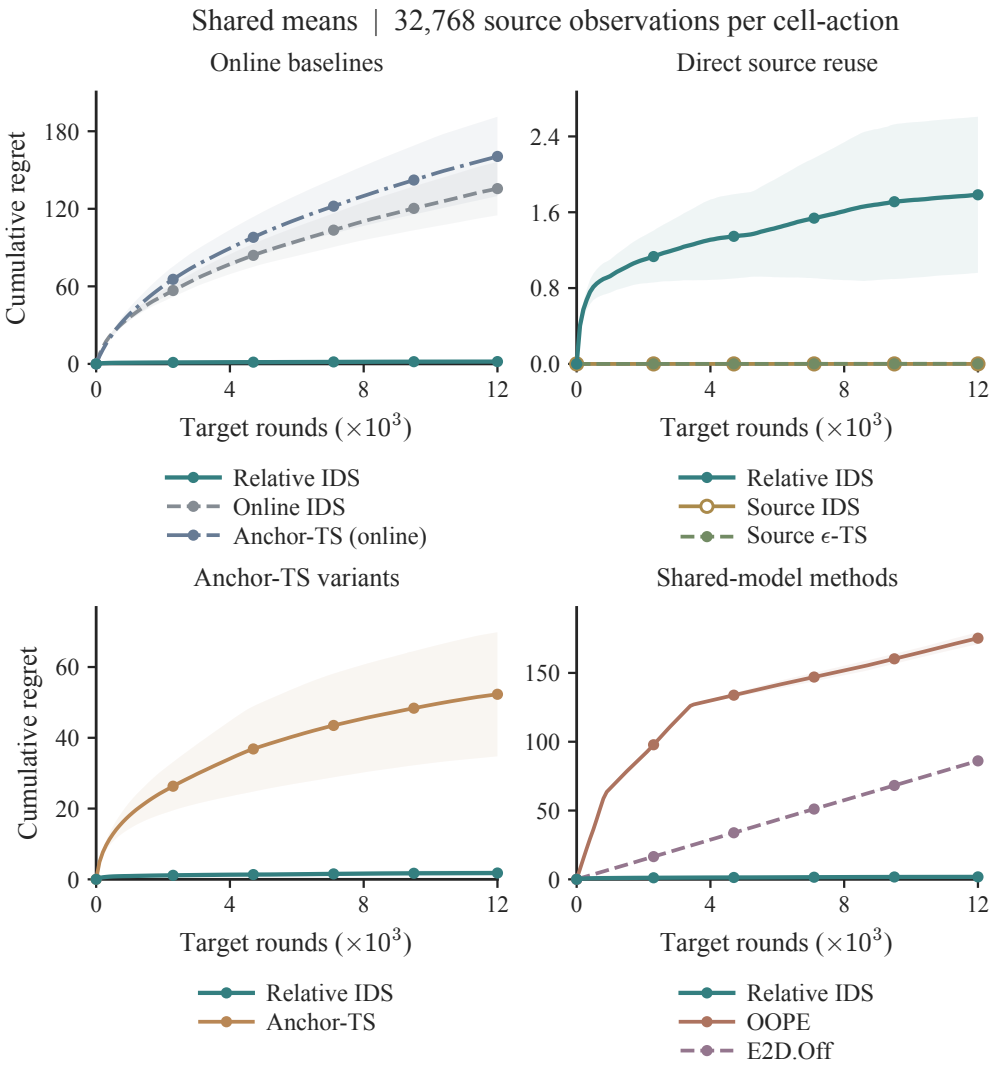


Figure 12: Learning curves in the shared-model suite with 32,768 source observations per cell–action. Each curve averages three gaps within each of 32 paired seeds. Panels separate online controls (including online Anchor-TS), direct reuse, source-assisted Anchor-TS and shared-model methods, with Relative IDS repeated for comparison. Shading shows pointwise approximate 95% intervals; vertical scales differ.

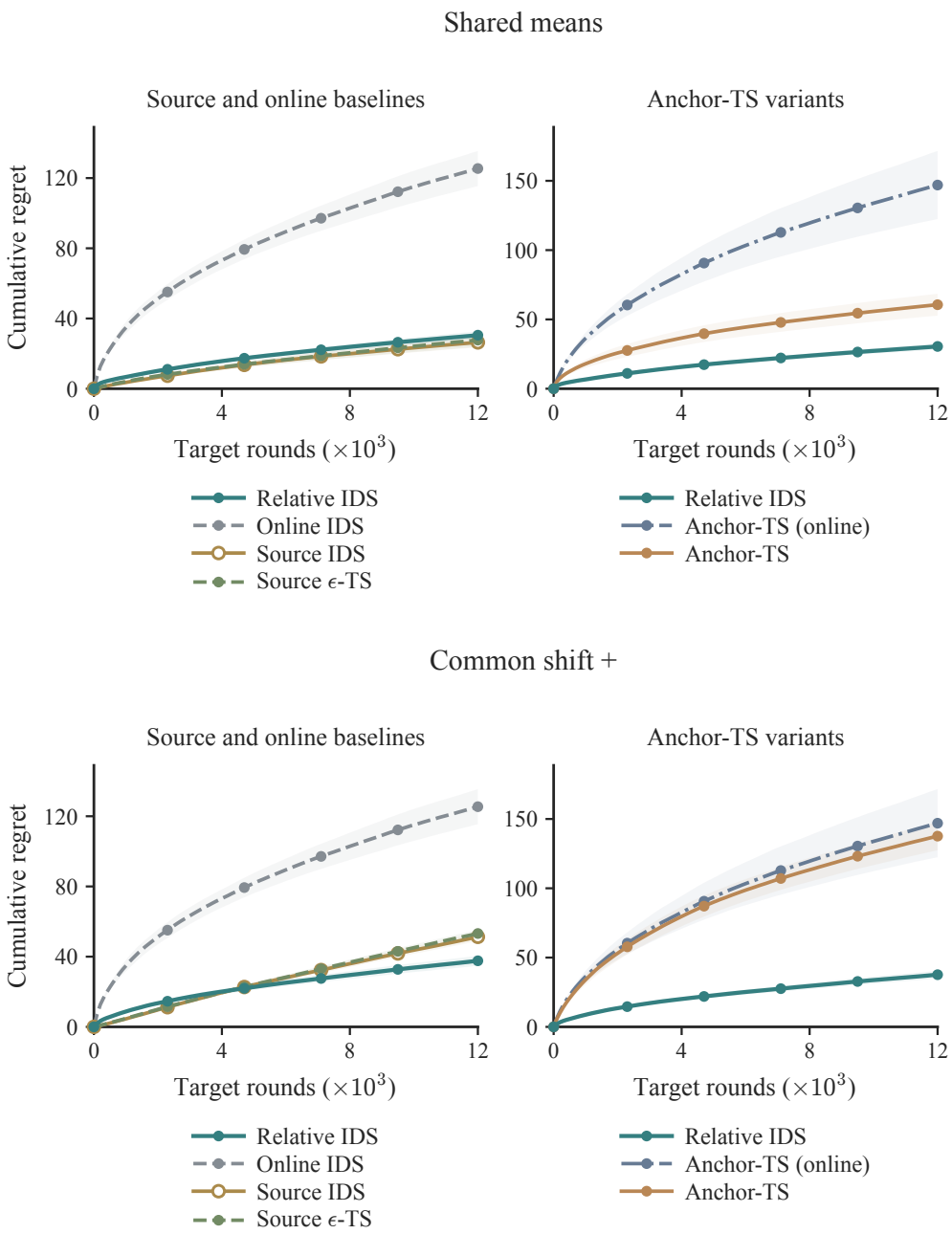


Figure 13: Learning curves under shared means (top) and a positive common shift (bottom). The original feedback worlds are reused; each curve averages nine gap–source-budget combinations within each of 32 paired seeds. Source and online controls appear on the left and Anchor-TS variants on the right. Shading shows pointwise approximate 95% intervals; vertical scales differ between panels.

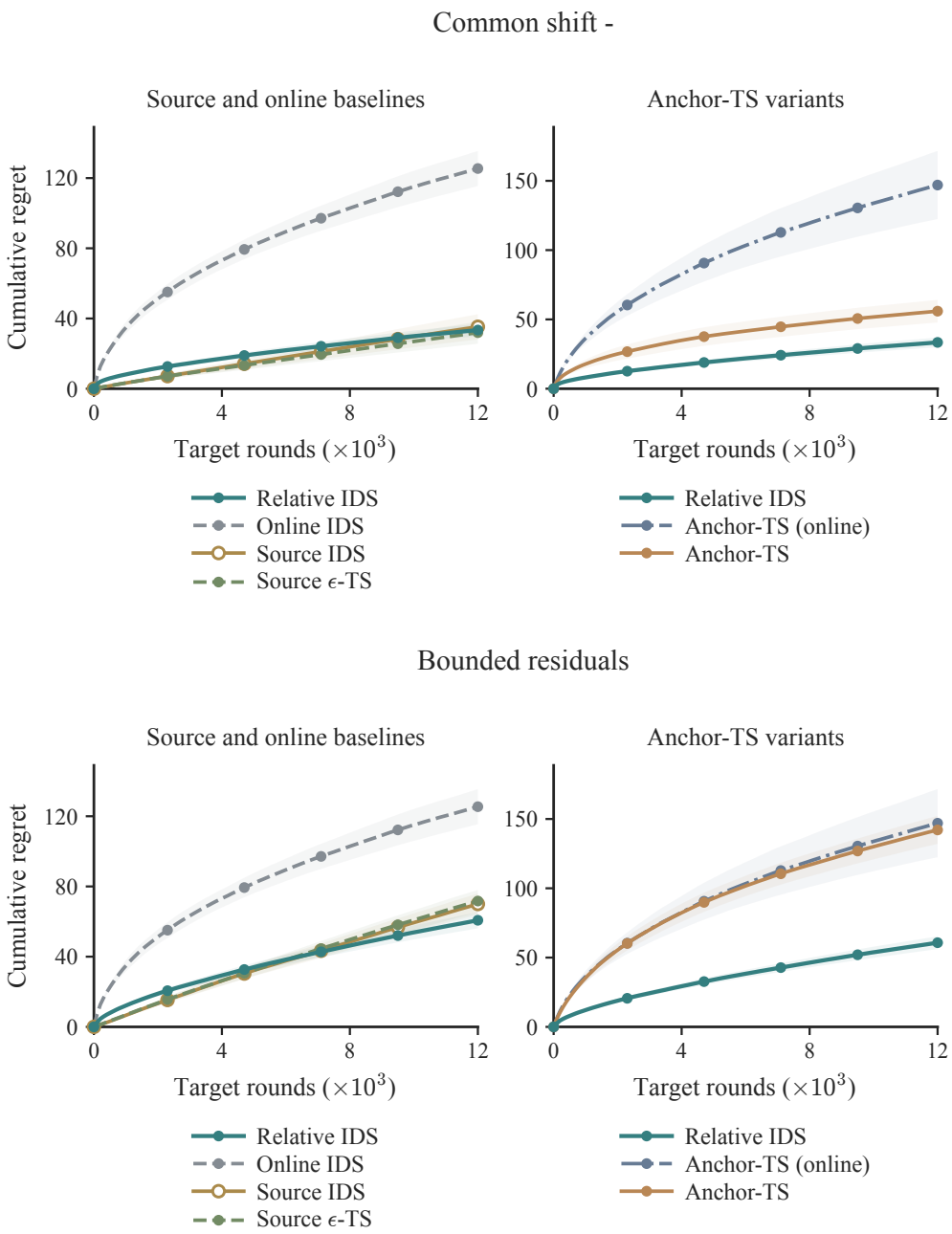


Figure 14: Learning curves under a negative common shift (top) and bounded residuals (bottom). The original feedback worlds are reused; each curve averages nine gap–source-budget combinations within each of 32 paired seeds. Source and online controls appear on the left and Anchor-TS variants on the right. Shading shows pointwise approximate 95% intervals; vertical scales differ between panels.

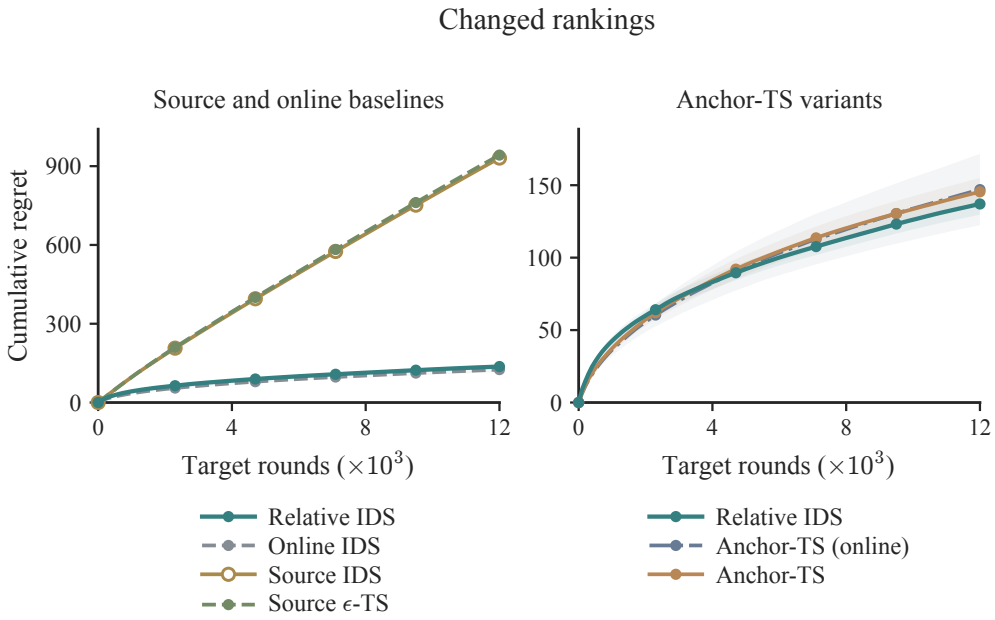


Figure 15: Learning curves under changed action rankings. The original feedback worlds are reused; each curve averages nine gap–source-budget combinations within each of 32 paired seeds. Source and online controls appear on the left and Anchor-TS variants on the right. Shading shows pointwise approximate 95% intervals; vertical scales differ between panels.

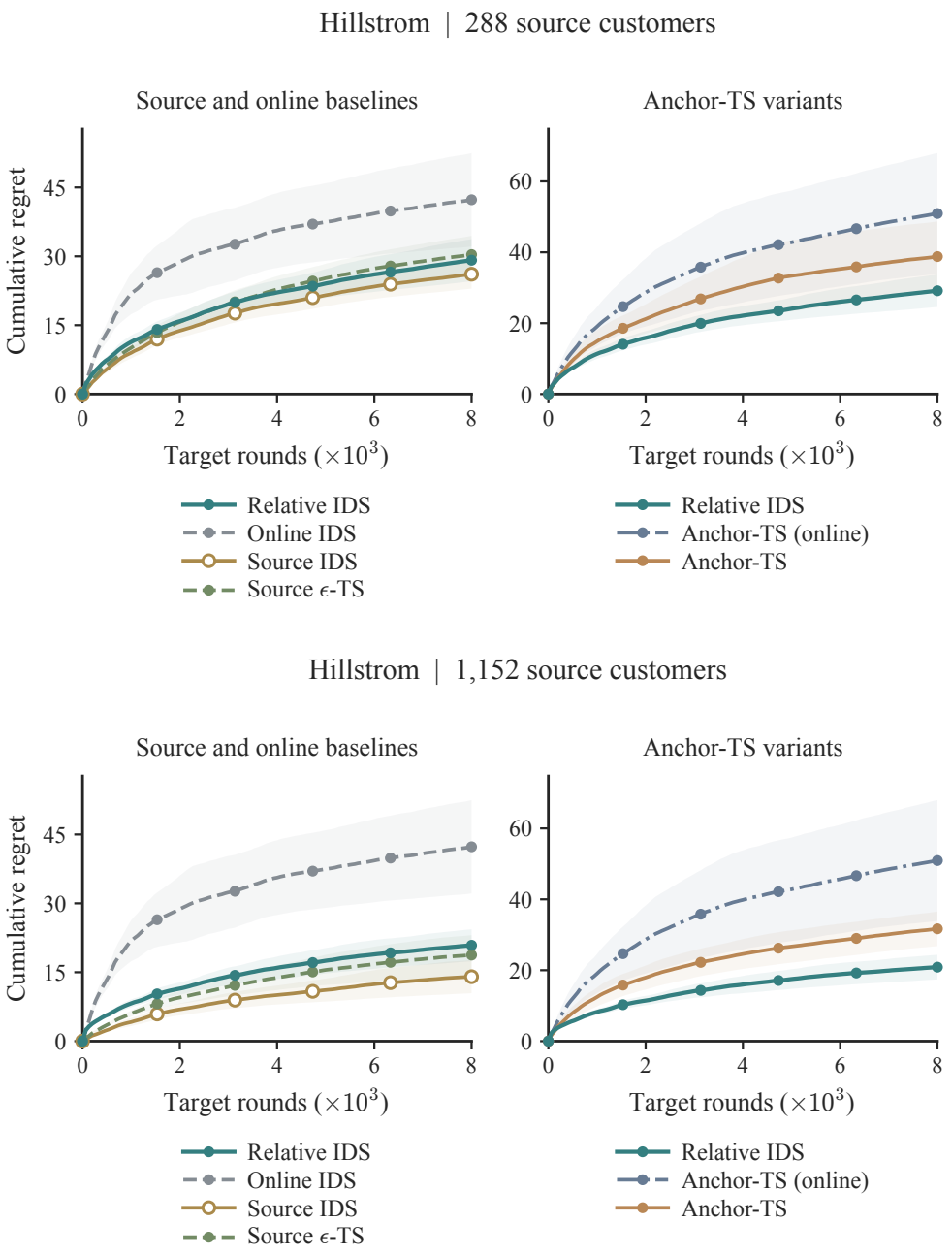


Figure 16: Hillstrom response-bank learning curves with 288 (top) and 1,152 (bottom) source customers. Curves average eight source draws within each of sixteen paired feedback seeds. Panels separate source and online controls from Anchor-TS variants. Shading shows pointwise approximate 95% intervals conditional on the fixed source subsamples and feedback bank; vertical scales differ between panels.

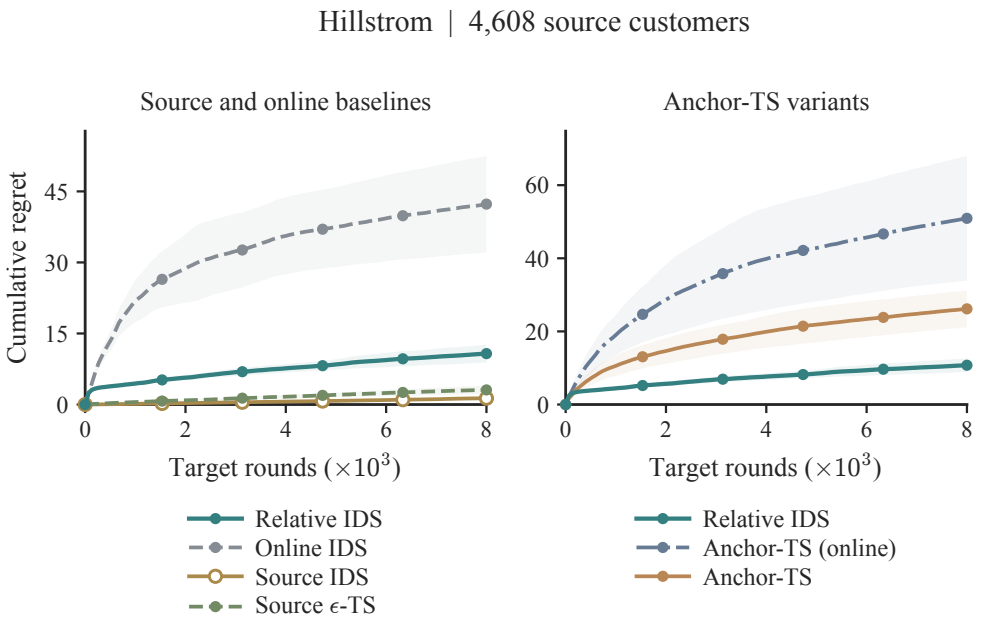


Figure 17: Hillstrom response-bank learning curves with 4,608 source customers. Curves average eight source draws within each of sixteen paired feedback seeds. Panels separate source and online controls from Anchor-TS variants. Shading shows pointwise approximate 95% intervals conditional on the fixed source subsamples and feedback bank; vertical scales differ between panels.

M.6 COMPLETE SETTING MEANS

The shared-model breakdown below uses the separate new-seed comparison with cyclic contexts. The five subsequent synthetic tables use the original iid-context worlds; their shared-mean regime is a distinct experiment. Online controls are identical across these five synthetic source regimes and are reported once in Table 20. All Hillstrom means, including the additional proposal controls, are already reported in Tables 16 and 18.

Table 19: Shared-model comparison, new seeds: complete setting means for our method and applicable controls. The Eq. (3) labels refer to OOPE's equation.

		Source per arm		
Method	Δ	128	2048	32768
Online IDS	.02	143.99	143.99	143.99
Online ϵ-TS, .75	.02	144.69	144.69	144.69
Online Beta-TS	.02	149.28	149.28	149.28
Exp-KL-MS	.02	151.72	151.72	151.72
LinMED	.02	145.83	145.83	145.83
Anchor-TS online, posterior, $\sigma = 1$	.02	140.63	140.63	140.63
Anchor-TS online, empirical, $\sigma = 1$	.02	142.24	142.24	142.24
Anchor-TS online, empirical, $\sigma = .5$	.02	150.24	150.24	150.24
E2D.Off, online	.02	176.57	176.57	176.57
Finite-class TS, online	.02	163.67	163.67	163.67
Source IDS	.02	119.69	42.60	0.00
Source ϵ-TS, $1/K$	.02	119.63	57.13	0.01
Source TS	.02	134.93	83.64	0.09
Source greedy	.02	142.50	54.38	0.00
MIN-UCB	.02	156.87	68.84	0.00
Directional MIN-UCB	.02	178.14	136.00	0.00
Anchor-TS source, posterior, $\sigma = 1$	.02	137.27	119.83	68.33
Anchor-TS source, empirical, $\sigma = 1$	.02	140.10	117.82	67.08
Anchor-TS source, empirical, $\sigma = .5$	.02	128.33	92.86	74.25
OOPE, $\sigma = 1$	.02	180.00	177.05	177.15
OOPE, $\sigma = .5$	.02	180.00	176.75	177.44
OOPE Eq.(3), $\sigma = .5$	.02	180.00	180.00	180.00
E2D.Off, source	.02	175.65	171.67	167.13
Finite-class TS, source	.02	157.01	101.56	0.01
Relative IDS	.02	116.71	48.88	5.35
Online IDS	.08	171.01	171.01	171.01
Online ϵ-TS, .75	.08	191.04	191.04	191.04
Online Beta-TS	.08	202.43	202.43	202.43
Exp-KL-MS	.08	270.50	270.50	270.50
LinMED	.08	256.14	256.14	256.14
Anchor-TS online, posterior, $\sigma = 1$	.08	184.48	184.48	184.48
Anchor-TS online, empirical, $\sigma = 1$	.08	200.68	200.68	200.68
Anchor-TS online, empirical, $\sigma = .5$	.08	230.33	230.33	230.33
E2D.Off, online	.08	530.72	530.72	530.72
Finite-class TS, online	.08	223.03	223.03	223.03
Source IDS	.08	41.39	0.00	0.00
Source ϵ-TS, $1/K$	.08	53.42	0.00	0.00
Source TS	.08	102.01	0.03	0.00
Source greedy	.08	127.50	0.00	0.00
MIN-UCB	.08	272.29	0.00	0.00
Directional MIN-UCB	.08	689.27	452.16	0.00
Anchor-TS source, posterior, $\sigma = 1$	.08	141.96	37.14	25.65
Anchor-TS source, empirical, $\sigma = 1$	.08	149.66	40.78	21.73
Anchor-TS source, empirical, $\sigma = .5$	.08	85.75	45.06	49.31

Continued on next page

Table 19: (continued)

Method	Δ	Source per arm 128	2048	32768
OOPE, $\sigma = 1$	.08	720.00	708.20	708.62
OOPE, $\sigma = .5$	.08	720.00	707.00	335.93
OOPE Eq.(3), $\sigma = .5$	.08	720.00	720.00	222.52
E2D.Off, source	.08	520.14	444.16	91.15
Finite-class TS, source	.08	129.65	0.10	0.00
Relative IDS	.08	59.85	12.44	0.00
Online IDS	.20	91.84	91.84	91.84
Online ϵ-TS, .75	.20	111.28	111.28	111.28
Online Beta-TS	.20	132.71	132.71	132.71
Exp-KL-MS	.20	169.11	169.11	169.11
LinMED	.20	233.14	233.14	233.14
Anchor-TS online, posterior, $\sigma = 1$	.20	130.59	130.59	130.59
Anchor-TS online, empirical, $\sigma = 1$	.20	143.28	143.28	143.28
Anchor-TS online, empirical, $\sigma = .5$	.20	100.98	100.98	100.98
E2D.Off, online	.20	883.39	883.39	883.39
Finite-class TS, online	.20	132.23	132.23	132.23
Source IDS	.20	0.18	0.00	0.00
Source ϵ-TS, $1/K$	.20	1.57	0.00	0.00
Source TS	.20	6.19	0.00	0.00
Source greedy	.20	0.00	0.00	0.00
MIN-UCB	.20	38.28	0.00	0.00
Directional MIN-UCB	.20	1580.32	1122.77	0.00
Anchor-TS source, posterior, $\sigma = 1$	.20	40.87	9.75	9.41
Anchor-TS source, empirical, $\sigma = 1$	.20	47.99	19.98	14.88
Anchor-TS source, empirical, $\sigma = .5$	.20	35.27	29.71	33.34
OOPE, $\sigma = 1$	.20	1800.00	1770.49	753.60
OOPE, $\sigma = .5$	.20	1800.00	1749.62	186.75
OOPE Eq.(3), $\sigma = .5$	.20	1785.40	1412.12	123.16
E2D.Off, source	.20	713.51	0.00	0.00
Finite-class TS, source	.20	6.49	0.00	0.00
Relative IDS	.20	24.79	0.00	0.00

Table 20: Shared means: complete setting means. Online controls are identical across all five source regimes; their budget-independent results are displayed once across the three source columns.

Method	Δ	Source per arm 128	2048	32768
Online IDS	.02		149.02	
Online ϵ-TS, .75	.02		137.09	
Online Beta-TS	.02		142.14	
Exp-KL-MS	.02		155.96	
LinMED	.02		143.14	
Anchor-TS online, posterior, $\sigma = 1$	.02		146.88	
Anchor-TS online, empirical, $\sigma = 1$	.02		141.36	
Anchor-TS online, empirical, $\sigma = .5$	.02		146.36	
Source IDS	.02	128.43	40.12	0.00
Source ϵ-TS, $1/K$	.02	120.95	50.69	0.00
Source TS	.02	143.53	74.47	0.02
Source greedy	.02	165.21	54.63	0.00
MIN-UCB	.02	157.27	68.13	0.00
Directional MIN-UCB	.02	178.09	133.09	0.00
Anchor-TS source, posterior, $\sigma = 1$	.02	137.56	117.63	62.93
Anchor-TS source, empirical, $\sigma = 1$	.02	144.21	118.09	67.85
Anchor-TS source, empirical, $\sigma = .5$	.02	125.59	95.81	69.52
Relative IDS	.02	123.91	44.72	5.33
Online ϵ-TS, $1/K$	.02		141.25	
Source ϵ-TS, .75	.02	137.06	67.37	0.01
Ablation: ϵ instead of IDS	.02	120.10	54.07	2.92
Online IDS	.08		147.38	
Online ϵ-TS, .75	.08		183.94	
Online Beta-TS	.08		206.37	
Exp-KL-MS	.08		280.77	
LinMED	.08		238.32	
Anchor-TS online, posterior, $\sigma = 1$	.08		206.38	
Anchor-TS online, empirical, $\sigma = 1$	.08		204.07	
Anchor-TS online, empirical, $\sigma = .5$	.08		182.99	
Source IDS	.08	68.73	0.00	0.00
Source ϵ-TS, $1/K$	.08	76.40	0.01	0.00
Source TS	.08	127.14	0.06	0.00
Source greedy	.08	270.44	0.00	0.00
MIN-UCB	.08	297.32	0.00	0.00
Directional MIN-UCB	.08	694.69	422.35	0.00
Anchor-TS source, posterior, $\sigma = 1$	.08	152.53	27.47	17.96
Anchor-TS source, empirical, $\sigma = 1$	.08	150.00	36.43	24.63
Anchor-TS source, empirical, $\sigma = .5$	.08	101.92	47.88	47.36
Relative IDS	.08	67.03	9.58	0.00
Online ϵ-TS, $1/K$	.08		206.32	
Source ϵ-TS, .75	.08	105.85	0.01	0.00
Ablation: ϵ instead of IDS	.08	69.55	7.33	0.00

Continued on next page

Table 20: (continued)

Method	Δ	Source per arm 128	2048	32768
Online IDS	.20	79.93		
Online ϵ-TS, .75	.20	108.51		
Online Beta-TS	.20	115.79		
Exp-KL-MS	.20	173.64		
LinMED	.20	235.86		
Anchor-TS online, posterior, $\sigma = 1$	.20	135.16		
Anchor-TS online, empirical, $\sigma = 1$	.20	141.08		
Anchor-TS online, empirical, $\sigma = .5$	.20	111.51		
Source IDS	.20	0.30	0.00	0.00
Source ϵ-TS, $1/K$	.20	1.14	0.00	0.00
Source TS	.20	5.03	0.00	0.00
Source greedy	.20	0.00	0.00	0.00
MIN-UCB	.20	50.24	0.00	0.00
Directional MIN-UCB	.20	1612.24	1066.64	0.00
Anchor-TS source, posterior, $\sigma = 1$	.20	41.40	8.44	7.62
Anchor-TS source, empirical, $\sigma = 1$	.20	49.23	12.24	11.24
Anchor-TS source, empirical, $\sigma = .5$	.20	21.83	17.08	18.46
Relative IDS	.20	23.91	0.00	0.00
Online ϵ-TS, $1/K$	.20	111.16		
Source ϵ-TS, .75	.20	4.32	0.00	0.00
Ablation: ϵ instead of IDS	.20	23.26	0.00	0.00

Table 21: Common shift +: source-assisted setting means. Identical online controls are reported once in Table 20.

Method	Δ	Source per arm 128	2048	32768
Source IDS	.02	139.00	137.24	0.04
Source ϵ-TS, $1/K$	.02	136.11	133.34	0.50
Source TS	.02	148.80	142.06	2.60
Source greedy	.02	144.38	60.09	0.00
MIN-UCB	.02	162.20	160.74	162.77
Directional MIN-UCB	.02	178.86	175.13	165.13
Anchor-TS source, posterior, $\sigma = 1$	.02	152.69	156.53	157.92
Anchor-TS source, empirical, $\sigma = 1$	.02	151.90	157.14	155.40
Anchor-TS source, empirical, $\sigma = .5$	.02	143.71	142.09	143.46
Relative IDS	.02	118.31	68.96	13.21
Source ϵ-TS, .75	.02	153.24	138.88	1.82
Ablation: ϵ instead of IDS	.02	118.82	73.49	18.58

Continued on next page

Table 21: (continued)

Method	Δ	Source per arm 128	2048	32768
Source IDS	.08	156.41	13.79	0.00
Source ϵ-TS, $1/K$	.08	159.03	23.94	0.00
Source TS	.08	238.07	58.06	0.00
Source greedy	.08	195.18	0.00	0.00
MIN-UCB	.08	377.47	381.86	373.55
Directional MIN-UCB	.08	697.75	619.45	655.17
Anchor-TS source, posterior, $\sigma = 1$	.08	253.53	323.46	321.47
Anchor-TS source, empirical, $\sigma = 1$	.08	263.69	314.06	318.27
Anchor-TS source, empirical, $\sigma = .5$	.08	165.88	189.11	193.26
Relative IDS	.08	81.44	21.31	0.00
Source ϵ-TS, .75	.08	233.81	47.94	0.00
Ablation: ϵ instead of IDS	.08	87.43	14.37	0.00
Source IDS	.20	15.24	0.00	0.00
Source ϵ-TS, $1/K$	.20	25.83	0.00	0.00
Source TS	.20	55.61	0.00	0.00
Source greedy	.20	0.00	0.00	0.00
MIN-UCB	.20	291.94	194.28	0.00
Directional MIN-UCB	.20	1642.21	1066.73	1642.04
Anchor-TS source, posterior, $\sigma = 1$	.20	164.51	188.08	191.91
Anchor-TS source, empirical, $\sigma = 1$	.20	171.17	191.58	190.84
Anchor-TS source, empirical, $\sigma = .5$	.20	82.89	89.80	88.45
Relative IDS	.20	35.63	0.00	0.00
Source ϵ-TS, .75	.20	47.75	0.00	0.00
Ablation: ϵ instead of IDS	.20	38.51	0.00	0.00

Table 22: Common shift -: source-assisted setting means. Identical online controls are reported once in Table 20.

Method	Δ	Source per arm 128	2048	32768
Source IDS	.02	148.41	56.74	0.00
Source ϵ-TS, $1/K$	.02	144.45	58.95	0.00
Source TS	.02	125.99	58.26	0.00
Source greedy	.02	149.69	58.28	0.00
MIN-UCB	.02	155.80	78.56	0.00
Directional MIN-UCB	.02	178.48	141.43	162.67
Anchor-TS source, posterior, $\sigma = 1$	.02	141.52	118.64	61.95
Anchor-TS source, empirical, $\sigma = 1$	.02	135.87	117.01	66.91
Anchor-TS source, empirical, $\sigma = .5$	.02	122.40	93.69	69.24
Relative IDS	.02	125.02	53.30	8.34
Source ϵ-TS, .75	.02	132.56	58.31	0.00
Ablation: ϵ instead of IDS	.02	117.11	61.46	6.36

Continued on next page

Table 22: (continued)

Method	Δ	Source per arm 128	2048	32768
Source IDS	.08	112.33	0.00	0.00
Source ϵ-TS, $1/K$	.08	84.07	0.00	0.00
Source TS	.08	55.07	0.00	0.00
Source greedy	.08	186.87	0.00	0.00
MIN-UCB	.08	281.56	0.00	0.00
Directional MIN-UCB	.08	696.65	363.61	649.45
Anchor-TS source, posterior, $\sigma = 1$	.08	152.59	26.84	18.37
Anchor-TS source, empirical, $\sigma = 1$	.08	142.43	36.18	23.03
Anchor-TS source, empirical, $\sigma = .5$	.08	75.89	44.22	42.29
Relative IDS	.08	65.01	15.32	0.00
Source ϵ-TS, .75	.08	64.32	0.00	0.00
Ablation: ϵ instead of IDS	.08	60.59	15.37	0.00
Source IDS	.20	0.00	0.00	0.00
Source ϵ-TS, $1/K$	.20	0.06	0.00	0.00
Source TS	.20	0.46	0.00	0.00
Source greedy	.20	0.00	0.00	0.00
MIN-UCB	.20	43.31	0.00	0.00
Directional MIN-UCB	.20	1626.51	896.01	1626.20
Anchor-TS source, posterior, $\sigma = 1$	.20	36.56	8.93	7.56
Anchor-TS source, empirical, $\sigma = 1$	.20	44.76	11.89	11.09
Anchor-TS source, empirical, $\sigma = .5$	.20	20.34	17.44	18.00
Relative IDS	.20	33.71	0.00	0.00
Source ϵ-TS, .75	.20	0.33	0.00	0.00
Ablation: ϵ instead of IDS	.20	34.43	0.00	0.00

Table 23: Bounded residuals: source-assisted setting means. Identical online controls are reported once in Table 20.

Method	Δ	Source per arm 128	2048	32768
Source IDS	.02	144.40	133.10	101.00
Source ϵ-TS, $1/K$	.02	142.72	131.91	101.00
Source TS	.02	153.63	137.48	102.64
Source greedy	.02	159.52	109.12	101.67
MIN-UCB	.02	160.77	163.61	163.23
Directional MIN-UCB	.02	178.53	178.07	176.47
Anchor-TS source, posterior, $\sigma = 1$	.02	154.53	158.08	156.51
Anchor-TS source, empirical, $\sigma = 1$	.02	153.34	154.83	155.97
Anchor-TS source, empirical, $\sigma = .5$	.02	141.64	140.02	140.92
Relative IDS	.02	134.62	108.63	109.38
Source ϵ-TS, .75	.02	151.90	136.62	102.38
Ablation: ϵ instead of IDS	.02	116.96	104.72	100.78

Continued on next page

Table 23: (continued)

Method	Δ	Source per arm 128	2048	32768
Source IDS	.08	160.73	64.24	6.74
Source ϵ-TS, $1/K$	.08	162.17	68.24	7.08
Source TS	.08	240.57	105.56	10.21
Source greedy	.08	263.69	22.52	14.78
MIN-UCB	.08	379.66	377.08	381.05
Directional MIN-UCB	.08	699.00	671.92	626.04
Anchor-TS source, posterior, $\sigma = 1$	.08	269.98	315.08	311.41
Anchor-TS source, empirical, $\sigma = 1$	.08	283.24	318.64	318.40
Anchor-TS source, empirical, $\sigma = .5$	.08	173.43	186.39	189.88
Relative IDS	.08	82.85	37.66	26.51
Source ϵ-TS, .75	.08	226.83	96.08	9.57
Ablation: ϵ instead of IDS	.08	96.54	40.86	30.87
Source IDS	.20	19.89	0.00	0.00
Source ϵ-TS, $1/K$	.20	32.08	0.00	0.00
Source TS	.20	61.33	0.00	0.00
Source greedy	.20	18.74	0.00	0.00
MIN-UCB	.20	304.45	294.46	161.54
Directional MIN-UCB	.20	1648.31	1471.84	1096.22
Anchor-TS source, posterior, $\sigma = 1$	.20	188.02	247.76	254.06
Anchor-TS source, empirical, $\sigma = 1$	.20	193.81	252.15	262.63
Anchor-TS source, empirical, $\sigma = .5$	.20	90.54	108.41	107.69
Relative IDS	.20	39.86	7.13	0.00
Source ϵ-TS, .75	.20	54.18	0.00	0.00
Ablation: ϵ instead of IDS	.20	42.53	4.22	0.00

Table 24: Changed rankings: source-assisted setting means. Identical online controls are reported once in Table 20.

Method	Δ	Source per arm 128	2048	32768
Source IDS	.02	234.56	240.00	240.00
Source ϵ-TS, $1/K$	.02	235.94	240.00	240.00
Source TS	.02	231.86	240.00	240.00
Source greedy	.02	240.00	240.00	240.00
MIN-UCB	.02	162.65	161.23	161.98
Directional MIN-UCB	.02	178.70	176.71	175.41
Anchor-TS source, posterior, $\sigma = 1$	.02	166.82	158.20	155.83
Anchor-TS source, empirical, $\sigma = 1$	.02	166.01	156.05	157.48
Anchor-TS source, empirical, $\sigma = .5$	.02	146.05	139.14	138.41
Relative IDS	.02	142.21	150.32	141.58
Source ϵ-TS, .75	.02	233.03	240.00	240.00
Ablation: ϵ instead of IDS	.02	142.57	143.36	139.69

Continued on next page

Table 24: (continued)

Method	Δ	Source per arm 128	2048	32768
Source IDS	.08	655.38	960.00	960.00
Source ϵ-TS, $1/K$	.08	701.79	960.00	960.00
Source TS	.08	705.65	960.00	960.00
Source greedy	.08	960.00	960.00	960.00
MIN-UCB	.08	379.01	380.10	381.35
Directional MIN-UCB	.08	698.41	688.01	680.63
Anchor-TS source, posterior, $\sigma = 1$	.08	339.11	315.97	320.60
Anchor-TS source, empirical, $\sigma = 1$	.08	334.37	321.81	323.04
Anchor-TS source, empirical, $\sigma = .5$	.08	181.28	187.77	189.37
Relative IDS	.08	191.09	153.20	173.44
Source ϵ-TS, .75	.08	688.14	960.00	960.00
Ablation: ϵ instead of IDS	.08	201.82	173.20	190.80
Source IDS	.20	303.13	2384.76	2400.00
Source ϵ-TS, $1/K$	.20	355.96	2386.46	2400.00
Source TS	.20	415.26	2380.02	2400.00
Source greedy	.20	2381.09	2400.00	2400.00
MIN-UCB	.20	302.69	313.93	312.24
Directional MIN-UCB	.20	1651.64	1604.75	1573.79
Anchor-TS source, posterior, $\sigma = 1$	.20	277.85	274.44	272.74
Anchor-TS source, empirical, $\sigma = 1$	.20	283.68	281.13	282.20
Anchor-TS source, empirical, $\sigma = .5$	.20	107.04	109.59	110.34
Relative IDS	.20	95.83	88.27	97.46
Source ϵ-TS, .75	.20	400.20	2382.80	2400.00
Ablation: ϵ instead of IDS	.20	108.49	109.56	104.44

N Real-Feedback Benchmarks Across Applications

We describe the response-bank construction used in Section F and report additional benchmarks of the same Relative IDS implementation on randomized treatment records and recommendation feedback. Each additional benchmark has two evaluation layers: learning with hidden, resampled factual feedback and evaluating the resulting fixed policies on separate factual records. The Hillstrom experiment in Table 16 has its separate source-customer protocol, given in Sections H and M. OBD contributes three campaigns from a second recommendation platform; these campaigns are one dataset, not three independent datasets.

N.1 Data, actions and population

Criteo Uplift. We use the corrected v2.1 release with 13979592 user records (Diemert et al., 2021). The action is the randomized advertising assignment `treatment`, not the post-assignment exposure indicator. Visits are the learning outcome. Medians of `f0` and `f2`, calculated on an independent design group, define at most four cells; cells with no design mass receive zero weight. The release combines trials and applies privacy-preserving label subsampling. Accordingly, our estimand is a standardized response in the released benchmark, not the original advertisers' population uplift. We do not substitute the observed treatment fraction for an unavailable original logging propensity.

International Stroke Trial. We use `IST_corrected.csv`, the corrected version-2 release of 19435 trial records (Sandercock et al., 2011). The two actions are the two values of the aspirin assignment `RXASP`; age below or above 75 and full alertness at randomization define four cells. The primary outcome is survival at the six-month follow-up, and the secondary outcome is full recovery at that follow-up. An unknown survival status is not coded as survival. Death implies no full recovery; other unknown recovery values remain missing. We use neither adherence variables nor predicted outcomes as pre-treatment features. The heparin assignment is not a baseline covariate.

Complete-outcome scores are accompanied by the range allowed by missing outcomes. This is a statistical benchmark, not a treatment recommendation.

KuaiRand. The Pure release contains roughly 27000 users, the complete random exposure log and normal recommendation records restricted to its fixed video pool (Gao et al., 2022). It is appropriate for our finite feedback-reservoir task, which does not require complete production trajectories. Source records are normal recommendations from April 8–21; target records are random exposures from April 22–May 8. We retain the two most frequent tabs among independent design users, breaking ties by tab index, and cross them with before/after noon. The four actions are empirical response reservoirs for videos of duration $(0, 15]$, $(15, 30]$, $(30, 60]$, and $(60, \infty)$ seconds. Thus an action selects a duration-stratified reservoir, not a specific video. The mixture of videos within each reservoir is part of the benchmark environment. The learning feedback is `long_view`; `is_like` provides a secondary evaluation outcome. We do not combine the different UI meanings of `is_click`, and we do not use full-month outcome aggregates as pre-decision features. Source, feedback-bank and reference users are disjoint. Standardized held-out log scores evaluate this reservoir contract; they are not production-policy off-policy evaluation (OPE) estimates for an arbitrary within-bucket recommendation rule.

Open Bandit Dataset. We use the full release of OBD (Saito et al., 2021), not the repository's small demonstration sample. We retain all items in each of the `all`, `men` and `women` campaigns, fixing position one in advance. The action is an item and the outcome is its recorded binary click. Source records come from the Bernoulli Thompson sampling (BTS) logger on November 24–26, 2019; feedback-bank records come from the random logger on November 27–28; reference records come from the random logger on November 29–30. We start with four context slots $h \in \{0, 1, 2, 3\}$, given by the first eight hexadecimal digits of `user_feature_0` modulo four. We then choose the finest supported partition in the fixed hierarchy h, $\lfloor h/2 \rfloor$, and a single group, where a partition is supported if each positive-weight group contains every action in the source, bank and reference records. This check uses covariate/action counts, not clicks or method rankings. Weights are source-window frequencies, with zero weight for empty slots. All records and all items are retained when groups are merged. These anonymized feature groups do not identify individual users.

Table 25: Additional benchmark populations and costs. Unit counts are distinct source/bank/reference customers, patients or users. N is the largest resampled source budget, not a count of new people. C is the paid target-calibration prefix for methods requiring bias bounds.

Dataset	Cells	Actions	Source	Bank	Reference	N	C
Criteo	2	2	4196508	4193411	4193499	2560	512
IST	4	2	5761	5803	5976	5120	1024
KuaiRand Pure	4	4	7456	8190	8247	10240	2048

Criteo has two positive-mass cells after the fixed median split. KuaiRand's rare tab has as few as one source record in a cell–action reservoir. Resampling that record supplies repeated feedback from the fixed empirical law, not new evidence about a population response. The full support counts accompany the code release.

Table 26: OBD support and costs. Source, bank and reference counts are actual position-one log rows, not independent users. All actions in each campaign are retained.

Campaign	Cells	Actions	Source	Bank	Reference	C
all	3	80	1653814	146931	127806	30720
men	3	34	585245	44730	41536	13056
women	2	46	1056790	93166	77457	11776

N.2 Feedback access and charged calibration

Before any policy is evaluated, records are partitioned using seed 143001. Criteo and IST assign individuals to the design, source, feedback-bank and reference splits with probabilities 10%, 30%,

30% and 30%. KuaiRand uses the same probabilities at the user level. Their context weights q_x are fixed from design metadata. OBD uses the policy/date partition and source-window weights specified above.

For a fixed cell–action, the source and target environments sample with replacement from their corresponding observed binary feedback reservoirs. Every sampled feedback value is linked to an original record. This defines Bernoulli laws with the empirical reservoir means, without making these means available to the learner or imputing any missing individual counterfactual. We report

$$R_T^{\mathrm{bank}} = \sum_{t=1}^{T} \left(\max_a \overline{Y}^{\mathrm{bank}}_{X_t,a} - \overline{Y}^{\mathrm{bank}}_{X_t,A_t} \right). \tag{22}$$

This bank regret measures learning in a fixed real-feedback environment; it is not observed regret from a newly deployed human experiment, and resampled draws are not counted as additional independent people.

For the additional benchmarks, each nonempty cell–action receives $b \in \{32, 128, 512\}$ source draws and 128 independent source-audit draws. All source-based comparators receive the combined $b + 128$ observations. The source-audit subset also calibrates the bias. For methods requiring a supplied bias bound, 128 target observations per nonempty cell–action form a round-robin calibration prefix, whose contexts follow a fixed balanced order. These observations are charged to the common $T = 8000$ target budget and update the learner. The remaining rounds draw contexts from q; every comparator, with or without a prefix, uses the same context stream and hidden potential feedback table. In Section F, by contrast, pilots wait for naturally arriving contexts, so regrets from the two protocols are not directly comparable. Pure-online methods and direct-source IDS, TS, ϵ-sampling and greedy methods choose actions autonomously from round one, without a forced calibration prefix. OBD uses $T = 64000$, because retaining all items with the same audit sample size requires a larger budget; its costs are given in Table 26, and its three campaigns are equally weighted only in the main summary row.

Let $\widehat{m}^s_{x,a}, \widehat{m}^t_{x,a}$ be the source and target audit means, $n_s = n_t = 128$, and $L = \log(4MK/0.01)$. Set

$$\begin{aligned}
e_s &= \sqrt{L/(2n_s)}, \qquad e_t = \sqrt{L/(2n_t)}, \\
\ell_{x,a} &= \max\{-1, \widehat{m}^s_{x,a} - \widehat{m}^t_{x,a} - e_s - e_t\}, \\
u_{x,a} &= \min\{1, \widehat{m}^s_{x,a} - \widehat{m}^t_{x,a} + e_s + e_t\}, \\
\widehat{B}_{x,a} &= \max\{|\ell_{x,a}|, |u_{x,a}|\}, \qquad \widehat{\beta}_x = \min\{2, \max_a u_{x,a} - \min_a \ell_{x,a}\}.
\end{aligned} \tag{23}$$

Conditional on the fixed reservoirs, a Hoeffding union bound over the $2MK$ audit means gives simultaneous coverage at least .99. On that event, $\widehat{B}$ bounds coordinate bias and $\widehat{\beta}_x$ bounds the range of source–target biases. These are computed from audit observations, not hidden target means. Relative IDS receives $\widehat{\beta}_x$ and uses $\max_a \widehat{B}_{x,a}$ as its shift scale b_c; MIN-UCB and source-assisted Anchor-TS receive $\widehat{B}$. Relative IDS and MIN-UCB use learning risk $.04/4$ in each cell. Reusing the source-audit observations in the source confidence intervals requires a union bound on the two events, not an independence claim between their resulting bounds. The post-calibration allocation is evaluated together with the prefix's actual regret; the prefix is not silently removed when comparing target-round efficiency. In particular, the theorem with an externally supplied β is not applied to the whole trajectory as if estimating that bound had been free.

We run 24 independently seeded source/feedback draws per budget, with seeds 143100–143123, and the 16 implementations listed in Tables 29 to 33. The IDS proposal and confidence code are unchanged. LinMED uses the valid Bernoulli mean-norm bound $\sqrt{K}$ rather than the synthetic model's $.75\sqrt{K}$; all other settings follow Section M. Because the real-feedback laws need not have uniform contexts, these experiments test allocation behavior rather than the precise minimax constants of the independent uniform-cell model. OOPE and E2D.Off are compared only in their declared shared-mean/realizable-class experiment in Section M; we do not construct a finite model class from hidden real-data means to manufacture realizability.

N.3 HELD-OUT REFERENCE EVALUATION

Only after all learning trajectories are fixed do we evaluate their policies on reference outcomes. For each budget and method, let $\bar{\pi}(a \mid x)$ average the 24 trajectories' within-cell action frequencies, including any calibration actions. For outcome j,

$$\widehat{V}_j(\bar{\pi}) = \sum_{x,a} q_x \bar{\pi}(a \mid x) \overline{Y}^{\mathrm{ref}}_{j,x,a}. \tag{24}$$

The reference score concerns the specified cells, outcomes and released populations. For the duration-stratified KuaiRand reservoirs it is a standardized log score, not an arbitrary recommendation policy's causal value.

Uncertainty is separated into two layers. First, paired-t intervals over the 24 simulation seeds measure Monte Carlo uncertainty conditional on the fixed reservoirs; they are pointwise, not population intervals or simultaneous bands. Second, for IST and KuaiRand, reference intervals use a unit-cluster sandwich covariance, with patients as units for IST and users for KuaiRand. Specifically, writing $n_{j,x,a}$ for the reference nonmissing count, the cluster contribution is

$$v_{u,j,x,a} = \frac{\sum_{i \in u: X_i = x, A_i = a} \mathbf{1}\{Y_{i,j} \text{ observed}\}(Y_{i,j} - \overline{Y}^{\mathrm{ref}}_{j,x,a})}{n_{j,x,a}}.$$

With U clusters, the joint covariance estimate is $\widehat{\Sigma} = \frac{U}{U-1} \sum_{u=1}^{U} v_u v_u^\top$. We use 10000 Gaussian multiplier draws and the .95 quantile of the largest standardized deviation over a fixed family of 210 statistics: the reference scores of 18 implementations and their 17 contrasts with Relative IDS, for three budgets and both outcomes. This family was fixed before two obsolete Anchor-TS configurations were removed and the four empirical-mean Anchor-TS configurations were rerun, and it was not recalibrated afterwards. Table 27 reports Relative IDS and its contrasts with Online IDS and Source IDS; these statistics are unchanged members of the family, so the displayed intervals keep the family's simultaneous level. Unlike the bounds in Sections I and J, this coverage is limited to the finite family; it gives no reference interval for the rerun Anchor-TS configurations. The resulting simultaneous reference intervals are asymptotic approximations, not the finite-sample certificates proved elsewhere in the paper. They condition on the trained policies and the chosen benchmark construction; they do not include uncertainty over hypothetical new source datasets or a different feature partition.

For missing reference outcomes, if a cell–action has s observed successes, n observed outcomes and m missing outcomes, its empirical response range is $[s/(n+m), (s+m)/(n+m)]$. Applying the nonnegative weights in (24) gives the reported missingness range; it is a separate sensitivity calculation, not a sampling interval.

Table 27: Independent reference evaluation at the largest source budget. Scores and differences are in percentage points. Intervals are approximate simultaneous 95% unit-cluster intervals, calibrated within each dataset over the fixed 210-statistic family of Section N.3. Positive differences favor Relative IDS. P/S denote the primary/secondary outcomes defined above.

Dataset	Outcome	Relative IDS score	vs. Online IDS	vs. Source IDS
IST	P	77.25 $[75.55, 78.95]$	-0.290 $[-0.690, +0.110]$	$+0.186$ $[-0.113, +0.486]$
IST	S	19.24 $[17.54, 20.94]$	-0.398 $[-0.950, +0.155]$	$+0.183$ $[-0.121, +0.486]$
KuaiRand Pure	P	10.05 $[9.50, 10.61]$	-0.289 $[-0.359, -0.219]$	$+0.957$ $[+0.768, +1.146]$
KuaiRand Pure	S	0.59 $[0.43, 0.75]$	-0.023 $[-0.040, -0.006]$	$+0.069$ $[+0.019, +0.119]$

The empirical missing-outcome ranges for Relative IDS are IST P: $[76.61, 77.40]$; IST S: $[18.93, 20.40]$. Other reported outcomes have no missing reference responses. These ranges address missingness, not sampling error.

Reference scores need not follow bank-regret rankings: the reference records have their own outcome means, and the evaluated Relative IDS policy includes calibration actions. On KuaiRand the two orderings agree, as Relative IDS scores 0.957 percentage points above Source IDS and 0.289 percentage points below Online IDS on the primary outcome, with both simultaneous intervals excluding zero; on IST, all four comparison intervals in Table 27 include zero.

OBD's fixed-record evaluation. OBD is evaluated conditional on the released feedback reservoirs. Within a cell–action, each interaction independently samples a record from its reservoir using fresh experiment randomness. This defines the empirical Bernoulli environment used in (22); it does not infer a new user's unobserved response to an item. The separate reference score uses (24) and the actual November 29–30 clicks. Its pointwise intervals are computed over the 24 training seeds, conditional on those fixed reference records. We do not treat individual log rows or the anonymized feature value as verified independent users and therefore do not attach a population cluster interval to OBD.

This evaluation also keeps logging identification distinct from learning in a feedback bank. The released random-logger propensity is marginally uniform at a position. Adjacent actions in the chronologically sorted logs exhibit substantial dependence, so the metadata alone does not establish conditional uniformity given the full chronological history. We therefore do not apply an adaptive chronological inverse propensity score (IPS) martingale certificate to these records. The independent sampling in our specified empirical environment, rather than a claim about the original logging filtration, justifies its Monte Carlo intervals and calibration calculation.

Table 28: OBD reference click scores at $b = 512$. Values and differences are in percentage points. Pointwise intervals measure only training-seed Monte Carlo uncertainty conditional on the fixed reference records; they are not population or causal OPE intervals. Positive differences favor Relative IDS.

Campaign	Relative IDS score	vs. Online IDS	vs. Source IDS
all	0.473 $[0.401, 0.545]$	$+0.042$ $[-0.034, +0.118]$	-0.041 $[-0.122, +0.039]$
men	0.501 $[0.464, 0.539]$	-0.001 $[-0.051, +0.049]$	-0.112 $[-0.163, -0.060]$
women	0.620 $[0.587, 0.654]$	$+0.066$ $[+0.029, +0.104]$	-0.007 $[-0.038, +0.024]$

N.4 All Methods and Source Budgets

In Tables 29 to 33, the three N columns correspond to $b = 32$, 128 and 512; N is the total number of source and audit draws, $b + 128$ per nonempty cell–action. At the largest source budget, Relative IDS has mean bank regret 33.79 on IST and 291.66 on OBD women, compared with 54.26 and 376.62 for Online IDS. On KuaiRand, its mean bank regret is 87.90 versus 156.45 for Source IDS, a benefit over direct source reuse in this setting. All these values include the calibration prefix. Splitting the observed trajectories at the end of the calibration prefix, over the same rounds for every method, locates the remaining gaps. Relative IDS has 12.97 more regret than Source IDS on IST and 62.28 more on OBD women, of which 9.60 and 60.34 arise in the prefix window; on OBD all, whose prefix takes 30,720 of 64,000 rounds, it has 270.44 more regret than Source IDS in that window and 56.40 less afterwards. On KuaiRand, the gap of 24.89 to Online IDS arises in both windows (9.32 and 15.56). The tables report all 16 methods at each budget; Figs. 18 and 19 show the learning curves at $b = 512$.

Table 29: IST: response-bank regret at $T = 8000$. Means and pointwise 95% Monte Carlo intervals use 24 seeds. N includes all available source/audit draws; online methods use none. Methods that require a bias bound spend 1024 target rounds on calibration; these rounds are included in regret.

Method	$N = 1280$	$N = 2048$	$N = 5120$
Relative IDS	37.10 [27.04, 47.17]	35.61 [26.97, 44.25]	33.79 [25.18, 42.39]
Online IDS	54.26 [38.26, 70.26]	54.26 [38.26, 70.26]	54.26 [38.26, 70.26]
Source IDS	29.94 [22.62, 37.26]	27.97 [20.62, 35.32]	20.82 [11.57, 30.07]
Online Beta-TS	43.24 [31.65, 54.82]	43.24 [31.65, 54.82]	43.24 [31.65, 54.82]
Exp-KL-MS	52.08 [46.57, 57.58]	52.08 [46.57, 57.58]	52.08 [46.57, 57.58]
LinMED	58.12 [42.94, 73.29]	58.12 [42.94, 73.29]	58.12 [42.94, 73.29]
Source TS	41.89 [35.76, 48.02]	45.09 [36.64, 53.54]	29.89 [23.84, 35.94]
Source ϵ-TS, $1/K$	30.34 [24.01, 36.67]	34.61 [27.09, 42.14]	23.69 [18.08, 29.30]
Source greedy	66.50 [43.89, 89.11]	74.17 [50.47, 97.87]	28.72 [9.40, 48.04]
MIN-UCB	60.57 [55.69, 65.46]	60.57 [55.69, 65.46]	60.57 [55.69, 65.46]
Anchor-TS source, posterior, $\sigma = 1$	53.13 [47.54, 58.72]	53.46 [47.96, 58.95]	53.25 [47.74, 58.76]
Anchor-TS source, empirical, $\sigma = 1$	54.06 [48.51, 59.62]	53.64 [48.01, 59.28]	54.46 [48.25, 60.66]
Anchor-TS source, empirical, $\sigma = .5$	43.74 [36.51, 50.97]	43.73 [36.49, 50.97]	43.73 [36.49, 50.97]
Anchor-TS online, posterior, $\sigma = 1$	49.50 [40.12, 58.87]	49.50 [40.12, 58.87]	49.50 [40.12, 58.87]
Anchor-TS online, empirical, $\sigma = 1$	40.41 [33.26, 47.57]	40.41 [33.26, 47.57]	40.41 [33.26, 47.57]
Anchor-TS online, empirical, $\sigma = .5$	43.29 [26.91, 59.67]	43.29 [26.91, 59.67]	43.29 [26.91, 59.67]

Table 30: KuaiRand Pure: response-bank regret at $T = 8000$. Means and pointwise 95% Monte Carlo intervals use 24 seeds. N includes all available source/audit draws; online methods use none. Methods that require a bias bound spend 2048 target rounds on calibration; these rounds are included in regret.

Method	$N = 2560$	$N = 4096$	$N = 10240$
Relative IDS	77.45 [61.66, 93.24]	87.27 [70.56, 103.97]	87.90 [69.67, 106.12]
Online IDS	63.01 [51.47, 74.55]	63.01 [51.47, 74.55]	63.01 [51.47, 74.55]
Source IDS	140.52 [136.28, 144.76]	149.61 [146.88, 152.35]	156.45 [154.44, 158.46]
Online Beta-TS	79.24 [67.35, 91.14]	79.24 [67.35, 91.14]	79.24 [67.35, 91.14]

Continued on next page

Table 30: (continued)

Method	$N = 2560$	$N = 4096$	$N = 10240$
Exp-KL-MS	100.92 [92.21, 109.63]	100.92 [92.21, 109.63]	100.92 [92.21, 109.63]
LinMED	71.72 [50.77, 92.68]	71.72 [50.77, 92.68]	71.72 [50.77, 92.68]
Source TS	144.99 [141.61, 148.36]	151.17 [148.86, 153.49]	157.56 [155.84, 159.29]
Source ϵ-TS, $1/K$	138.47 [133.91, 143.04]	145.77 [142.63, 148.92]	156.75 [154.36, 159.13]
Source greedy	201.29 [184.14, 218.43]	196.39 [179.47, 213.32]	214.85 [205.47, 224.23]
MIN-UCB	140.17 [137.32, 143.02]	140.17 [137.32, 143.02]	140.17 [137.32, 143.02]
Anchor-TS source, posterior, $\sigma = 1$	130.02 [124.84, 135.19]	129.97 [124.77, 135.16]	129.97 [124.77, 135.16]
Anchor-TS source, empirical, $\sigma = 1$	130.27 [124.97, 135.56]	130.23 [124.91, 135.54]	130.23 [124.91, 135.54]
Anchor-TS source, empirical, $\sigma = .5$	106.89 [96.82, 116.95]	106.89 [96.82, 116.95]	106.89 [96.82, 116.95]
Anchor-TS online, posterior, $\sigma = 1$	102.18 [91.23, 113.13]	102.18 [91.23, 113.13]	102.18 [91.23, 113.13]
Anchor-TS online, empirical, $\sigma = 1$	95.52 [84.64, 106.40]	95.52 [84.64, 106.40]	95.52 [84.64, 106.40]
Anchor-TS online, empirical, $\sigma = .5$	77.31 [60.56, 94.06]	77.31 [60.56, 94.06]	77.31 [60.56, 94.06]

Table 31: OBD all, position 1, all 80 actions: bank regret at $T = 64000$. Means and pointwise 95% Monte Carlo intervals use 24 seeds. Methods that require a bias bound pay 30720 target rounds for calibration; the cost is included.

Method	$N = 38400$	$N = 61440$	$N = 153600$
Relative IDS	705.65 [661.92, 749.38]	703.70 [658.58, 748.82]	794.08 [738.86, 849.31]
Online IDS	799.60 [772.66, 826.54]	799.60 [772.66, 826.54]	799.60 [772.66, 826.54]
Source IDS	556.48 [463.06, 649.90]	467.96 [374.37, 561.55]	580.05 [479.37, 680.72]
Online Beta-TS	950.34 [931.71, 968.97]	950.34 [931.71, 968.97]	950.34 [931.71, 968.97]
Exp-KL-MS	923.72 [906.44, 941.00]	923.72 [906.44, 941.00]	923.72 [906.44, 941.00]
LinMED	635.29 [567.58, 702.99]	635.29 [567.58, 702.99]	635.29 [567.58, 702.99]
Source TS	625.45 [557.11, 693.79]	521.97 [423.76, 620.19]	560.64 [469.46, 651.81]
Source ϵ-TS, $1/K$	558.29 [458.45, 658.14]	625.15 [501.48, 748.82]	690.16 [568.39, 811.93]
Source greedy	1314.03 [1253.20, 1374.86]	1201.55 [1086.62, 1316.48]	1221.88 [1129.86, 1313.91]

Continued on next page

Table 31: (continued)

Method	$N = 38400$	$N = 61440$	$N = 153600$
MIN-UCB	1368.88 [1367.66, 1370.11]	1368.88 [1367.66, 1370.11]	1368.88 [1367.66, 1370.11]
Anchor-TS source, posterior, $\sigma = 1$	1369.36 [1368.01, 1370.71]	1370.36 [1369.11, 1371.60]	1370.29 [1369.01, 1371.57]
Anchor-TS source, empirical, $\sigma = 1$	1369.35 [1368.00, 1370.69]	1370.38 [1369.13, 1371.64]	1370.31 [1369.03, 1371.60]
Anchor-TS source, empirical, $\sigma = .5$	1329.85 [1326.30, 1333.40]	1329.77 [1326.17, 1333.37]	1329.77 [1326.17, 1333.37]
Anchor-TS online, posterior, $\sigma = 1$	1346.60 [1343.98, 1349.21]	1346.60 [1343.98, 1349.21]	1346.60 [1343.98, 1349.21]
Anchor-TS online, empirical, $\sigma = 1$	1347.58 [1344.68, 1350.48]	1347.58 [1344.68, 1350.48]	1347.58 [1344.68, 1350.48]
Anchor-TS online, empirical, $\sigma = .5$	1230.69 [1213.55, 1247.83]	1230.69 [1213.55, 1247.83]	1230.69 [1213.55, 1247.83]

Table 32: OBD men, position 1, all 34 actions: bank regret at $T = 64000$. Means and pointwise 95% Monte Carlo intervals use 24 seeds. Methods that require a bias bound pay 13056 target rounds for calibration; the cost is included.

Method	$N = 16320$	$N = 26112$	$N = 65280$
Relative IDS	379.76 [360.61, 398.92]	378.24 [358.88, 397.61]	360.39 [335.61, 385.18]
Online IDS	388.39 [372.36, 404.42]	388.39 [372.36, 404.42]	388.39 [372.36, 404.42]
Source IDS	336.44 [322.31, 350.58]	349.91 [332.68, 367.14]	314.03 [287.24, 340.82]
Online Beta-TS	457.63 [449.56, 465.70]	457.63 [449.56, 465.70]	457.63 [449.56, 465.70]
Exp-KL-MS	457.47 [448.06, 466.89]	457.47 [448.06, 466.89]	457.47 [448.06, 466.89]
LinMED	366.38 [331.82, 400.94]	366.38 [331.82, 400.94]	366.38 [331.82, 400.94]
Source TS	431.85 [415.39, 448.30]	432.19 [418.57, 445.81]	398.91 [380.63, 417.20]
Source ϵ-TS, $1/K$	335.34 [308.21, 362.47]	321.50 [285.87, 357.13]	323.27 [284.61, 361.93]
Source greedy	709.12 [639.11, 779.13]	718.03 [650.91, 785.15]	729.36 [648.39, 810.33]
MIN-UCB	645.29 [644.58, 645.99]	645.29 [644.58, 645.99]	645.29 [644.58, 645.99]
Anchor-TS source, posterior, $\sigma = 1$	645.55 [644.83, 646.27]	646.29 [645.70, 646.89]	646.08 [645.48, 646.68]
Anchor-TS source, empirical, $\sigma = 1$	645.52 [644.78, 646.25]	646.31 [645.72, 646.90]	646.08 [645.48, 646.68]
Anchor-TS source, empirical, $\sigma = .5$	630.63 [629.17, 632.10]	629.96 [628.53, 631.39]	629.94 [628.50, 631.37]

Continued on next page

Table 32: (continued)

Method	$N = 16320$	$N = 26112$	$N = 65280$
Anchor-TS online, posterior, $\sigma = 1$	632.63 [631.05, 634.22]	632.63 [631.05, 634.22]	632.63 [631.05, 634.22]
Anchor-TS online, empirical, $\sigma = 1$	631.25 [629.28, 633.21]	631.25 [629.28, 633.21]	631.25 [629.28, 633.21]
Anchor-TS online, empirical, $\sigma = .5$	587.69 [582.76, 592.63]	587.69 [582.76, 592.63]	587.69 [582.76, 592.63]

Table 33: OBD women, position 1, all 46 actions: bank regret at $T = 64000$. Means and pointwise 95% Monte Carlo intervals use 24 seeds. Methods that require a bias bound pay 11776 target rounds for calibration; the cost is included.

Method	$N = 14720$	$N = 23552$	$N = 58880$
Relative IDS	342.25 [317.65, 366.85]	322.24 [297.79, 346.69]	291.66 [260.65, 322.67]
Online IDS	376.62 [356.41, 396.82]	376.62 [356.41, 396.82]	376.62 [356.41, 396.82]
Source IDS	296.19 [274.31, 318.07]	277.53 [249.40, 305.65]	229.39 [192.49, 266.28]
Online Beta-TS	429.27 [413.54, 444.99]	429.27 [413.54, 444.99]	429.27 [413.54, 444.99]
Exp-KL-MS	427.25 [416.50, 437.99]	427.25 [416.50, 437.99]	427.25 [416.50, 437.99]
LinMED	239.18 [204.64, 273.73]	239.18 [204.64, 273.73]	239.18 [204.64, 273.73]
Source TS	390.52 [376.70, 404.34]	370.19 [351.60, 388.79]	315.61 [288.32, 342.90]
Source ϵ-TS, $1/K$	235.49 [193.73, 277.26]	247.69 [203.70, 291.69]	234.21 [189.29, 279.13]
Source greedy	506.92 [433.22, 580.63]	491.62 [415.53, 567.71]	449.46 [375.73, 523.20]
MIN-UCB	559.74 [559.15, 560.34]	559.74 [559.15, 560.34]	559.74 [559.15, 560.34]
Anchor-TS source, posterior, $\sigma = 1$	559.38 [558.96, 559.79]	559.79 [559.28, 560.31]	559.63 [559.09, 560.18]
Anchor-TS source, empirical, $\sigma = 1$	559.38 [558.96, 559.79]	559.80 [559.29, 560.31]	559.64 [559.09, 560.18]
Anchor-TS source, empirical, $\sigma = .5$	549.18 [548.07, 550.30]	548.90 [547.78, 550.02]	548.89 [547.77, 550.02]
Anchor-TS online, posterior, $\sigma = 1$	553.00 [551.68, 554.32]	553.00 [551.68, 554.32]	553.00 [551.68, 554.32]
Anchor-TS online, empirical, $\sigma = 1$	552.40 [551.50, 553.31]	552.40 [551.50, 553.31]	552.40 [551.50, 553.31]
Anchor-TS online, empirical, $\sigma = .5$	529.97 [526.67, 533.27]	529.97 [526.67, 533.27]	529.97 [526.67, 533.27]

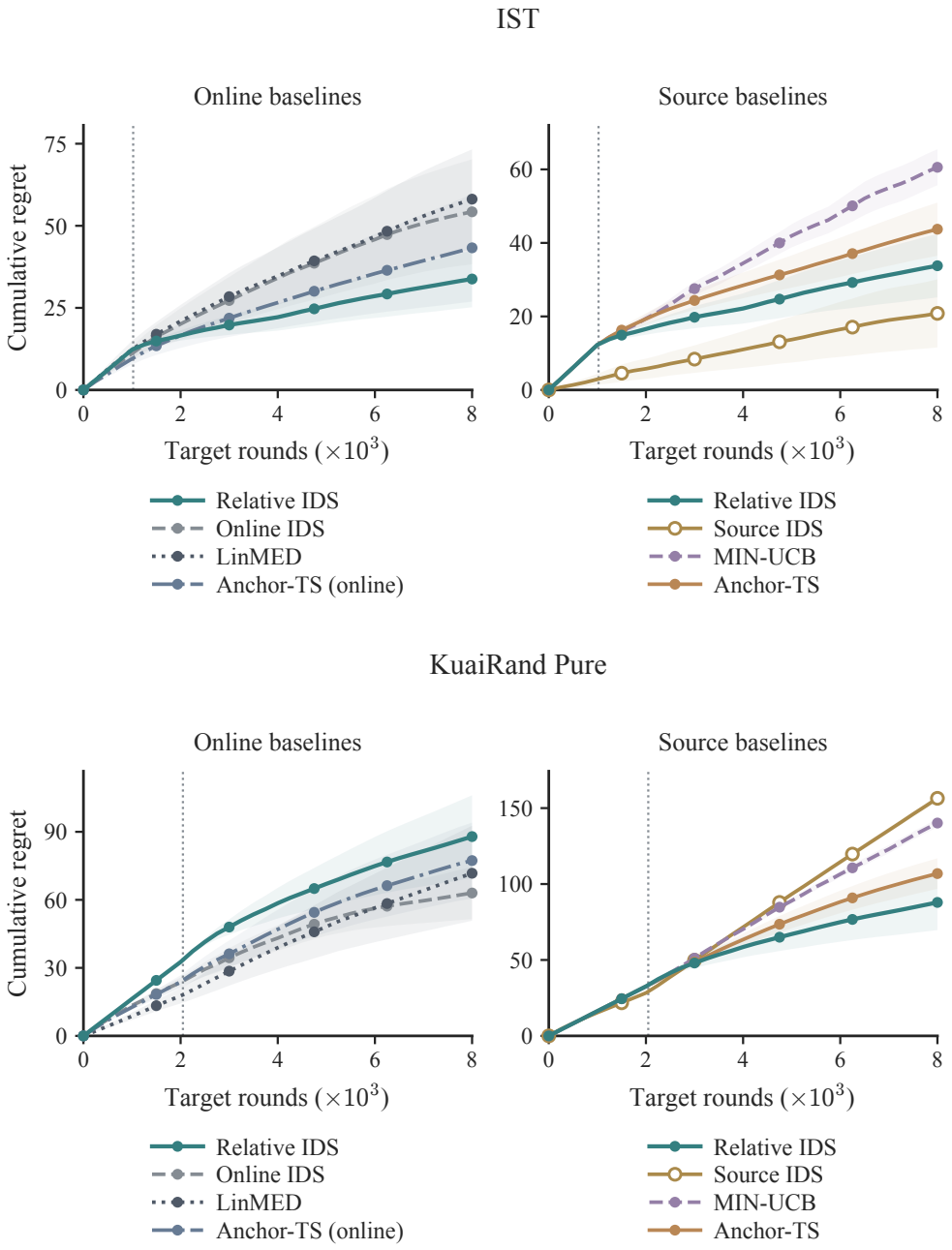


Figure 18: Learning curves on IST (top) and KuaiRand Pure (bottom) at $T = 8000$, with $b = 512$ plus the declared source audit. Columns compare Relative IDS with online baselines (left) and source baselines (right). Curves include calibration regret; dotted vertical lines mark the end of the paid calibration prefix. Shading shows pointwise approximate 95% Monte Carlo intervals over 24 paired seeds, conditional on each fixed bank. Vertical scales differ between panels. Anchor-TS uses empirical means with $\sigma = .5$. The preceding tables give all method and budget results.

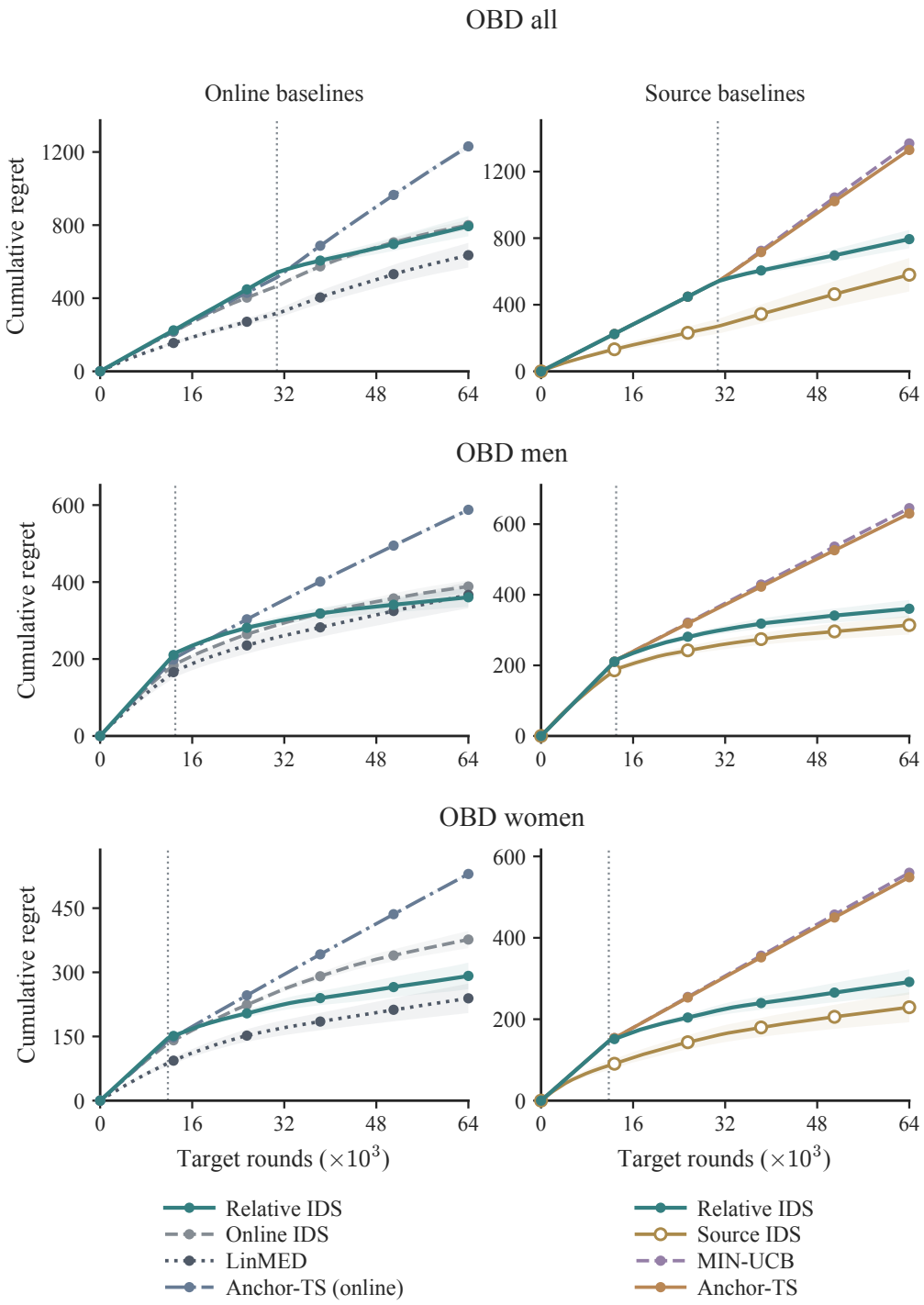


Figure 19: Learning curves on OBD all, men and women (top to bottom) at $T = 64000$, with $b = 512$ plus the declared source audit. Columns compare Relative IDS with online baselines (left) and source baselines (right), with one shared legend per column. Curves include calibration regret; dotted vertical lines mark the end of the paid calibration prefix. Shading shows pointwise approximate 95% Monte Carlo intervals over 24 paired seeds, conditional on each fixed bank. Vertical scales differ between panels. Anchor-TS uses empirical means with $\sigma = .5$. The preceding tables give all method and budget results.