FAST A/B-TESTING: EFFICIENT POLICY EVALUATION VIA TREE-COUPLED FEEDBACK SHARING

Anonymous authors
Paper under double-blind review

ABSTRACT

Online platforms increasingly compare many adaptive decision policies—ranking systems, recommendation algorithms, pricing rules, and language-model agents—while each reward-bearing interaction can be costly or risky. The standard recipe is a direct A/B/n test, which runs each of J candidate policies over an independent T-horizon and therefore uses JT outcomes. We introduce Tree-Coupled A/B Testing (`TCAB`), an exact feedback-sharing design for reducing the outcome usage while preserving the marginal trajectories. At a high level, the policies are associated to a tree structure and are executed by any topological order. The observed outcome is shared along an edge when the two policies on this edge can be coupled. Every policy retains exactly its standalone finite-horizon trajectory law, even though the policies are deliberately dependent. If $D_{e,t}$ records a mismatch on tree edge e at round t, the number of outcomes satisfies the pathwise identity $N(T) = T + \sum_{t,e} D_{e,t} \leq JT$. This cost is conditionally optimal for the selected tree, and we propose a heuristic for empirically finding the optimal tree. For fixed J, sublinear pseudo-regret of every policy can imply $\mathbb{E}[N(T)] = T + o(T)$. We also obtain finite-sample variance bounds for pairwise policy contrasts. Experiments on language-model evaluation and adaptive bandit policies demonstrate substantial improvements—around 20-50% outcome save and improved policy evaluation precision.

1 INTRODUCTION

Online controlled experiments are central to production infrastructure at major digital platforms. Google, Microsoft/Bing, and LinkedIn have described systems that support overlapping experiments and continuous product iteration (Tang et al., 2010; Kohavi et al., 2013; Xu et al., 2015). A cross-industry summit reported that the thirteen participating organizations had collectively tested more than one hundred thousand treatment variants in the preceding year (Gupta et al., 2019). In modern machine-learning evaluation, product teams compare many different recommender-system configurations, ranking policies, prompting strategies, large-language-model (LLM) agents, and model checkpoints, which unavoidably requires massive expensive human feedback (Chiang et al., 2024).

The classical A/B/n experiment compares J candidate policies over a target horizon T by running J independent trajectories and thereby invokes JT feedback interactions.[1] This cost can be operationally important when an experimental interaction can reduce user experience, consume expert labels, invoke a costly model, or simply delay a deployment decision. We refer the readers to (Kohavi et al., 2009; 2020; Larsen et al., 2024; Quin et al., 2024) for a complete survey of this classical methodology and its extensive applications in the digital industry.

There is nevertheless a large source of redundancy. Candidate policies generated by neighboring hyperparameters, model checkpoints, or business rules frequently make the same decision on the same context. When two policies realize the same complete context–action pair, they require the same conditional reward law and can therefore receive the same physical outcome. Purposefully coupling their trajectories can remove duplicated reward noise and save an interaction. The challenge is to do so without changing either policy's adaptive trajectory distribution.

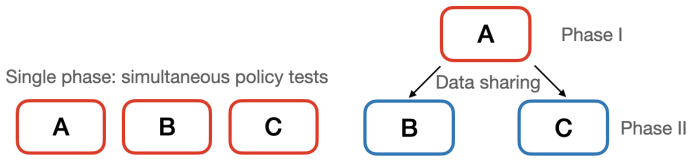
[1]Throughout this work, interactions and queries are used interchangeably.

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.

Contextuality makes this challenge fundamentally different from replaying a past observation. Even under i.i.d. contexts, the realized contexts attached to the observations of a selected action can rarely repeat later, disabling straightforward feedback reuse. Moreover, in applications such as ranked slates, auctions, and recommendation, the outcome of choosing one item may depend on the entire displayed slate or user state, not only on the chosen item's local feature. We therefore consider a full-context reward kernel $Q_a(\cdot \mid x)$ and couple the policies' complete one-step *context–action laws*, not the (context, action) tuple itself. A maximal coupling in our case matches the two complete pairs with the largest probability permitted by their total-variation distance.

Moving from two policies to $J \geq 3$ creates a second challenge. Pairwise maximal couplings need not be jointly compatible: in general there is no single coupling that maximizes the equality probability of every pair simultaneously (Angel & Spinka, 2019). We resolve this incompatibility with a policy tree. Pairwise couplings prescribed on the $J - 1$ edges of an acyclic graph can always be glued into one joint law. A mismatched edge starts a new reward lineage; matched edges transmit the same (context, action, reward) triple. Consequently, the number of actual reward queries is exactly one plus the number of mismatched edges in each round.

Our algorithm, Tree-Coupled A/B Testing (`TCAB`), is round-synchronous. It uses either a fixed policy tree throughout or a time-varying tree. At each round, the algorithm samples the root policy, traverses the tree by depth, and transmits outcomes through matched edges. It still exposes practical parallelism: all children at the same depth can be run concurrently once their parents are available. Adaptive policies remain sequential across rounds, as they must, while nonadaptive policies can additionally parallelize across rounds.



(a) Standard A/B/C test (b) Tree-based with data sharing

Figure 1: In standard A/B tests, policies are run independently and simultaneously. In `TCAB`, the policies are run in phases per round, which are determined by the tree, and reuse data from previous phases.

The resulting guarantees are finite-horizon and model-free. Every policy trajectory has exactly its standalone law, so all values and policy contrasts are unbiased. The expected query cost is T plus cumulative total variation (TV) over the tree edges. This identity yields a precise design principle: a baseline-centered star is simple and makes every baseline comparison maximal, while an oracle minimum-spanning tree minimizes the *current-round* cost among edge-local tree designs.

Contributions. Our main contributions are as follows.

1. We formulate exact finite-horizon comparison of J arbitrary history-dependent, possibly randomized contextual policies under i.i.d. full contexts and nonparametric reward kernels $Q_a(\cdot \mid x)$.
2. We propose `TCAB` and prove the pathwise and expected cost identities

$$N(T) = T + \sum_{t=1}^{T} \sum_{e \in E_t} D_{e,t}, \qquad \mathbb{E}[N(T)] = T + \sum_{t=1}^{T} \mathbb{E}\left[\sum_{e \in E_t} \delta_{e,t} \right],$$

   where $D_{e,t} \in \{0, 1\}$ denotes the mismatch, and $\delta_{e,t}$ denotes the history-dependent TV distance. Within the natural class of conditionally exact edge-local designs, `TCAB` is optimal on every selected tree. A baseline star and a calibrated minimum-spanning tree (MST) give two concrete specializations.
3. We establish a finite-sample regret-to-cost bound for potentially time-varying tree sequences. For fixed J, almost-sure uniqueness of the oracle action and $o(T)$ pseudo-regret for every policy imply $T + o(T)$ expected reward queries.
4. We prove finite-sample variance bounds for every pairwise policy contrast. The bounds extend to time-varying trees and isolate the roles of edge mismatches and realized pseudo-regret.

5. Across two language-model evaluation tasks and an adaptive search-bandit task, TCAB consistently improves the empirical cost–precision trade-off relative to independent A/B/n baselines.

2 Related literature

Large-scale experimentation and comparison of learning policies. Industrial experimentation systems at Google, Bing, and LinkedIn motivate the need to compare many candidate changes under limited experimental traffic (Tang et al., 2010; Kohavi et al., 2013; Xu et al., 2015; Gupta et al., 2019). The closest methodological work is Meng et al. (2026), whose Artificial Replay construction compares two *context-free* stochastic bandit algorithms with exact marginal distributions. Our problem faces two distinct difficulties: (1) decisions are contextual, so feedback cannot be shared straightforwardly from another policy's history; crucially, the tuple (context, action) may rarely repeat, and direct reuse is extremely inefficient. Instead, we match the context–action laws rather than the realized tuples themselves. (2) $J \geq 3$ policies cannot in general be made pairwise maximally coupled simultaneously.

Contextual replay and off-policy evaluation. Li et al. (2011) give exact replay for a contextual-bandit algorithm from a uniformly randomized log, while inverse-propensity, doubly robust, and related methods evaluate policies from fixed logged data (Dudík et al., 2011; Swaminathan & Joachims, 2015; Wang et al., 2017; Su et al., 2020; Zhan et al., 2021). These approaches begin with an exogenous logging distribution and hence require suitable overlap or weighting. In contrast, TCAB prospectively constructs coupled trajectories of runnable candidate policies: insufficient overlap increases the number of reward queries but does not change the target policy law.

Coupling multiple policies. For two laws, maximal coupling achieves disagreement probability equal to TV distance (Thorisson, 2000; Lindvall, 2002), and coupled rejection sampling gives an exact implementation under density-ratio access (Corenflos & Särkkä, 2022). With three or more marginals, simultaneous pairwise maximality can fail (Angel & Spinka, 2019). TCAB exploits the complementary fact that pairwise couplings specified on the edges of a tree are jointly compatible. For interested readers, Appendix A discusses multiple-comparison/adaptive-inference work, multi-marginal transport, policy-selection work, and contextual-bandit regret in more details.

3 Problem formulation

In this section, we formalize the problem setup and the target we aim to estimate through the controlled experiment.

3.1 Stochastic contextual environment

Let $T \geq 1$ be the target horizon, $\mathcal{A} = [K]$ the action space, and $\mathcal{X}$ the context space. Contexts arrive as

$$X_t \sim P_X \quad \text{i.i.d. over } t. \tag{3.1}$$

For a slate, X_t may contain the user and all candidate-item features; changing feasible actions can likewise be encoded in X_t. For each action $a \in \mathcal{A}$, a reward query at the context-action (x, a) returns a draw from the underlying kernel $Q_a(\cdot \mid x)$, with mean $\mu_a(x) := \int r\, Q_a(\,\mathrm{d}r \mid x)$.

Assumption 3.1 (I.i.d. stationary rewards)**.** Conditional on the queried context–action pairs, reward queries are independent draws from their corresponding kernels $Q_a(\cdot \mid x)$.

3.2 Candidate policies and estimands

Fix $J \geq 2$ policies indexed by $\mathcal{I} := \{0, 1, \ldots, J-1\}$. Policy j is formally defined as a sequence of measurable kernels $\pi_{j,t}(\cdot \mid h, x) \in \Delta(\mathcal{A})$ with history $H^j_{t-1} = (X^j_s, A^j_s, R^j_s)_{s<t}$; any persistent internal state affecting future decisions is included in this state. In a standalone run, one simply draws

$$X^j_t \sim P_X, \qquad A^j_t \sim \pi_{j,t}(\cdot \mid H^j_{t-1}, X^j_t), \qquad R^j_t \sim Q_{A^j_t}(\cdot \mid X^j_t). \tag{3.2}$$

Define the T-horizon policy values as

$$S_j(T) := \sum_{t=1}^{T} R_t^j, \qquad V_j(T) := \mathbb{E}[S_j(T)], \qquad V(T) := (V_0(T), \dots, V_{J-1}(T))^\top. \tag{3.3}$$

For a baseline policy r, let the value contrast be

$$\theta^{(r)}(T) := \big(V_j(T) - V_r(T)\big)_{j \neq r}. \tag{3.4}$$

Independent runs spend JT reward queries. We seek exact horizon-T trajectories for all policies using a random total $N(T) \leq JT$, while preserving unbiased estimation of these values and baseline contrasts. To this end, we will make the following assumption: through the experiment, we are allowed to observe the context of an arriving user, evaluate the candidate policies, and then decide whether to route the user into the experiment or normal traffic. If a user query is determined to be routed to normal traffic, it is not affected by the experiment and hence will not be counted in $N(T)$.

Assumption 3.2 (Operational access)**.** Before requesting a reward, the experiment can draw an unlabeled $x \sim P_X$, evaluate the relevant policy kernels at x, and then decides whether to (1) reject it without updating either policy, or (2) accept it and request a reward.

Neither the density of P_X nor the reward kernel is evaluated in this assumption. We only count reward-bearing experimental interactions. For readers interested in the number of users we will reject in `TCAB`, Proposition 4.1 shows that the expected number of rejections is at most $(J-1)T$.

4 Tree-Coupled A/B Testing

The `TCAB` algorithm is formally described in Algorithm 1. It allows the experiment to use either a fixed policy tree or a time-varying one. For generality, let $\mathcal{T}_t = (\mathcal{I}, E_t)$ denote the policy tree at the start of round t. Write $p_t(v)$ and $\mathrm{dep}_t(v)$ for the parent and depth of a nonroot vertex v respectively.

Recall that $\mathcal{X} \times \mathcal{A}$ denotes the context-action space, (X_t^j, A_t^j) the context-action pair for policy j at the t-th round of its trajectory. Conditional on policy history h_j, define the context–action joint law

$$\nu_{j,t}^{h_j}(\,\mathrm{d}x, a) := P_X(\,\mathrm{d}x)\, \pi_{j,t}(a \mid h_j, x). \tag{4.1}$$

For policies i, j, denote the history-conditional TV distance of their joint laws by

$$\delta_{ij,t}(h_i, h_j) := \mathrm{TV}\Big(\nu_{i,t}^{h_i}, \nu_{j,t}^{h_j}\Big) = \frac{1}{2} \int_{\mathcal{X}} \|\pi_{i,t}(\cdot \mid h_i, x) - \pi_{j,t}(\cdot \mid h_j, x)\|_1 \, P_X(\,\mathrm{d}x). \tag{4.2}$$

A maximal coupling of these two laws matches their complete pairs with probability $1 - \delta_{ij,t}$. At round t, `TCAB` samples the root policy r from its law $\nu_{r,t}^{H_{t-1}^r}$ conditioned on the history H_{t-1}^r and recursively maximally couples each child to its parent: When the parent policy i admits the triple (X_t^i, A_t^i, R_t^i), the child policy j will inherit this feedback and reuse $R_t^j = R_t^i$ with probability

$$q_{ij,t}(X_t^i, A_t^i) = \min\left\{1, \frac{\pi_{j,t}(A_t^i \mid H_{t-1}^j, X_t^i)}{\pi_{i,t}(A_t^i \mid H_{t-1}^i, X_t^i)}\right\}. \tag{4.3}$$

Otherwise, instead of naively accepting a new sample, the child policy must match the *residual* of its own standalone law. Operationally, the child will repeatedly draw proposals

$$X \sim P_X, A \sim \pi_{j,t}(\cdot \mid H_{t-1}^j, X), \tag{4.4}$$

and accept with probability

$$s_{ij,t}(X, A) := 1 - \min\left\{1, \frac{\pi_{i,t}(A \mid H_{t-1}^i, X)}{\pi_{j,t}(A \mid H_{t-1}^j, X)}\right\}. \tag{4.5}$$

Proposition 4.1 (Proposal overhead)**.** *Let $U_{ij,t}$ denote the number of proposed context-action before one is accepted by (4.5). For an edge (i, j) with $\delta_{ij,t} > 0$, conditional on entering the residual branch, the number of child proposals is geometric with mean $1/\delta_{pv,t}$. Since the residual branch itself is entered with probability $\delta_{pv,t}$,*

$$\mathbb{E}[U_{ij,t}] = \begin{cases} 1, & \text{if } \delta_{ij,t} > 0; \\ 0, & \text{if } \delta_{ij,t} = 0. \end{cases}$$

Hence the total expected residual-proposal overhead is at most $(J-1)T$.

Algorithm 1 Tree-Coupled A/B Testing (TCAB)

Require: Horizon T, policies $(\pi_j)_{j\in\mathcal{I}}$, context sampler, reward interface

Initialize $H_0^j \leftarrow \varnothing$ for every $j \in \mathcal{I}$
for $t = 1, \dots, T$ **do**
  Select policy tree $\mathcal{T}_t = (\mathcal{I}, E_t)$ with root r_t
  Draw $(X_t^{r_t}, A_t^{r_t}) \sim \nu_{r_t,t}^{H_{t-1}^{r_t}}$ and reward $R_t^{r_t}$
  for depth $d = 1, \dots, \max_v \mathrm{dep}_t(v)$ **do**
    In parallel over all v with $\mathrm{dep}_t(v) = d$, generate (X_t^v, A_t^v, R_t^v) from the edge sampler (4.3–4.5) using parent $p_t(v)$
  Simultaneously append (X_t^j, A_t^j, R_t^j) to every policy history
return values $(S_j(T))_{j\in\mathcal{I}}$

The proposition separates two notions of cost. Small TV means that reward sharing is frequent, but a rare residual event can require many screened proposals conditional on occurring. Their product is controlled: the expected number of residual context proposals per active edge-round is at most one. Our objective $N(T)$ nevertheless counts only reward-bearing experimental interactions, because those are the resource that feedback sharing is designed to reduce.

Tree choices. We use two concrete variants with fixed trees in the experiments. First, `TCAB.STAR` fixes a star centered at a designated baseline policy, consisting of $J-1$ baseline–alternative edges. Second, if all current pairwise distances in (4.2) were known, a minimum-spanning tree (MST) would minimize the *current-round* sum of tree-edge distances by Theorem 5.3. In practice, we use `TCAB.MST` that estimates fixed pairwise scores from pilot data and freezes the resulting MST before the main experiment. These two choices are discussed in more details in Section 5.2.

5 Exactness and Reward-Query Efficiency

This section establishes the main guarantees of TCAB. We show that TCAB maintains the exact trajectory distributions with provably less queries and vanishing variance. The subsections are organized as follows: we first show in Theorem 5.1 that the generated trajectory for each policy admits the exact marginal distribution from a standalone run as if it is run independently, then we characterize the expected query cost of TCAB in Theorem 5.3 which is asymptotically no-extra-cost when the candidate policies are no-regret by Theorem 5.5; finally, Theorem 5.7 provides the variance analysis and shows when the variance vanishes.

To state these results, for each edge $e = \{i, j\} \in E_t$, define

$$D_{e,t} := \mathbf{1}\{(X_t^i, A_t^i) \neq (X_t^j, A_t^j)\}, \quad N_t := \text{number of queries at round } t, \quad N(T) := \sum_{t=1}^{T} N_t.$$

The feedback inheritence in TCAB gives

$$N_t = 1 + \sum_{e\in E_t} D_{e,t}, \qquad N(T) = T + \sum_{t=1}^{T}\sum_{e\in E_t} D_{e,t} \quad \text{pathwise.} \tag{5.1}$$

Since every tree has $J-1$ edges,

$$T \le N(T) \le JT \quad \text{almost surely.} \tag{5.2}$$

5.1 Exact Trajectory Marginals and Unbiased Contrasts

As a first guarantee for TCAB, we show that our tree-coupled feedback sharing design does not alter the marginal trajectory distributions for the candidate policies.

Theorem 5.1 (Exact marginal trajectories)**.** *Under Assumption 3.1 and 3.2, for every finite T, J, every policy tree sequence $(\mathcal{T}_t)_{t\le T}$, and every collection of measurable history-dependent policies, the TCAB trajectory*

$$H_T^j = (X_t^j, A_t^j, R_t^j)_{t=1}^T$$

has exactly the same distribution as the standalone process in (3.2), for every policy $j \in \mathcal{I}$.

Theorem 5.1 shows that, while the generated trajectories are dependent across policies, no individual policy can statistically distinguish its TCAB trajectory from a standalone run. Consequently, the trajectory-based estimands are unbiased.

Corollary 5.2 (Unbiased values and baseline contrasts)**.** *If rewards have finite first moments, then*

$$\mathbb{E}[S_j(T)] = V_j(T), \qquad \mathbb{E}[S_j(T) - S_i(T)] = V_j(T) - V_i(T)$$

for every $j, i \in \mathcal{I}$.

5.2 Exact query cost

Due to the residual sampling we use in (equation 4.3), it is rather straightforward to show that, at round t and edge $\{i, j\}$, a reward observation is inherited with the probability $1 - \delta_{ij,t} = 1 - \mathrm{TV}(\nu_{i,t}^{H_{t-1}^i}, \nu_{j,t}^{H_{t-1}^j})$, where H_{t-1}^i and H_{t-1}^j are the realized trajectories of policy i and j up to round t. Then we have the following exact characterization of the query cost for TCAB.

Theorem 5.3 (Conditional cost identity and edge-local optimality)**.** *For TCAB,*[2]

$$\mathbb{E}[N_t \mid \mathcal{F}_{t-1}] = 1 + \sum_{e \in E_t} \delta_{e,t}, \quad \textit{and} \quad \mathbb{E}[N(T)] = T + \mathbb{E}\left[\sum_{t=1}^{T} \sum_{e \in E_t} \delta_{e,t}\right]. \tag{5.3}$$

It is clear from Theorem 5.3 that $\mathbb{E}[N(T)]$ crucially depends on the choice of the trees $(\mathcal{T}_t)_t$. For simplicity, we use a fixed policy tree through our numerical experiments in Section 6, with the following two variants. Nonetheless, it is beneficial if one can choose the tree $\mathcal{T}_t$ as an approximate minimizer of the TV characterization $\arg\min_{\mathcal{T}_t} \sum_{(i,j) \in E_t} \delta_{ij,t}$ at each round.

Star tree In practice, there is typically a baseline policy r that has been run for a long time, and the experiment aims to compare other $J - 1$ alternatives to it. A simple heuristic is to fix r as the root and $(r, j) \in E$ as the only edges. Additionally, given the offline logs of the baseline r, it is possible to reconstruct the trajectory of the root from the offline data in TCAB, which shaves the linear term T in the cost in (5.3). This gives the `TCAB.STAR` variant in Section 6.

Corollary 5.4 (Baseline-centered star)**.** *For the fixed star rooted at r, $\mathbb{E}[N(T)] = T + \sum_{t=1}^{T} \sum_{j \neq r} \mathbb{E}[\delta_{rj,t}]$.*

TV-minimizing tree Given the characterization of $\mathbb{E}[N(T)]$ in (5.3), the optimal fixed tree is

$$\arg\min_{\mathcal{T}} \mathbb{E}\left[\sum_{t=1}^{T} \sum_{(i,j) \in E} \delta_{ij,t}\right]. \tag{5.4}$$

The history-dependent TV is unknown. However, if pilot data is available, one can calibrate the policy tree by evaluating the policies offline and empirically approximating the objective in (5.4). Then the empirical minimizer $\widehat{\mathcal{T}}_*$ to (5.4) is fixed throughout the experiment. Note that the root needs not to be the baseline in practice. This gives the `TCAB.MST` variant in Section 6.

5.3 Low regret implies near-one-trajectory cost

For learning policies, pseudo-regret is a popular metric that measures the cumulative reward difference between the policy of interest and the hindsight optimal oracle. This subsection shows that, if the candidate policies are "good" in the sense of achieving low regrets, then the query cost $\mathbb{E}[N(T)]$ of TCAB can collapse to T asymptotically. The idea is intuitive: no-regret policies eventually converge to selecting the hindsight optimal action, thus we are able to share most of the feedback in TCAB. For general policies with potentially non-vanishing regrets, Theorem 5.5 bounds the query cost in the policy regrets.

[2] We use $\delta_{e,t}$ to denote the TV $\delta_{ij,t}$ for edge $e = \{i, j\}$.

Assume rewards lie in $[0,1]$ for this and the next subsection. Define the optimal mean reward $\mu_*(x) := \max_{a\in\mathcal{A}} \mu_a(x)$ and the reward gap $\Delta_a(x) := \mu_*(x) - \mu_a(x)$. When the maximizer is unique, let the maximizer and the min reward gap be $a^*(x) := \arg\max_a \mu_a(x)$ and $\Delta(x) := \min_{a\neq a^*(x)} \Delta_a(x)$. Then policy j's pseudo-regret is $\mathcal{R}_j(T) := \mathbb{E}\Big[\sum_{t=1}^T \Delta_{A_t^j}(X_t^j)\Big]$. Let $d_{j,t}$ be the degree of j in tree $\mathcal{T}_t$ at round t, and define the degree-weighted regret

$$\mathcal{R}_{\deg}(T) := \mathbb{E}\left[\sum_{t=1}^T \sum_{j\in\mathcal{I}} d_{j,t}\Delta_{A_t^j}(X_t^j)\right] \leq (J-1)\sum_{j\in\mathcal{I}} \mathcal{R}_j(T) \tag{5.5}$$

since $d_{j,t} \leq J-1$.

Theorem 5.5 (Regret-to-cost bound)**.** *Suppose the optimal action $a^*(X)$ is unique P_X-almost surely and define $F_\Delta(\epsilon) := \mathbb{P}\{\Delta(X) \leq \epsilon\}$. For every predictable tree sequence $\{\mathcal{T}_t\}_t$ and every $\epsilon > 0$,*

$$\mathbb{E}[N(T)] - T \leq 2(J-1)TF_\Delta(\epsilon) + \frac{\mathcal{R}_{\deg}(T)}{\epsilon}. \tag{5.6}$$

Consequently, if $\mathcal{R}_j(T) = o(T)$ for every policy j, then

$$\mathbb{E}[N(T)] = T + o(T). \tag{5.7}$$

Almost-sure uniqueness is needed only to force $F_\Delta(\epsilon) \to 0$ as $\epsilon \downarrow 0$. Without a common tie-breaking rule, two zero-regret policies can disagree indefinitely on a positive-probability tie set.

The following result gives (1) an explicit rate when the margin $F_\Delta(\epsilon) \to 0$ at some exponential rate and (2) if there is a bounded gap $\Delta_{\min}$ between the best and the second best actions, the total queries $N(T)$ is bounded by the degree-weighted regret.

Corollary 5.6 (Margin and uniform-gap rates)**.** *If $F_\Delta(\epsilon) \leq C\epsilon^\beta$ for all sufficiently small ϵ, with constants $C, \beta > 0$, then*

$$\mathbb{E}[N(T)] - T = O\Big(\big((J-1)T\big)^{1/(\beta+1)}\mathcal{R}_{\deg}(T)^{\beta/(\beta+1)}\Big). \tag{5.8}$$

If $\Delta(X) \geq \Delta_{\min} > 0$ P_X-almost surely, then

$$\mathbb{E}[N(T)] - T \leq \frac{\mathcal{R}_{\deg}(T)}{\Delta_{\min}}. \tag{5.9}$$

5.4 Variance reduction for policy contrasts

Now we consider the variance of the policy contrast estimators from `TCAB`. For policy j, define the realized regret $L_j(T) := \sum_{t=1}^T \Delta_{A_t^j}(X_t^j)$. For policies i, j, let $P_{ij,t}$ denote their unique path in the round-t tree $\mathcal{T}_t$ and set $B_{ij}(T) := \mathbb{E}\Big[\sum_{t=1}^T \sum_{e\in P_{ij,t}} \delta_{e,t}\Big]$. The the variance of the estimated policy contrasts from `TCAB` depends on the variance of the policies' regret variance and $B_{ij}(T)$:

Theorem 5.7 (Pairwise-contrast variance)**.** *For every policy pair $i, j \in \mathcal{I}$,*

$$\mathrm{Var}(S_j(T) - S_i(T)) \leq 15B_{ij}(T) + 6\mathrm{Var}(L_i(T)) + 6\mathrm{Var}(L_j(T)). \tag{5.10}$$

In particular, if every policy admits $o(T)$ expected regret and $o(T)$ regret variance, then

$$T^{-1}\mathrm{Var}(S_j(T) - S_i(T)) \longrightarrow 0.$$

Theorem 5.7 is the main inferential guarantee: a round contributes no reward noise to the i–j comparison whenever the entire tree path between the two policies matches. Any fixed zero-sum linear comparison can be written as a finite linear combination of pairwise contrasts; Appendix F.1 records the resulting general extension.

6 NUMERICAL EXPERIMENTS

This section evaluates the empirical cost–precision trade-off of `TCAB` using the two specializations `TCAB.STAR` and `TCAB.MST` in three settings. Throughout this section, the tree we use for `TCAB` is fixed over the experiment. For every horizon and Monte Carlo replication, we run one faithful horizon-T tree experiment and record its realized reward-query cost $N(T)$; we do not average multiple tree trajectories within a reported experimental unit. We compare against independent A/B/n testing with the full JT budget (`AB.FULL`) and against a matched-budget independent A/B/n testing. The latter allocates $q_T = \lfloor N(T)/J \rfloor$ observations or learning rounds to each policy. For history-free policies, scaling the resulting per-round estimate by T remains unbiased for the horizon-T cumulative target; for adaptive policies, the corresponding extrapolation is generally biased and is interpreted separately below.

Across the three experiments, `TCAB` improves the cost–precision frontier while preserving the target contrasts. At matched query budgets, it reduces median contrast error and variance relative to independent A/B/n designs, detects smaller policy differences, and uses substantially fewer reward observations. The results show that STAR and MST coupling can make multi-policy comparisons practical under constrained experimental budgets.

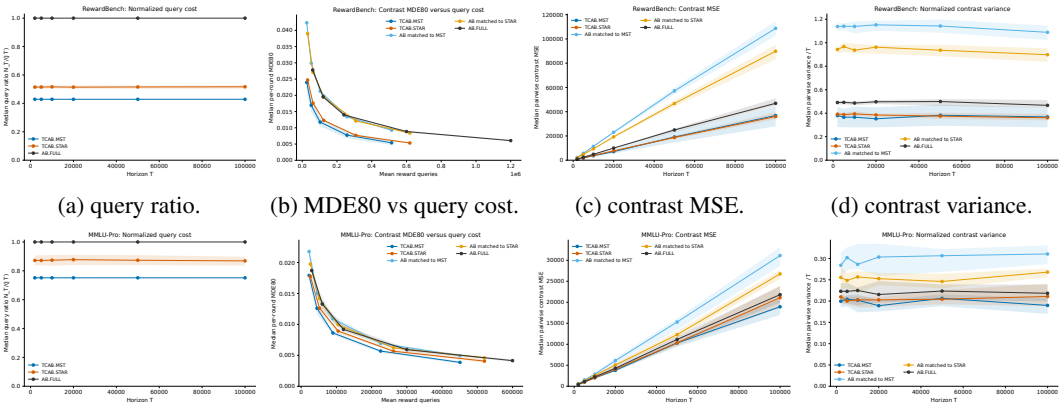
6.1 LLM AGENTS AS NONADAPTIVE POLICIES

RewardBench RewardBench is a benchmark for assessing whether language models correctly rank a human–preferred response above a rejected response (Lambert et al., 2025). We use all 2,985 examples in its filtered evaluation split, which span chat, difficult chat, safety, mathematical reasoning, and code, and sample examples according to weights that reproduce the benchmark's official category–level aggregation. Each example constitutes a context X containing a prompt and a preferred–rejected response pair, and the action space is binary: $A = 0$ selects the preferred response and $A = 1$ selects the rejected response. The policies are $J = 12$ language models with complete scores in the pinned RewardBench results repository. For policy j, we convert its cached scores into the deterministic action $A_j(X) = 0$ when it assigns the higher score to the preferred response and $A_j(X) = 1$ otherwise, using a fixed tie rule. We then model a noisy preference observation by assigning reward $R \sim \text{Bernoulli}(0.9)$ when the selected response is preferred and $R \sim \text{Bernoulli}(0.1)$ otherwise.

MMLU-Pro MMLU-Pro is a challenging multiple-choice benchmark for evaluating language understanding and reasoning across 14 academic domains, with questions containing up to ten candidate answers (Wang et al., 2024). We use its pinned test split as a finite empirical context distribution: a context X is a question together with its answer options, sampled uniformly with replacement, and the action space $A = 0, \ldots, 9$ indexes the candidate answers. We construct $J = 6$ deterministic policies from three open-weight instruction models—SmolLM2-360M-Instruct, Qwen2.5-0.5B-Instruct, and TinyLlama-1.1B-Chat—each evaluated under plain and chat-formatted prompt variants. For every model–prompt combination and question, we score the answer letters and cache the highest-scoring valid option as the policy action. The reward is $R(X, A) = \mathbf{1}\{A = Y_X\}$, where Y_X is the correct answer, so each policy's value equals its test-set accuracy.

The minimum detectable effect at 80% power (MDE80) is the smallest per-round difference between two policies that a two-sided 5% test detects with 80% probability under the testing procedure used in the experiment. Lower values indicate better sensitivity. Figure 2 shows that `TCAB` resolves smaller differences at a given reward-query budget. Both `TCAB.STAR` and `TCAB.MST` achieve cumulative mean-squared error (MSE) and variance comparable to `AB.FULL`, and substantially smaller than independent A/B/n runs with matched budgets, while using about 40% and 80% of the full query budget on RewardBench and MMLU-Pro, respectively.[3] On RewardBench, the reported median contrast MSE and variance reductions relative to matched-budget A/B/n are about 67% for `TCAB.MST` and 59% for `TCAB.STAR`. On MMLU-Pro, the corresponding reductions are about 31% and 20%.

[3] For history-free, time-invariant policies, every edge TV distance is a constant δ_e, so Theorem 5.3 gives $\mathbb{E}[N(T)] = T(1 + \sum_{e \in E} \delta_e)$. Hence the normalized query ratio is constant in T, as observed.



(a) query ratio. (b) MDE80 vs query cost. (c) contrast MSE. (d) contrast variance.

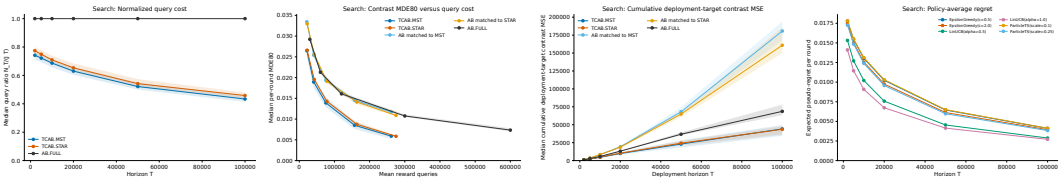
(e) query ratio. (f) MDE80 vs query cost. (g) contrast MSE. (h) contrast variance.

Figure 2: Query cost, inferential sensitivity, and estimation accuracy for RewardBench (top) and MMLU-Pro (bottom). Lines report medians and shaded regions report the interquartile range across pairwise policy contrasts over 500 independent runs. Lower MDE80 indicates sensitivity to smaller policy differences.

6.2 Adaptive online learning policies

To further highlight the benefit of TCAB, we consider a benchmark with adaptive policies. By Theorem 5.5, when the policies asymptotically learn the optimal action, we are able to share most of the observations and attain $\mathbb{E}[N(T)] = T + o(T)$.

MSLR-Search We construct a semi-synthetic search environment from MSLR-WEB10K, a learning-to-rank dataset containing query–document relevance judgments and 136 ranking features (Microsoft Research, 2010). At each round, a query is sampled uniformly and ten candidate documents are drawn from its precomputed top-20 pool; the resulting document-feature vectors form the context, and the action selects one document for the top search position. Selecting document a produces a Bernoulli click whose mean is a ridge-regression relevance score, constrained to $[0.05, 0.95]$. We compare six adaptive policies—LinUCB with exploration parameters 0.5 and 1.0, decaying ϵ-greedy with constants 0.5 and 2.0, and finite-particle linear Thompson sampling with scales 0.1 and 0.25. Because these policies update from their observed clicks, this experiment evaluates TCABin a genuinely history-dependent setting; the semi-synthetic ridge-regression score permits accurate hindsight estimates of policy values and cumulative regret.



(a) query ratio. (b) MDE80 vs query cost. (c) contrast MSE. (d) policy regret.

Figure 3: Query cost, inferential sensitivity, and estimation accuracy for MSLR-Search. (d) plots the average pseudo-regret of each policy.

Consistent with Theorem 5.5, the normalized query cost decreases as the adaptive policies' per-round regrets fall, from roughly 80% toward 50% of the full JT budget over the reported horizon. Plot (b) shows a widening MDE80 advantage at matched cost. For a realized tree cost $N(T)$, the matched-budget independent baseline can advance each of the J adaptive policies for only $q_T = \lfloor N(T)/J \rfloor$ rounds. To display an estimate against the horizon-T cumulative target, the experiment needs to rescale it by $\widetilde{S}_j(T) \approx S_j(q_T)T/q_T$. This extrapolation is generally biased during learning because the average reward over the first q_T rounds need not equal the average reward over the first T rounds; the same issue does not arise for the exact horizon-T TCAB trajectory.

AI USE STATEMENT

In this work, we used generative AI tools for building and implementing the datasets and the numerical experiments, generating the figures, polishing the writing and proofs, and giving feedback on the experiments. We have not used generative AI tools for formulating the framework and the mathematical claims, providing critical mathematical ingredients, giving qualitative analysis, or interpreting the results, and translation, synthetic data generation, and dataset reformation are not applicable to this work. Additionally, we used generative AI tools for identifying relevant literature. We have reviewed all AI-assisted work. We checked LLM-assisted writing and proofs and verified their correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1638–1646. PMLR, 2014.

Omer Angel and Yinon Spinka. Pairwise optimal coupling of multiple random variables. *arXiv preprint arXiv:1903.00632*, 2019.

Siddhartha Banerjee, Sean R. Sinclair, Milind Tambe, Lily Xu, and Christina Lee Yu. Artificial replay: A meta-algorithm for harnessing historical data in bandits. *arXiv preprint arXiv:2210.00025*, 2022.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021. doi: 10.1287/mnsc.2020.3605.

Jennifer Brennan, Yahu Cong, Yiwei Yu, Lina Lin, Yajun Peng, Changping Meng, Ningren Han, Jean Pouget-Abadie, and David M. Holtz. Reducing symbiosis bias through better A/B tests of recommendation algorithms. In *Proceedings of the ACM Web Conference 2025*, pp. 3702–3715. ACM, 2025. doi: 10.1145/3696410.3714738.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 2024.

Adrien Corenflos and Simo Särkkä. The coupled rejection sampler. *arXiv preprint arXiv:2201.09585*, 2022.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Charles W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955. doi: 10.1080/01621459.1955.10501294.

Avishek Ghosh and Abishek Sankararaman. Breaking the $\sqrt{T}$ barrier: Instance-independent logarithmic regret in stochastic contextual linear bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7531–7549. PMLR, 2022.

Somit Gupta, Ron Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumitha Chandran, Nanyu Chen, Dominic Coey, et al. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, 21(1):20–35, 2019. doi: 10.1145/3331651.3331655.

Vitor Hadad, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15):e2014602118, 2021. doi: 10.1073/pnas.2014602118.

Osama A. Hanna, Lin Yang, and Christina Fragouli. Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1791–1821. PMLR, 2023.

Maximilian Kasy and Anja Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021. doi: 10.3982/ECTA17527.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1168–1176. ACM, 2013. doi: 10.1145/2487575.2488217.

Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvari. Confident off-policy evaluation and selection through self-normalized importance weighting. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 640–648. PMLR, 2021.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Nicholas Larsen, Jonathan Stallrich, Srijan Sengupta, Alex Deng, Ron Kohavi, and Nathaniel T Stevens. Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician*, 78(2):135–149, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, 2020.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 297–306. ACM, 2011. doi: 10.1145/1935826.1935878.

Shuangning Li, Chonghuan Wang, and Jingyan Wang. Choosing the better bandit algorithm under data sharing: When do A/B experiments work? *arXiv preprint arXiv:2507.11891*, 2025.

Torgny Lindvall. *Lectures on the Coupling Method*. Dover Publications, Mineola, NY, 2002.

Huiling Meng, Ningyuan Chen, and Xuefeng Gao. Design experiments to compare multi-armed bandit algorithms. *arXiv preprint arXiv:2603.05919*, 2026.

Microsoft Research. Microsoft learning to rank datasets, 2010. URL https://www.microsoft.com/en-us/research/project/mslr/. MSLR-WEB10K.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41:693–721, 2013. doi: 10.1214/13-AOS1101.

Federico Quin, Danny Weyns, Matthias Galster, and Camila Costa Silva. A/b testing: A systematic literature review. *Journal of Systems and Software*, 211:112011, 2024.

David Simchi-Levi and Chonghuan Wang. Multi-armed bandit experimental design: Online decision-making and adaptive inference. *Management Science*, 71(6):4828–4846, 2025. doi: 10.1287/mnsc.2023.00492.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudik. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9167–9176. PMLR, 2020.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823. PMLR, 2015.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 17–26. ACM, 2010. doi: 10.1145/1835804.1835810.

Hermann Thorisson. *Coupling, Stationarity, and Regeneration*. Springer, New York, 2000.

Runzhe Wan, Branislav Kveton, and Rui Song. Safe exploration for efficient policy evaluation and comparison. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22491–22511. PMLR, 2022.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3589–3597. PMLR, 2017.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yuxiao Wen, Jingkai Huang, Zhengyuan Zhou, et al. Perishable online inventory control with context-aware demand distributions. In *NeurIPS 2025 Workshop MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making*.

Yuxiao Wen, Yanjun Han, and Zhengyuan Zhou. Joint value estimation and bidding in repeated first-price auctions. *arXiv preprint arXiv:2502.17292*, 2025.

Yuxiao Wen, Yanjun Han, and Zhengyuan Zhou. Optimal arm elimination algorithms for combinatorial bandits. *Twenty-Ninth Annual Conference on Artificial Intelligence and Statistics*, 2026.

Weiqiang Wu, Jing Yang, and Cong Shen. Stochastic linear contextual bandits with diverse contexts. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2392–2401. PMLR, 2020.

Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2227–2236. ACM, 2015. doi: 10.1145/2783258.2788602.

Mengjiao Yang, Bo Dai, Ofir Nachum, George Tucker, and Dale Schuurmans. Offline policy selection under uncertainty. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4376–4396. PMLR, 2022.

Ruohan Zhan, Vitor Hadad, David A. Hirshberg, and Susan Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2125–2135. ACM, 2021.

A Additional related literature

Multiple comparisons, adaptive inference, and policy selection. Classical many-to-one procedures compare several static treatments with one control while accounting for multiplicity (Dunnett, 1955). Adaptive-design work studies observation allocation or valid inference under adaptively collected data (Kasy & Sautmann, 2021; Hadad et al., 2021; Wen et al., 2025; Simchi-Levi & Wang, 2025). Work on policy comparison and selection includes safe data collection for evaluating several policies (Wan et al., 2022), high-confidence off-policy selection (Kuzborskij et al., 2021), and ranking policies from a fixed experience data set (Yang et al., 2022). These problems differ from ours because the candidates are not jointly coupled to reduce the total costs, while their exact finite-horizon path laws must all be preserved.

Artificial replay and data sharing. The phrase "artificial replay" is also used by Banerjee et al. (2022) for incorporating an exogenous historical data set to warm-start a bandit rather than coupling several prospective policy trajectories. A related industry literature studies interference when recommendation algorithms share training data: Brennan et al. (2025) document "symbiosis bias," and Li et al. (2025) analyzes when data sharing can reverse the ranking of two bandit algorithms. `TCAB` instead shares an observation only through an exact coupling event, so the induced cross-policy dependence is part of the experimental design rather than an unmodeled training interaction.

Contextual-bandit regret. Sublinear regret is available for stochastic contextual bandits under many structural conditions, including smooth covariate effects (Perchet & Rigollet, 2013), covariate diversity (Wu et al., 2020; Bastani et al., 2021), linear reductions with plentiful contexts (Ghosh & Sankararaman, 2022; Hanna et al., 2023; Wen et al., 2026), and general policy classes (Agarwal et al., 2014; Wen et al.); see Lattimore & Szepesvári (2020). `TCAB` treats the candidate policies as black boxes: such structural assumptions are needed only to verify the regret hypotheses used in Section 5, not for replay exactness or the query-cost identity.

B Coupling preliminaries

We recall the two coupling facts used by `TCAB`: share the overlap of two laws, then realize the resulting edge couplings on a tree.

B.1 Two-marginal maximal coupling

Let ν_0, ν_1 be probability measures on a standard Borel space Z, with densities p_0, p_1 relative to a common measure ρ. Their shared mass is

$$\lambda(\mathrm{d}z) := \min\{p_0(z), p_1(z)\}\rho(\mathrm{d}z), \qquad \delta := 1 - \lambda(\mathsf{Z}) = \mathrm{TV}(\nu_0, \nu_1).$$

Lemma B.1 (Canonical maximal coupling)**.** *With probability $1-\delta$, draw $Z \sim \lambda/(1-\delta)$ and set $Z_0 = Z_1 = Z$. Otherwise, draw (Z_0, Z_1) from any coupling of the residual laws*

$$\frac{\nu_0 - \lambda}{\delta} \quad \textit{and} \quad \frac{\nu_1 - \lambda}{\delta}.$$

Branches of probability zero are omitted. Then $Z_i \sim \nu_i$ and

$$\mathbb{P}(Z_0 \neq Z_1) = \delta.$$

No coupling has a strictly smaller disagreement probability.

Proof. The shared and residual branches contribute λ and $\nu_i - \lambda$ to marginal i, giving ν_i. The residual laws are supported on the disjoint sets $\{p_0 > p_1\}$ and $\{p_1 > p_0\}$, so disagreement occurs exactly on the residual branch.

For optimality, any coupling satisfies

$$\nu_0(B) - \nu_1(B) \le \mathbb{P}(Z_0 \neq Z_1)$$

for every measurable set B. Taking the supremum over B gives the lower bound $\delta = \mathrm{TV}(\nu_0, \nu_1)$. The fact that our coupling attains this lower bound yields its optimality. See also Thorisson (2000); Lindvall (2002). □

Lemma B.2 (One-sided rejection sampler). *Draw $Z_0 \sim \nu_0$ and set $Z_1 = Z_0$ with probability $\min\{1, p_1(Z_0)/p_0(Z_0)\}$. On rejection, draw independent proposals $Y \sim \nu_1$ until one is accepted with probability*

$$1 - \min\{1, p_0(Y)/p_1(Y)\}.$$

Set Z_1 to the accepted proposal. This gives a maximal coupling using only sampling and density-ratio access. For $\delta > 0$, the proposal count conditional on the residual branch is geometric with mean $1/\delta$; for $\delta = 0$, that branch is absent. Zero-denominator ratios are immaterial on sampling-null sets.

Proof. The acceptance rules extract the shared mass and the child's residual mass:

$$p_0 \min\{1, p_1/p_0\} = \min\{p_0, p_1\},$$
$$p_1\big[1 - \min\{1, p_0/p_1\}\big] = (p_1 - p_0)_+.$$

Thus the shared branch contributes λ; on rejection, the parent and accepted child have laws $(\nu_0 - \lambda)/\delta$ and $(\nu_1 - \lambda)/\delta$. Lemma B.1 gives maximality. Each residual proposal succeeds with probability $\int (p_1 - p_0)_+ \, \mathrm{d}\rho = \delta$, giving the geometric law. □

B.2 Tree gluing

The following lemma shows that TCAB attains the maximal coupling jointly over the edges of each tree in the tree sequence $\{\mathcal{T}_t\}_t$.

Lemma B.3 (Tree gluing lemma). *Let $\mathcal{T} = (\mathcal{I}, E)$ be a finite rooted tree and ν_j probability laws on a standard Borel space Z. Prescribe a coupling γ_{pv} of ν_p, ν_v for each parent–child edge (p, v). These couplings can all be realized as edge marginals of one joint law. In particular, every tree edge can be maximally coupled at once.*

Proof. Standard Borelness gives a conditional kernel K_{pv} for each edge:

$$\gamma_{pv}(\mathrm{d}z_p, \mathrm{d}z_v) = \nu_p(\mathrm{d}z_p) K_{pv}(z_p, \mathrm{d}z_v).$$

Draw the root from ν_r, then each child from $K_{pv}(Z_p, \cdot)$ using fresh randomness. If the parent has law ν_p, the edge has law γ_{pv} and the child has law ν_v. Induction proves the claim. The tree structure matters because each child has only one parent and hence one prescribed conditional law. □

C Proof of the implementation proposition

In the remaining proofs, $\mathcal{F}_{t-1}$ contains all information available before round-t policy draws, including past histories and the selected tree. Tree selection uses only past information and independent randomization, not current-round observations. The tree's paths and degrees are therefore $\mathcal{F}_{t-1}$-measurable. We suppress the history arguments of $\delta_{ij,t}$ for notational simplicity.

Proof of Proposition 4.1. Fix a parent–child edge (i, j) and condition on $\mathcal{F}_{t-1}$. Relative to $P_X \otimes \mathrm{counting}$, the two densities are the policy probabilities. Lemma B.2 thus gives the sampler in Equations (4.3–4.5).

Count residual proposals including the accepted one, and set $U_{ij,t} = 0$ on the shared branch. For $\delta_{ij,t} > 0$, the residual branch has probability $\delta_{ij,t}$ and conditional expected length $1/\delta_{ij,t}$. These factors cancel, giving

$$\mathbb{E}[U_{ij,t} \mid \mathcal{F}_{t-1}] = \mathbf{1}\{\delta_{ij,t} > 0\} \le 1.$$

Summing over $J - 1$ edges and T rounds proves the bound. Rejected proposals require no reward query. □

D Proofs of exactness and query cost

Proof of Theorem 5.1. We verify the standalone transition conditional on each policy's history, then induct over time.

Condition on $\mathcal{F}_{t-1}$. The root has the required context–action law by construction. Whenever a parent has its required law, Lemma B.2 gives the child's required law. Induction down the tree therefore yields

$$\mathcal{L}\big((X_t^j, A_t^j) \mid \mathcal{F}_{t-1}\big) = \nu_{j,t}^{H_{t-1}^j} \qquad \text{for every } j. \tag{D.1}$$

It remains to check the rewards. No current-round acceptance decision uses a current-round reward, so we may equivalently generate all context–action pairs first and then draw the rewards. Delete the mismatched tree edges. Each remaining component has one common pair (x, a) and receives one fresh draw from $Q_a(\cdot \mid x)$, shared by its nodes. Consequently,

$$\mathcal{L}\big(R_t^j \mid \mathcal{F}_{t-1}, X_t^j = x, A_t^j = a\big) = Q_a(\cdot \mid x). \tag{D.2}$$

Sharing changes dependence across policies, but not this conditional reward law.

Together, these two identities give the standalone transition in (3.2). Their right-hand sides depend on the past only through H_{t-1}^j, so the same transition holds after conditioning only on that policy's history. Starting from the empty histories and iterating over t completes the proof. □

Proof of Corollary 5.2. Theorem 5.1 gives $S_j(T)$ its standalone distribution and hence expectation. Linearity of expectation gives every contrast identity; independence between policies is not needed. □

Proof of Theorem 5.3. Deleting one edge from a tree increases its number of connected components by one. Since `TCAB` queries one reward per matched component, the round's cost is exactly $1 + \sum_{e \in E_t} D_{e,t}$. Every edge is maximally coupled conditional on $\mathcal{F}_{t-1}$, so

$$\mathbb{E}[D_{e,t} \mid \mathcal{F}_{t-1}] = \delta_{e,t}.$$

Taking conditional expectations gives the round-wise identity in equation (5.3). Summing over rounds and taking expectations gives the total-cost identity. Neither step requires independence between edge mismatches. □

Edge-local optimality and the oracle MST. For completeness, fix the pre-round information and consider any design with the same conditional standalone marginals. Suppose its root requests a fresh reward, and each child either copies its parent's reward at the same complete context–action pair or requests a fresh reward; no other sharing is allowed. On every edge e, Lemma B.1 forces a mismatch with probability at least $\delta_{e,t}$. Each mismatch requires a new query, so the conditional expected cost is at least $1 + \sum_{e \in E_t} \delta_{e,t}$. `TCAB` attains this bound simultaneously on all edges.

If all current pairwise distances are known, minimizing this cost over trees is exactly the minimum-spanning-tree problem. This minimization is stated in (5.4) and gives rise to the heuristic variant `TCAB.MST` as discussed in Section 5.2.

E Proofs of regret-based efficiency

Proof of Theorem 5.5. Two policies are close in TV whenever both usually choose the same oracle action. At context x, apply the triangle inequality through the point mass at $a^*(x)$, then integrate over P_X. The conditional marginals in (D.1) give

$$\delta_{ij,t} \le \mathbb{P}\big(A_t^i \ne a^*(X_t^i) \mid \mathcal{F}_{t-1}\big) + \mathbb{P}\big(A_t^j \ne a^*(X_t^j) \mid \mathcal{F}_{t-1}\big).$$

Each policy appears in exactly $d_{j,t}$ edges. Summing this inequality over edges and rounds, and applying Theorem 5.3, yields

$$\mathbb{E}[N(T)] - T \le \mathbb{E}\left[\sum_{t=1}^{T} \sum_{j \in \mathcal{I}} d_{j,t} \mathbf{1}\{A_t^j \ne a^*(X_t^j)\}\right]. \tag{E.1}$$

Charge each nonoptimal choice either to a small-gap context or to at least ϵ regret:

$$\mathbf{1}\{A_t^j \ne a^*(X_t^j)\} \le \mathbf{1}\{\Delta(X_t^j) \le \epsilon\} + \frac{\Delta_{A_t^j}(X_t^j)}{\epsilon}.$$

Since the degrees are $\mathcal{F}_{t-1}$-measurable and each context has conditional law P_X,

$$\mathbb{E}\left[\sum_j d_{j,t}\mathbf{1}\{\Delta(X_t^j) \le \epsilon\} \,\middle|\, \mathcal{F}_{t-1}\right] = F_\Delta(\epsilon)\sum_j d_{j,t} = 2(J-1)F_\Delta(\epsilon).$$

Multiply the preceding pointwise bound by $d_{j,t}$, sum, and use Equation (E.1). The regret terms give $\mathcal{R}_{\deg}(T)/\epsilon$, proving Equation (5.6).

For the asymptotic claim, fix $\epsilon > 0$, divide by T, and let $T \to \infty$. Fixed J and sublinear regrets imply $\mathcal{R}_{\deg}(T) = o(T)$, so

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N(T)] - T}{T} \le 2(J-1)F_\Delta(\epsilon).$$

Now let $\epsilon \downarrow 0$. Uniqueness of the optimal action and finiteness of the action set imply $\Delta(X) > 0$ almost surely, so the right-hand side tends to zero. Together with $N(T) \ge T$, this proves Equation (5.7). □

Proof of Corollary 5.6. Since $F_\Delta \le 1$, we may enlarge C so that the margin condition holds for every $\epsilon > 0$, not just sufficiently small ϵ. Theorem 5.5 then gives

$$\mathbb{E}[N(T)] - T \le 2C(J-1)T\epsilon^\beta + \frac{\mathcal{R}_{\deg}(T)}{\epsilon}.$$

For positive $\mathcal{R}_{\deg}(T)$, balance the cost of small-gap contexts against the regret term by taking

$$\epsilon \asymp \left(\frac{\mathcal{R}_{\deg}(T)}{(J-1)T}\right)^{1/(\beta+1)}.$$

This gives Equation (5.8). If $\mathcal{R}_{\deg}(T) = 0$, let $\epsilon \downarrow 0$ in the same bound.

Under a uniform gap, each nonoptimal choice incurs at least $\Delta_{\min}$ regret. Applying this directly to Equation (E.1) gives (5.9). □

F Proofs of the variance results

Throughout this section, we assume that rewards lie in $[0, 1]$, as stated in Section 5.4.

Proof of Theorem 5.7. Adding the one-round regret to a policy's reward removes its mean loss relative to the oracle. The conditional laws proved above give, for every j,

$$\mathbb{E}\Big[R_t^j + \Delta_{A_t^j}(X_t^j) \mid \mathcal{F}_{t-1}\Big] = \int \mu_*(x) P_X(\,\mathrm{d}x).$$

Fix i, j and take the difference of these adjusted rewards:

$$W_t := R_t^j + \Delta_{A_t^j}(X_t^j) - R_t^i - \Delta_{A_t^i}(X_t^i).$$

The common conditional mean cancels, so (W_t) is a martingale-difference sequence.

If every edge on $P_{ij,t}$ matches, the endpoints share the same context, action, and reward, and hence $W_t = 0$. Each adjusted reward lies in $[0, 2]$, so otherwise $|W_t| \le 2$. Therefore

$$W_t^2 \le 4\sum_{e\in P_{ij,t}} D_{e,t}.$$

Martingale differences are uncorrelated across rounds. Using $\mathbb{E}[D_{e,t} \mid \mathcal{F}_{t-1}] = \delta_{e,t}$ and the predictability of the path gives

$$\mathrm{Var}\left(\sum_{t=1}^T W_t\right) = \sum_{t=1}^T \mathbb{E}[W_t^2] \le 4B_{ij}(T).$$

Finally,

$$S_j(T) - S_i(T) = \sum_{t=1}^{T} W_t + L_i(T) - L_j(T).$$

Apply $\mathrm{Var}(U+V) \le 2\mathrm{Var}(U) + 2\mathrm{Var}(V)$, first to this decomposition and then to $L_i(T) - L_j(T)$, to obtain

$$\mathrm{Var}(S_j(T) - S_i(T)) \le 8B_{ij}(T) + 4\mathrm{Var}(L_i(T)) + 4\mathrm{Var}(L_j(T)).$$

This is slightly stronger than (5.10).

For the asymptotic conclusion, retain the fixed-J and almost-sure uniqueness assumptions of Theorem 5.5. Every path is contained in the corresponding tree, so

$$B_{ij}(T) \le \mathbb{E}\left[\sum_{t=1}^{T} \sum_{e \in E_t} \delta_{e,t}\right] = \mathbb{E}[N(T)] - T = o(T).$$

The conclusion follows when both regret variances are also $o(T)$. Uniqueness is needed only for this implication from low regret to low cost, not for the finite-sample variance bound. □

F.1 General fixed zero-sum contrasts

Let $S(T) := (S_0(T), \dots, S_{J-1}(T))^\top$ and fix $c \in \mathbb{R}^J$ with $\sum_j c_j = 0$. By Theorem 5.1, $c^\top S(T)$ is unbiased for $c^\top V(T)$. The zero-sum condition lets us write this contrast as a weighted sum of comparisons with any fixed reference policy.

Theorem F.1 (Fixed zero-sum contrasts)**.** *Fix an anchor $r \in \mathcal{I}$. Then*

$$\mathrm{Var}\big(c^\top S(T)\big) \le \left(\sum_{j \ne r} |c_j| \sqrt{15B_{jr}(T) + 6\mathrm{Var}(L_j(T)) + 6\mathrm{Var}(L_r(T))}\right)^2. \tag{F.1}$$

If J, c are fixed, $\mathbb{E}[N(T)] - T = o(T)$, and $\mathrm{Var}(L_j(T)) = o(T)$ for every policy, then $\mathrm{Var}(c^\top S(T)) = o(T)$.

Proof. Since $\sum_j c_j = 0$,

$$c^\top S(T) = \sum_{j \ne r} c_j \{S_j(T) - S_r(T)\}.$$

Center both sides and apply the triangle inequality in L^2:

$$\sqrt{\mathrm{Var}(c^\top S(T))} \le \sum_{j \ne r} |c_j| \sqrt{\mathrm{Var}(S_j(T) - S_r(T))}.$$

Substituting Theorem 5.7 proves Equation (F.1). For the asymptotic claim, $B_{jr}(T) \le \mathbb{E}[N(T)] - T = o(T)$. Each square-root term is therefore $o(\sqrt{T})$, and the square of a fixed finite sum of these terms is $o(T)$. □

Corollary F.2 (Simple second-moment condition)**.** *Let*

$$M_j(T) := \sum_{t=1}^{T} \mathbf{1}\{A_t^j \ne a^*(X_t^j)\}.$$

For fixed J, if $\mathbb{E}[N(T)] - T = o(T)$ and $\mathbb{E}[M_j(T)^2] = o(T)$ for every j, then every fixed zero-sum contrast has variance $o(T)$.

Proof. Each nonoptimal action incurs at most one unit of regret, so $0 \le L_j(T) \le M_j(T)$. Hence

$$\mathrm{Var}(L_j(T)) \le \mathbb{E}[L_j(T)^2] \le \mathbb{E}[M_j(T)^2] = o(T).$$

Apply Theorem F.1. □