Fast A/B/n Testing: Exact Multi-Policy Comparison via Tree-Coupled Feedback Sharing

Yuxiao Wen*

August 14, 2026

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. A direct A/B/n design gives each of J policies its own horizon-T trajectory and therefore uses JT outcomes. We introduce Tree-Coupled A/B Testing (`TCAB`), an exact feedback-sharing design for arbitrary history-dependent contextual-bandit policies. At each round, a predictable tree connects the current policy histories; every parent–child context–action law is maximally coupled, and one reward is shared within each component of matched tree edges. 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 reward queries satisfies the pathwise identity $N(T) = T + \sum_{t,e} D_{e,t}$ and hence equals T plus cumulative tree-edge total variation in expectation. This cost is conditionally optimal among exact edge-local designs on the selected tree, and a current-round minimum-spanning tree is myopically optimal among tree designs. For fixed J, sublinear pseudo-regret of every policy and almost-sure uniqueness of the oracle action imply $\mathbb{E}[N(T)] = T + o(T)$, versus JT for independent runs. We also obtain finite-sample variance bounds for pairwise policy contrasts. Experiments on reward-model evaluation, multiple-choice language-model evaluation, and adaptive search policies demonstrate substantial improvements in the cost–precision frontier.

Keywords: Online experimentation; A/B testing; contextual bandits; maximal coupling; adaptive policy comparison; feedback sharing.

1 Introduction

Online controlled experiments are production infrastructure at major digital platforms. Google, Microsoft/Bing, and LinkedIn have described systems that support overlapping experiments and continuous product iteration [Kohavi et al., 2013, Tang et al., 2010, Xu et al., 2015]. A cross-industry summit involving thirteen organizations reported that the participating organizations had collectively tested more than one hundred thousand treatment variants in the preceding year [Gupta et al., 2019]. The same pressure now appears in machine-learning evaluation: product teams compare many recommender-system configurations, ranking policies, prompting strategies, and model checkpoints, while human or online feedback remains expensive. For example, the original Chatbot Arena study accumulated more than 240,000 human preference votes to compare language models [Chiang et al., 2024].

*Yuxiao Wen is with the Courant Institute of Mathematical Sciences, New York University, email: yuxiaowen@nyu.edu.

A useful distinction is between reusing *traffic* and reusing *feedback*. Google's overlapping-experiment infrastructure, for example, allows compatible experiments in different layers to share the same underlying traffic while preserving the randomization needed within each experiment [Tang et al., 2010]. Our question is complementary and operates inside a multi-policy comparison: when several candidate policies would produce the same contextual decision, can one realized outcome be used by several policies without changing the finite-horizon law of any of them? This distinction matters because traffic multiplexing alone does not remove duplicated reward-bearing interactions among the alternatives in a single policy-comparison problem.

The classical A/B experiment is designed for static treatments. Modern systems instead often compare *policies*: a policy observes the current query or user context, chooses an action, receives only the selected action's outcome, and may update all later decisions from its adaptive history. To compare J candidate policies over the same target horizon T, the direct A/B/n design runs J independent trajectories and uses JT reward-bearing interactions. This cost is operationally important when an experimental action can reduce user experience, consume expert labels, invoke a costly model, or delay a deployment decision.

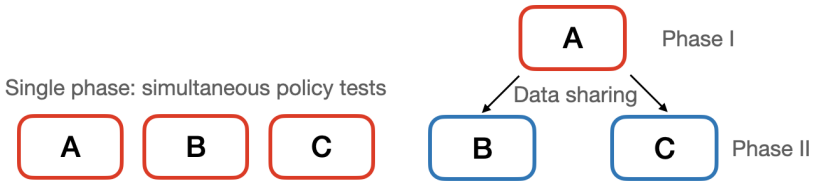
There is nevertheless a large source of redundancy. Candidate policies generated by neighboring hyperparameters, model checkpoints, or business rules frequently make the same decision on the same context. When two policies realize the same complete context–action pair, they require the same conditional 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.

Contextuality makes this challenge fundamentally different from replaying an arm label. Even under i.i.d. contexts, the contexts attached to the observations of a selected action depend on the policy's selection rule and history. 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 model a full-context reward kernel $Q_a(\cdot \mid x)$ and couple the policies' complete one-step *context–action laws*. A maximal coupling makes two complete pairs equal with the largest probability permitted by their total-variation distance; the shared branch uses one reward and the residual branch opens a fresh query.

Moving from two policies to $J \geq 3$ creates a second obstruction. 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 and 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 broken edge starts a new reward lineage; matched edges transmit the same context, action, and reward. Consequently, the number of reward queries is exactly one plus the number of broken edges in each round.

Our algorithm, Tree-Coupled A/B Testing (`TCAB`), is round-synchronous. At the beginning of round t, the experiment may select any tree measurable with respect to the histories available through round $t-1$. It then samples the root, traverses the tree by depth, queries one outcome per matched-edge component, and updates all policies simultaneously. This outer-loop-over-time formulation directly permits predictable, time-adaptive trees. It also exposes practical parallelism: all children at the same depth can be coupled concurrently once their parents are available, and the reward queries for distinct components can be issued concurrently. Adaptive policies remain sequential across rounds, as they must, while nonadaptive policies can additionally parallelize across rounds.

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 over the selected tree edges. This identity yields a precise design



(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.

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. We emphasize that this is a myopic tree-optimality statement, not a claim of global optimality among arbitrary multi-marginal couplings or over the full adaptive horizon.

The method is most useful when candidate policies are related rather than arbitrary. Examples include neighboring model checkpoints, nearby hyperparameter settings, alternative ranking or recommendation rules that agree on most users, and learning algorithms that increasingly concentrate on the same good actions. In the first regime, the relevant tree-edge total-variation distances are small from the outset; in the second, our regret analysis shows that the excess query cost can vanish relative to T. Conversely, if candidate policies almost never agree on their complete context–action pairs, `TCAB` remains exact but offers little query saving. This makes the theory directly diagnostic: the same edge disagreements that determine cost also indicate when feedback sharing will or will not help.

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 full-context reward kernels $Q_a(\cdot \mid x)$.

2. We introduce a round-synchronous, predictably time-adaptive version of `TCAB`. Maximal couplings on all selected tree edges coexist, all same-depth coupling operations and all component-level reward queries admit parallel execution, and every policy retains exactly its standalone trajectory law.

3. We prove the pathwise and expected cost identities

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

   Within the natural class of conditionally exact edge-local designs, `TCAB` is optimal on every selected tree. A baseline star and a current-round minimum-spanning tree give two concrete specializations.

4. We establish a finite-sample regret-to-cost bound for predictable 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; a margin condition yields an explicit rate.

5. 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; a general fixed zero-sum extension is given in the appendix.

6. Across two language-model evaluation tasks and an adaptive search-bandit task, `TCAB` improves the empirical cost–precision trade-off relative to independent A/B/n baselines with matched or full budgets.

2 Related literature

Large-scale online experimentation. Online controlled experimentation has become core infrastructure at large digital platforms. Google's overlapping-experiment architecture was designed to run more experiments on limited traffic by allowing compatible experiments in different layers to overlap on the same users or queries [Tang et al., 2010]; Microsoft/Bing and LinkedIn describe related large-scale experimentation systems and organizational challenges [Kohavi et al., 2013, Xu et al., 2015]. A cross-industry summit emphasizes the scale of the resulting experimentation programs and the operational pressure for faster, more reliable decisions [Gupta et al., 2019]. These systems motivate our resource question but solve a different problem: they multiplex traffic across experiments, whereas `TCAB` shares realized feedback across the candidate policies within one multi-policy comparison while preserving each policy's standalone trajectory law.

Artificial Replay and comparison of learning algorithms. The closest methodological work is Meng et al. [2026], which introduces Artificial Replay for comparing two context-free stochastic bandit algorithms and proves exact marginal, interaction-cost, and variance guarantees. Our setting introduces two obstacles absent from the two-policy context-free construction. First, a contextual policy selects actions after observing a query, so exact reuse requires coupling the complete context–action law rather than only an arm label. Second, for $J \geq 3$ distributions, all pairwise maximal couplings need not be compatible; our policy tree selects $J - 1$ pairs whose maximal couplings can be glued simultaneously. The phrase "artificial replay" is also used by Banerjee et al. [2022] for incorporating an exogenous historical data set to warm-start a bandit. That problem concerns how one learner uses historical observations, rather than how several prospective policy trajectories share newly generated feedback.

Multiple comparisons, adaptive experiments, and policy selection. Classical many-to-one procedures compare several static treatments with a common control while accounting for multiplicity [Dunnett, 1955]. Adaptive-design work studies how observations should be allocated, or how valid inference can be maintained, under adaptively collected data [Hadad et al., 2021, Kasy and Sautmann, 2021, Simchi-Levi and Wang, 2025]. Related work on policy comparison and selection includes safe exploration 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 works optimize allocation or inference under a given data-generating scheme. Our objective is different: construct a joint prospective experiment in which every adaptive candidate has exactly the finite-horizon path law it would have had in isolation, but duplicated reward queries are shared whenever coupling permits.

Contextual replay and off-policy evaluation. Li et al. [2011] give an exact replay evaluator for a contextual-bandit algorithm from a uniformly randomized log: a logged event is retained when the target algorithm chooses the logged action, so the retained adaptive history has the same law as an online target trajectory. Inverse-propensity, doubly robust, self-normalized, and related methods form a broader off-policy-evaluation literature [Dudík et al., 2011, Su et al., 2020, Swaminathan and Joachims, 2015, Wang et al., 2017, Zhan et al., 2021]. Those methods begin with an exogenous log

and therefore require support, weighting, modeling, or a bias–variance compromise. In `TCAB`, the candidate policies are runnable and the experiment prospectively generates residual observations when policies fail to couple. Lack of overlap therefore raises the number of reward queries rather than invalidating the target policy trajectory.

Data sharing and interference between learning algorithms. A distinct line of work asks what happens when algorithms in an A/B experiment share training data. Brennan et al. [2025] document "symbiosis bias" in recommendation experiments, and Li et al. [2025] analyze settings in which data sharing can alter or even reverse the ranking of two bandit algorithms. Such sharing changes what the learning algorithms observe and can therefore change the estimand. `TCAB` instead introduces dependence through an exact coupling: an observation is shared only on a branch for which the recipient policy's conditional transition law remains correct. Hence cross-policy dependence is intentional and accounted for, while each individual policy's path law is preserved.

Maximal coupling, common random numbers, and multi-marginal transport. For two probability laws, maximal coupling attains the total-variation lower bound on disagreement [Lindvall, 2002, Thorisson, 2000]; one-sided rejection constructions provide exact implementations under density-ratio access [Corenflos and Särkkä, 2022]. With three or more marginals, simultaneous pairwise maximality can fail [Angel and Spinka, 2019]. Our tree construction uses the complementary fact that prescribed pairwise couplings on the edges of an acyclic graph can always be glued, so exactly $J-1$ strategically selected pairs can be maximal without approximation. Sharing contexts and outcomes is also related to common-random-number methods for simulation comparison [Glasserman, 2003, Kleijnen, 1975, Nelson and Matejcik, 1995]. The adaptive-policy setting is more delicate because shared randomness enters a learner's history and changes later actions unless the correct conditional transition is maintained recursively. At a fixed round, globally minimizing the number of distinct realized context–action pairs over all J marginals can be viewed as a multi-marginal optimal-transport problem [Pass, 2015, Villani, 2009]. We use the tree restriction for exactness, transparency, and an implementable local sampler; we do not claim to solve the unrestricted multi-marginal problem.

Bandits and no-regret online learning. Our efficiency guarantees connect directly to the extensive literature on no-regret sequential decision-making. In stochastic multi-armed bandits, classical upper-confidence-bound algorithms achieve logarithmic instance-dependent regret [Auer et al., 2002], while Thompson sampling admits logarithmic instance-dependent and near-optimal worst-case regret guarantees [Agrawal and Goyal, 2012, 2013a]. No-regret guarantees extend broadly to contextual and structured settings, including contextual bandits with general policy classes [Agarwal et al., 2014, Beygelzimer et al., 2011], linear contextual bandits [Abbasi-Yadkori et al., 2011, Agrawal and Goyal, 2013b], and stochastic contextual models under structural conditions such as smooth covariate effects, covariate diversity, and plentiful contexts [Bastani et al., 2021, Ghosh and Sankararaman, 2022, Hanna et al., 2023, Perchet and Rigollet, 2013, Wen et al., 2026a, Wu et al., 2020]; see Lattimore and Szepesvári [2020] for a general treatment. Such online-learning models arise in a wide range of applications, including pricing [Cohen et al., 2020, Kleinberg and Leighton, 2003], advertising [Hu et al., 2026, Weed et al., 2016, Wen et al., 2025], and search and recommendation systems [Li et al., 2010, Wen et al., 2026b]. `TCAB` does not require the candidate policies to belong to any particular model class and treats them as black boxes: these structural assumptions are needed only to establish a particular policy's regret rate, not for coupling exactness or the query-cost identity. The connection arises through our query-efficiency result: whenever the

compared policies have sublinear pseudo-regret and increasingly concentrate on a common optimal action, their pairwise disagreement is sublinear, and `TCAB` requires only $T + o(T)$ reward queries for any fixed number of policies, rather than the JT queries required by independent evaluation. Thus, advances in no-regret learning that make candidate policies converge more rapidly toward the oracle translate directly into greater feedback sharing under `TCAB`.

3 Problem formulation

3.1 Stochastic contextual environment

Let $T \geq 1$ be the target horizon, $\mathcal{A} = [K]$ the action space, and $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ a standard Borel full-context space. Contexts arrive as

$$X_t \sim P_X \quad \text{i.i.d. over } t. \tag{1}$$

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 $a \in \mathcal{A}$, a reward query at (x, a) returns a draw from the full-context 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 full-context outcomes)**.** *Conditional on the queried context–action pairs, reward queries are independent draws from their corresponding kernels $Q_a(\cdot \mid x)$; rejected unlabeled context proposals generate no reward and cause no carryover or interference.*

The full-context specification allows the selected action's outcome to depend on the entire user/slate state and guarantees that two policies matching on (X, A) require the same conditional reward law.

3.2 Candidate policies and estimands

Fix $J \geq 2$ policies indexed by $\mathcal{I} := \{0, 1, \ldots, J-1\}$. Policy j is a sequence of measurable kernels $\pi_{j,t}(\cdot \mid h, x) \in \Delta(\mathcal{A})$ with history $H_{t-1}^j = (X_s^j, A_s^j, R_s^j)_{s<t}$; any persistent internal state affecting future decisions is included in this state. In a standalone run,

$$X_t^j \sim P_X, \qquad A_t^j \sim \pi_{j,t}(\cdot \mid H_{t-1}^j, X_t^j), \qquad R_t^j \sim Q_{A_t^j}(\cdot \mid X_t^j). \tag{2}$$

Define

$$S_j(T) := \sum_{t=1}^{T} R_t^j, \qquad V_j(T) := \mathbb{E}[S_j(T)], \qquad V(T) := (V_0(T), \ldots, V_{J-1}(T))^\top. \tag{3}$$

For a baseline r, let

$$\theta^{(r)}(T) := (V_j(T) - V_r(T))_{j \neq r}. \tag{4}$$

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.

3.3 Operational access

Assumption 3.2 (Pairwise context–action coupling access)**.** *Before requesting a reward, the experiment can draw an unlabeled $x \sim P_X$, side-effect-free sample/evaluate the relevant policy kernels at x, reject a proposal without updating either policy, and request a reward only after a context–action pair is accepted (or reuse an already queried reward after an exact match). Equivalently, the experiment can implement the pairwise maximal-coupling primitive used below.*

More explicitly, for an ordered parent–child pair (p, v) at current histories (h_p, h_v), the experiment must be able to (i) draw $x \sim P_X$ before requesting a reward; (ii) sample from and evaluate $\pi_{p,t}(\cdot \mid h_p, x)$ and $\pi_{v,t}(\cdot \mid h_v, x)$ without advancing either policy state; (iii) reject a proposal and return the underlying query to the default workflow without updating either policy; and (iv) after acceptance, either query $Q_a(\cdot \mid x)$ or copy an already queried reward when the two complete pairs match. The ability to screen contexts without consuming the reward-bearing resource is the operational distinction between a context proposal and an experimental reward query.

We count the latter resource. Depending on the application, it may correspond to exposing a user to an experimental action and observing a click or purchase, collecting a human preference label, invoking an expensive downstream model or tool, or obtaining another outcome that is materially more costly than evaluating the candidate policies on an unlabeled context. The number of screened proposals can also matter computationally; Section 4.2 gives its exact rejection-sampling behavior separately from the reward-query count $N(T)$.

Neither the density of P_X nor the reward kernel Q_a needs to be evaluated. What is needed is policy-probability access on a proposed context. For deterministic policies, these probabilities reduce to evaluating the policies' chosen actions, and the coupling implementation simplifies accordingly.

4 Tree-Coupled A/B Testing

Let $\mathcal{F}_{t-1}$ contain all policy histories and experimental randomness revealed through round $t-1$. At the start of round t, choose an $\mathcal{F}_{t-1}$-measurable rooted spanning tree $\mathcal{T}_t = (\mathcal{I}, E_t, r_t)$. Thus the tree may react to past data but cannot depend on unrevealed round-t information. Write $p_t(v)$ and $\mathrm{dep}_t(v)$ for the parent and depth of a nonroot vertex v.

4.1 Conditional context–action laws and tree coupling

Let $\mathcal{Z} := \mathcal{X} \times \mathcal{A}$, $Z_t^j = (X_t^j, A_t^j)$, and, conditional on policy history h_j, define the policy's next complete-pair law

$$\nu_{j,t}^{h_j}(\,\mathrm{d}x, a) := P_X(\,\mathrm{d}x)\,\pi_{j,t}(a \mid h_j, x). \tag{5}$$

For policies i, j let

$$\delta_{ij,t}(h_i, h_j) := \mathsf{TV}\left(\nu_{i,t}^{h_i}, \nu_{j,t}^{h_j}\right) = \frac{1}{2}\int_{\mathcal{X}} \left\|\pi_{i,t}(\cdot \mid h_i, x) - \pi_{j,t}(\cdot \mid h_j, x)\right\|_1 P_X(\,\mathrm{d}x). \tag{6}$$

A maximal coupling of these two laws matches their complete pairs with probability $1 - \delta_{ij,t}$. Conditional on $\mathcal{F}_{t-1}$ and $\mathcal{T}_t$, `TCAB` samples the root from its law (5) and recursively maximally couples each child to its parent. Since the selected edges form a tree, the $J-1$ prescribed edge couplings can be glued into a single joint law while preserving every node marginal; the formal gluing statement is proved in Appendix A.

The use of the complete pair $Z = (X, A)$ is essential. Sharing merely because two policies select the same arm can be invalid in a contextual problem: the distribution of the context attached to that arm is itself policy dependent. With the full-context reward model $Q_a(\cdot \mid x)$, equality of (X, A) is exactly the event on which two policies require the same conditional reward law.

4.2 One-sided rejection implementation

Every law in (5) is dominated by $\rho := P_X \otimes$ counting, with density

$$p_{j,t}^{h_j}(x, a) = \pi_{j,t}(a \mid h_j, x).$$

Consider an oriented tree edge (p, v) and fix the current histories (h_p, h_v). A parent-first maximal coupling first draws $Z_p \sim \nu_{p,t}^{h_p}$ and lets the child inherit this same pair with probability

$$q_{pv,t}(Z_p) := \min\left\{1, \frac{p_{v,t}^{h_v}(Z_p)}{p_{p,t}^{h_p}(Z_p)}\right\} = \min\left\{1, \frac{\pi_{v,t}(A_p \mid h_v, X_p)}{\pi_{p,t}(A_p \mid h_p, X_p)}\right\}. \tag{7}$$

The denominator is positive almost surely on a pair generated by the parent. The accepted mass is the overlap measure $\min\{p_{p,t}^{h_p}, p_{v,t}^{h_v}\}\rho$.

If the parent pair is rejected, the child must be drawn from the residual part of its own law. Operationally, repeatedly draw

$$X \sim P_X, \qquad A \sim \pi_{v,t}(\cdot \mid h_v, X),$$

and accept the proposal with probability

$$s_{pv,t}(X, A) := 1 - \min\left\{1, \frac{\pi_{p,t}(A \mid h_p, X)}{\pi_{v,t}(A \mid h_v, X)}\right\}. \tag{8}$$

The accepted residual has normalized density

$$\frac{(p_{v,t}^{h_v} - p_{p,t}^{h_p})_+}{\delta_{pv,t}(h_p, h_v)}$$

with respect to ρ. Thus the procedure is an exact maximal coupling. The common context measure cancels from both ratios, which is why the experiment never needs to know a density for P_X; the reward kernel also does not enter the coupling decision.

For deterministic policies, (7) reduces to an intuitive rule: the parent's pair is inherited exactly when the child would choose the same action on the parent's context. After a mismatch, contexts are proposed until the two current decision rules disagree in the direction required by the residual law. This is also useful operationally because model scoring or policy evaluation can often be done before committing a query to the experimental treatment.

Proposition 4.1 (Exact rejection implementation and proposal overhead)**.** *Let $U_{pv,t}$ denote the number of sampled proposals before one is accepted by (8). Under Assumption 3.2, (7)–(8) implement the edge couplings used by* `TCAB`*. For an edge (p, v) with $\delta_{pv,t} > 0$, 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 $\delta_{pv,t}$, its unconditional expected proposal contribution is at most one:*

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

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

The proposition separates two notions of cost. Small total variation 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 theoretical objective $N(T)$ nevertheless counts only reward-bearing experimental interactions, because those are the resource that feedback sharing is designed to reduce.

Algorithm 1: Round-synchronous Tree-Coupled A/B Testing (`TCAB`)

Input: Horizon T, policies $(\pi_j)_{j\in\mathcal{I}}$, predictable tree rule $(\Psi_t)_{t\le T}$, context sampler, reward interface

Initialize $H_0^j \leftarrow \varnothing$ for every $j \in \mathcal{I}$;
for $t = 1, \dots, T$ **do**
  Select $\mathcal{T}_t = (\mathcal{I}, E_t, r_t) \leftarrow \Psi_t((H_{t-1}^j)_{j\in\mathcal{I}})$;
  Draw $Z_t^{r_t} \sim \nu_{r_t,t}^{H_{t-1}^{r_t}}$;
  for $d = 1, \dots, \max_v \mathrm{dep}_t(v)$ **do**
    In parallel over v at depth d, maximally couple Z_t^v to $Z_t^{p_t(v)}$ using (7)–(8);
  Form the matched-edge components $\mathcal{C}_t$;
  In parallel over $C \in \mathcal{C}_t$, query one $R_t^C \sim Q_{A_t^C}(\cdot \mid X_t^C)$ and assign it to every $j \in C$;
  Simultaneously append (X_t^j, A_t^j, R_t^j) to every policy history;
return $(S_j(T))_{j\in\mathcal{I}}$ *and the desired policy contrasts.*

4.3 Reward inheritance, matched components, and parallelism

After all context–action pairs at round t are generated, retain the tree edges on which the endpoints match and let $\mathcal{C}_t$ be the resulting connected components. Equality propagates along a matched path, so every policy in a component C has the same pair $Z_t^C = (X_t^C, A_t^C)$. `TCAB` draws one reward

$$R_t^C \sim Q_{A_t^C}(\cdot \mid X_t^C)$$

and assigns that same physical draw to every policy in C. Distinct components receive conditionally independent fresh rewards. Equivalently, a child inherits its parent's reward on a matched edge and opens a new reward lineage on a mismatch.

The round-synchronous order is important for adaptive policies: all policies first generate their round-t observations and then update to round $t+1$. Within a round, however, every child at the same tree depth can be coupled in parallel once its parent pair is available, and the reward queries for distinct matched components can also be issued in parallel. For nonadaptive policies, different rounds may additionally be parallelized because there is no history dependence across t.

4.4 Tree choices and when they help

A *star* rooted at a designated baseline r is the simplest choice. It makes every baseline–alternative pair maximal and has depth one, so all alternatives can be processed in parallel after the baseline pair is available. Its cost, however, charges disagreement with the baseline $J-1$ times; it is therefore most attractive when the incumbent policy is representative of the alternatives or when baseline-versus-alternative contrasts are the main inferential targets. We denote this specialization by `TCAB.STAR`.

A general tree can connect similar policies through intermediate policies. If all current pairwise distances $\delta_{ij,t}(H_{t-1}^i, H_{t-1}^j)$ were available, a minimum-spanning tree minimizes their sum and hence the current-round expected query cost within the edge-local tree class; Section 5 states this formally. Deterministic tie-breaking makes such a rule predictable. The guarantee is deliberately myopic: today's joint coupling can affect cross-policy dependence in future histories, so a sequence of greedy MSTs need not be globally horizon-optimal.

In practice, current conditional TV distances are rarely available exactly. Our `TCAB.MST` implementation uses pilot data or a simulator to estimate a fixed pairwise similarity score—for example,

average complete-pair mismatch over pilot trajectories—constructs the corresponding MST, and freezes it before the main experiment. For adaptive policies the pilot score can depend on the pilot history distribution and coupling, so pilot construction, random seeds, and separation from the main experiment should be reported.

These two choices illustrate where the method gains most. A star is natural for product launches that compare many challengers to a deployed incumbent. An MST is natural when candidates form clusters—for example, nearby checkpoints, prompt variants, exploration parameters, or recommendation rules—because an intermediate policy can relay feedback between alternatives that are not both close to the baseline. The exact cost identity below makes this intuition quantitative: only disagreement on the selected tree edges creates additional reward queries.

5 Exactness and reward-query efficiency

For $e = \{i, j\} \in E_t$, define

$$D_{e,t} := \mathbb{1}\{Z_t^i \neq Z_t^j\}, \qquad N_t := \text{number of reward queries at round } t, \qquad N(T) := \sum_{t=1}^{T} N_t.$$

The component construction gives

$$N_t = 1 + \sum_{e \in E_t} D_{e,t}, \qquad N(T) = T + \sum_{t=1}^{T} \sum_{e \in E_t} D_{e,t} \quad \text{pathwise.} \tag{9}$$

Since every tree has $J - 1$ edges,

$$T \leq N(T) \leq JT \quad \text{almost surely.} \tag{10}$$

5.1 Exact trajectory marginals and unbiased contrasts

Theorem 5.1 (Exact marginal trajectories for predictable tree sequences)**.** *Under Assumption 3.1, for every finite T, J, every predictable rooted-tree sequence $(\mathcal{T}_t)_{t \leq T}$, 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 (2), for every $j \in \mathcal{I}$.

The theorem is a complete path-law statement. Policies are dependent across j, but no individual policy can statistically distinguish its `TCAB` trajectory from a standalone run.

Corollary 5.1 (Unbiased values and baseline contrasts)**.** *If rewards have finite first moments, then*

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

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

5.2 Exact query cost and the scope of optimality

The lower-bound statement requires a conditional notion of exactness. Marginal equality of a policy's complete trajectory alone does not imply that its next-step law remains correct after conditioning on other policies' histories.

Definition 5.1 (Conditionally exact design)**.** *A multi-policy design is* conditionally exact *if, for every round t and policy j,*

$$\mathcal{L}(Z_t^j \mid \mathcal{F}_{t-1}) = \nu_{j,t}^{H_{t-1}^j} \quad \textit{almost surely,}$$

and the reward assigned to j, conditional on $\mathcal{F}_{t-1}$ and $Z_t^j = (x, a)$, has law $Q_a(\cdot \mid x)$. The design is edge-local *on a rooted tree if a nonroot policy can inherit an existing reward lineage only from its parent and only when the two complete pairs coincide; otherwise the child opens a fresh reward query that may be inherited by its descendants.*

Theorem 5.2 (Conditional cost identity and edge-local optimality)**.** *Let*

$$\delta_{e,t} := \mathsf{TV}\left(\nu_{i,t}^{H_{t-1}^i}, \nu_{j,t}^{H_{t-1}^j}\right), \qquad e = \{i, j\} \in E_t.$$

For `TCAB`*,*

$$\mathbb{E}[N_t \mid \mathcal{F}_{t-1}] = 1 + \sum_{e \in E_t} \delta_{e,t}, \qquad \mathbb{E}[N(T)] = T + \sum_{t=1}^{T} \mathbb{E}\left[\sum_{e \in E_t} \delta_{e,t}\right]. \tag{11}$$

Conditional on the same pre-round information and selected tree, every conditionally exact edge-local design has expected round-t cost at least $1 + \sum_{e \in E_t} \delta_{e,t}$. Thus `TCAB` *is conditionally optimal in that class.*

This theorem does not claim global instancewise optimality among arbitrary multi-marginal designs. A non-tree design may sometimes share a reward across nonadjacent policies and beat every tree. The tree restriction buys an explicit sampler, simultaneous edge maximality, and exact cost accounting.

Corollary 5.2 (Baseline-centered star)**.** *For the fixed star rooted at r,*

$$\mathbb{E}[N(T)] = T + \sum_{t=1}^{T} \sum_{j \neq r} \mathbb{E}\left[\mathsf{TV}\left(\nu_{r,t}^{H_{t-1}^r}, \nu_{j,t}^{H_{t-1}^j}\right)\right]. \tag{12}$$

Within the class of conditionally exact baseline-local designs, `TCAB.STAR` *minimizes each round's conditional expected cost.*

Corollary 5.3 (Myopic MST optimality within tree designs)**.** *Conditional on $\mathcal{F}_{t-1}$, suppose every pairwise distance $\delta_{ij,t}$ is available. The smallest round-t conditional expected cost among conditionally exact edge-local tree designs is*

$$1 + \min_{\mathcal{T} \text{ spanning tree}} \sum_{\{i,j\} \in E(\mathcal{T})} \delta_{ij,t}. \tag{13}$$

A predictably selected minimum-spanning tree followed by tree-maximal coupling attains this value. The result is current-round optimality; it does not establish full-horizon optimality of the greedy MST sequence.

When a fixed star uses a baseline trajectory from historical logs, the leading T queries may be avoided only if the log itself has the exact standalone law required here and the policies can be replayed against the stored baseline history. This observation does not automatically extend to a changing-root adaptive tree.

5.3 Low regret implies near-one-trajectory cost

Assume rewards lie in $[0, 1]$ for this subsection. Define

$$\mu_*(x) := \max_{a \in \mathcal{A}} \mu_a(x), \qquad \Delta_a(x) := \mu_*(x) - \mu_a(x).$$

When the maximizer is unique, let

$$a^*(x) := \arg\max_a \mu_a(x), \qquad \Delta(x) := \min_{a \neq a^*(x)} \Delta_a(x).$$

Policy j's pseudo-regret is

$$\mathcal{R}_j(T) := \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{A_t^j}(X_t^j)\right]. \tag{14}$$

Let $d_{j,t}$ be the degree of j in $\mathcal{T}_t$ and define the degree-weighted regret

$$\mathcal{R}_{\deg}(T) := \mathbb{E}\left[\sum_{t=1}^{T} \sum_{j \in \mathcal{I}} d_{j,t} \Delta_{A_t^j}(X_t^j)\right]. \tag{15}$$

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

$$\mathcal{R}_{\deg}(T) \leq (J-1) \sum_{j \in \mathcal{I}} \mathcal{R}_j(T). \tag{16}$$

Theorem 5.3 (Finite-sample regret-to-cost bound)**.** *Suppose the oracle action is unique P_X-almost surely and define*

$$F_\Delta(\varepsilon) := \mathbb{P}\{\Delta(X) \leq \varepsilon\}.$$

For every predictable tree sequence and every $\varepsilon > 0$,

$$\mathbb{E}[N(T)] - T \leq 2(J-1)TF_\Delta(\varepsilon) + \frac{\mathcal{R}_{\deg}(T)}{\varepsilon}. \tag{17}$$

Consequently, if J is fixed and $\mathcal{R}_j(T) = o(T)$ for every j, then

$$\mathbb{E}[N(T)] = T + o(T). \tag{18}$$

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

Corollary 5.4 (Margin and uniform-gap rates)**.** *If $F_\Delta(\varepsilon) \leq C\varepsilon^\beta$ for all sufficiently small ε, with $C, \beta > 0$, then*

$$\mathbb{E}[N(T)] - T = O\Big(((J-1)T)^{1/(\beta+1)}\, \mathcal{R}_{\deg}(T)^{\beta/(\beta+1)}\Big), \tag{19}$$

where the constant depends only on C and β. If $\Delta(X) \geq \Delta_{\min} > 0$ almost surely, then

$$\mathbb{E}[N(T)] - T \leq \frac{\mathcal{R}_{\deg}(T)}{\Delta_{\min}}. \tag{20}$$

5.4 Variance reduction for policy contrasts

Assume rewards lie in $[0, 1]$. Define 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

$$B_{ij}(T) := 15 \sum_{t=1}^{T} \mathbb{E}\left[\sum_{e \in P_{ij,t}} D_{e,t}\right] + 6\,\mathrm{Var}(L_i(T)) + 6\,\mathrm{Var}(L_j(T)). \tag{21}$$

Theorem 5.4 (Finite-sample pairwise-contrast variance)**.** *For every $i, j \in \mathcal{I}$,*

$$\mathrm{Var}(S_j(T) - S_i(T)) \le B_{ij}(T). \tag{22}$$

In particular, if the expected number of mismatches along the time-varying paths $P_{ij,t}$ is $o(T)$ and both realized-regret variances are $o(T)$, then

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

Theorem 5.4 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 E.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. 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 design. 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.

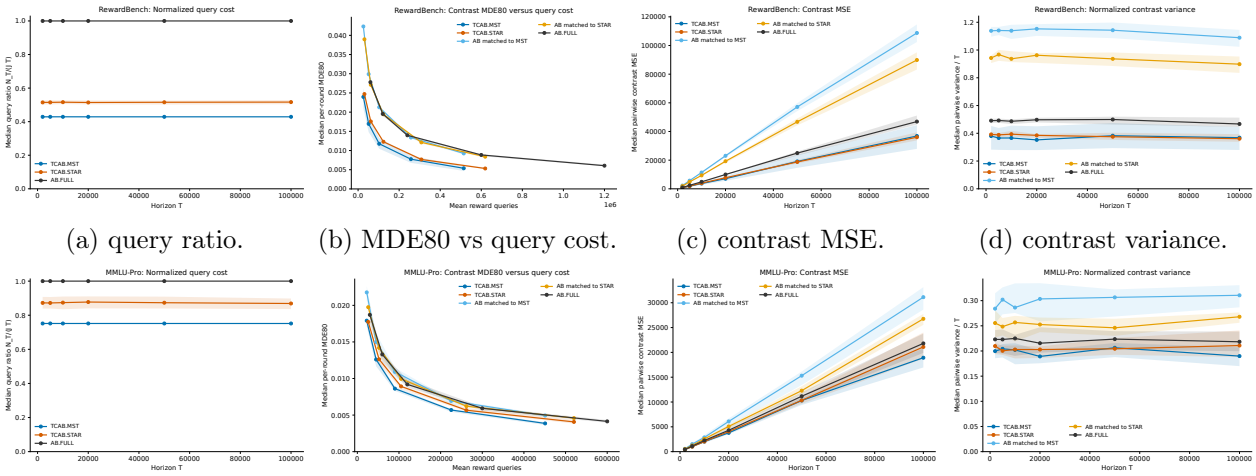
Across the three experiments, `TCAB` improves the cost–precision frontier while preserving the target contrasts. At matched query budgets, it reduces median contrast error and variance relative to independent A/B/n designs, detects smaller policy differences, and uses substantially fewer reward observations. 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.



(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.

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 80% and 40% of the full query budget on RewardBench and MMLU-Pro, respectively.[1] On RewardBench, the reported me-

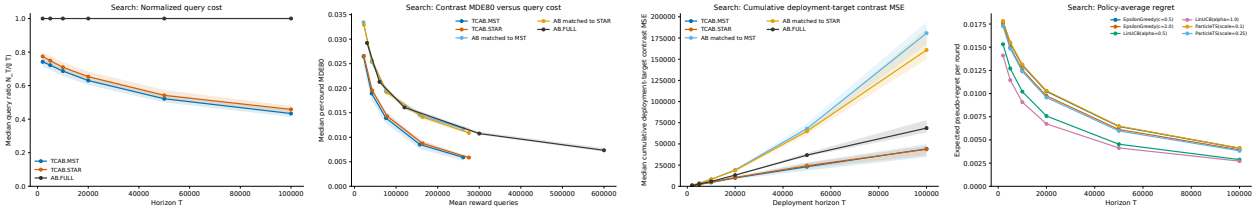
[1]For history-free, time-invariant policies, every edge distance is a constant δ_e, so Theorem 5.2 gives $\mathbb{E}[N(T)] =$

dian 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%.

6.2 Adaptive online learning policies

To further highlight the benefit of `TCAB`, we consider a benchmark with adaptive policies. By Theorem 5.3, 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, Qin and Liu, 2013]. 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 `TCAB` in 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 5.3, 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 rescales it as $TS_j(q_T)/q_T$. 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.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

$T\{1 + \sum_{e \in E} \delta_e\}$. Hence the normalized query ratio is constant in T, as observed.

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. 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*, pages 1638–1646. PMLR, 2014.

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

S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 99–107. PMLR, 2013a.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135. PMLR, 2013b.

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

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352.

S. Banerjee, S. R. Sinclair, M. Tambe, L. Xu, and C. L. Yu. Artificial replay: A meta-algorithm for harnessing historical data in bandits. *arXiv preprint arXiv:2210.00025*, 2022.

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

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 19–26. PMLR, 2011.

J. Brennan, Y. Cong, Y. Yu, L. Lin, Y. Peng, C. Meng, N. Han, J. Pouget-Abadie, and D. M. Holtz. Reducing symbiosis bias through better A/B tests of recommendation algorithms. In *Proceedings of the ACM Web Conference 2025*, pages 3702–3715. ACM, 2025. doi: 10.1145/3696410.3714738.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, and I. 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*, pages 8359–8388. PMLR, 2024.

M. C. Cohen, I. Lobel, and R. Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

A. Corenflos and S. Särkkä. The coupled rejection sampler. *arXiv preprint arXiv:2201.09585*, 2022.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011.

C. 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.

A. Ghosh and A. 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*, pages 7531–7549. PMLR, 2022.

P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2003.

S. Gupta, R. Kohavi, D. Tang, Y. Xu, R. Andersen, E. Bakshy, N. Cardin, S. Chandran, N. Chen, D. 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.

V. Hadad, D. A. Hirshberg, R. Zhan, S. Wager, and S. 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.

O. A. Hanna, L. Yang, and C. 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*, pages 1791–1821. PMLR, 2023.

Z. Hu, Y. Wen, Y. Yao, J. Zhang, and Z. Zhou. Learning to bid with unknown private values in budget-constrained first-price auctions. *arXiv preprint arXiv:2605.09448*, 2026.

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

J. P. C. Kleijnen. Antithetic variates, common random numbers and optimal computer time allocation in simulation. *Management Science*, 21(10):1176–1185, 1975. doi: 10.1287/mnsc.21.10.1176.

R. Kleinberg and T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1168–1176. ACM, 2013. doi: 10.1145/2487575.2488217.

I. Kuzborskij, C. Vernade, A. Gyorgy, and C. 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*, pages 640–648. PMLR, 2021.

N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, 2025.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, 2020.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

L. Li, W. Chu, J. Langford, and X. 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*, pages 297–306. ACM, 2011. doi: 10.1145/1935826.1935878.

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

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

H. Meng, N. Chen, and X. Gao. Design experiments to compare multi-armed bandit algorithms. *arXiv preprint arXiv:2603.05919*, 2026.

Microsoft Research. Microsoft learning to rank datasets. MSLR-WEB10K, 2010.

B. L. Nelson and F. J. Matejcik. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science*, 41(12):1935–1945, 1995. doi: 10.1287/mnsc.41.12.1935.

B. Pass. Multi-marginal optimal transport: Theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1771–1790, 2015. doi: 10.1051/m2an/2015020.

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

T. Qin and T. Liu. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013. URL http://arxiv.org/abs/1306.2597.

D. Simchi-Levi and C. 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.

Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. 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*, pages 9167–9176. PMLR, 2020.

A. Swaminathan and T. 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*, pages 814–823. PMLR, 2015.

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

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

C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009. doi: 10.1007/978-3-540-71050-9.

R. Wan, B. Kveton, and R. 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*, pages 22491–22511. PMLR, 2022.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Y.-X. Wang, A. Agarwal, and M. 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*, pages 3589–3597. PMLR, 2017.

J. Weed, V. Perchet, and P. Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR, 2016.

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

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

Y. Wen, Z. Hu, Y. Han, Y. Yao, and Z. Zhou. The (marginal) value of a search ad: An online causal framework for repeated second-price auctions. In *Forty-third International Conference on Machine Learning*, 2026b.

W. Wu, J. Yang, and C. 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*, pages 2392–2401. PMLR, 2020.

Y. Xu, N. Chen, A. Fernandez, O. Sinno, and A. 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*, pages 2227–2236. ACM, 2015. doi: 10.1145/2783258.2788602.

M. Yang, B. Dai, O. Nachum, G. Tucker, and D. 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*, pages 4376–4396. PMLR, 2022.

R. Zhan, V. Hadad, D. A. Hirshberg, and S. 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*, pages 2125–2135. ACM, 2021.

A Coupling preliminaries

A.1 Two-marginal maximal coupling

Let $(\mathsf{Z}, \mathcal{Z})$ be standard Borel and let ν_0, ν_1 be dominated by ρ, with densities p_0, p_1. Define

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

When $\delta > 0$, define

$$\widetilde{\nu}_i(\,\mathrm{d}z) := \frac{\nu_i(\,\mathrm{d}z) - \lambda(\,\mathrm{d}z)}{\delta}, \qquad i \in \{0, 1\}.$$

Lemma A.1 (Canonical maximal coupling)**.** *With probability α, draw Z from λ/α and set $Z_0 = Z_1 = Z$; this branch is absent when $\alpha = 0$. With probability δ, draw (Z_0, Z_1) from any coupling of $(\widetilde{\nu}_0, \widetilde{\nu}_1)$; this branch is absent when $\delta = 0$. Then $Z_i \sim \nu_i$ and*

$$\mathbb{P}(Z_0 \neq Z_1) = \delta = \mathsf{TV}(\nu_0, \nu_1).$$

No coupling has a smaller disagreement probability.

Proof. The marginal decomposition $\nu_i = \lambda + \delta\widetilde{\nu}_i$ gives the required laws. The residual measures are mutually singular because their densities are supported on $\{p_0 > p_1\}$ and $\{p_1 > p_0\}$, respectively, so the residual branch cannot agree with positive probability. Also

$$\mathsf{TV}(\nu_0, \nu_1) = 1 - \int \min\{p_0, p_1\}\,\mathrm{d}\rho = \delta.$$

For any coupling, the common-value submeasure is dominated by both marginals and therefore has mass at most $\int \min\{p_0, p_1\}\,\mathrm{d}\rho = \alpha$. Hence $\mathbb{P}(Z_0 \neq Z_1) \geq \delta$. See Thorisson [2000] and Lindvall [2002]. □

Lemma A.2 (One-sided rejection sampler)**.** *Suppose samples from ν_0, ν_1 and the density ratios are available. Draw $Z_0 \sim \nu_0$. Set $Z_1 = Z_0$ with probability $\min\{1, p_1(Z_0)/p_0(Z_0)\}$. Otherwise repeatedly draw $Y \sim \nu_1$ until it is accepted with probability*

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

and set $Z_1 = Y$. This is a maximal coupling. The first-stage match probability is α. If $\delta > 0$, conditional on entering the residual branch, the number of ν_1 proposals is geometric with success probability δ and mean $1/\delta$; when $\delta = 0$, the branch is never entered.

Proof. The accepted first-stage measure is

$$p_0(z)\min\{1, p_1(z)/p_0(z)\}\rho(\,\mathrm{d}z) = \lambda(\,\mathrm{d}z).$$

A residual proposal is accepted with unnormalized density

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

whose total mass is δ. Thus the accepted residual has law $\widetilde{\nu}_1$ and the proposal count has the stated geometric law. □

A.2 Tree gluing

Lemma A.3 (Tree gluing lemma)**.** *Let $\mathcal{T} = (\mathcal{I}, E)$ be a finite rooted tree on a standard Borel space Z. For every directed edge $e = (p(v), v)$, let γ_e be a coupling of $\nu_{p(v)}$ and ν_v. Then there is a joint coupling of $(\nu_j)_{j\in\mathcal{I}}$ whose marginal on every edge is γ_e. In particular, maximal couplings can be attained simultaneously on all tree edges.*

Proof. Disintegrate

$$\gamma_e(\,\mathrm{d}z_p,\,\mathrm{d}z_v) = \nu_p(\,\mathrm{d}z_p)K_e(z_p,\,\mathrm{d}z_v),$$

which is possible on standard Borel spaces. Draw $Z_r \sim \nu_r$ and then recursively draw

$$Z_v \sim K_{(p(v),v)}(Z_{p(v)}, \cdot),$$

conditionally independently across siblings if desired. Induction over depth gives every node marginal ν_v, and the construction gives the prescribed edge law. Acyclicity removes compatibility constraints that can arise on graphs with cycles. □

B Proof of the implementation proposition

Proof of Proposition 4.1. Fix a round, an oriented edge (p, v), and the two current histories. Apply Lemma A.2 to

$$\nu_p(\,\mathrm{d}x, a) = P_X(\,\mathrm{d}x)\pi_{p,t}(a \mid h_p, x), \qquad \nu_v(\,\mathrm{d}x, a) = P_X(\,\mathrm{d}x)\pi_{v,t}(a \mid h_v, x),$$

with dominating measure $P_X \otimes$ counting. The common P_X factor cancels, yielding exactly (7) and (8). Lemma A.3 glues these edge laws into the round-t tree coupling used in Algorithm 1. The residual branch is entered with probability $\delta_{pv,t}$ and, when positive, its conditional proposal mean is $1/\delta_{pv,t}$; its unconditional contribution is therefore one. Summing over at most $J-1$ edges in each of T rounds proves the proposal bound. Reward sampling is performed only after the complete-pair coupling and therefore does not enter the acceptance ratios. □

C Proofs of exactness and query cost

Proof of Theorem 5.1. Fix t and condition on $\mathcal{F}_{t-1}$. The tree $\mathcal{T}_t$ and every policy history are then fixed. By the canonical edge construction and Lemma A.3,

$$\mathcal{L}(Z_t^j \mid \mathcal{F}_{t-1}) = \nu_{j,t}^{H_{t-1}^j} \tag{23}$$

for every j. In particular, for measurable $B \subseteq \mathcal{X}$ and $a \in \mathcal{A}$,

$$\mathbb{P}(X_t^j \in B, A_t^j = a \mid \mathcal{F}_{t-1}) = \int_B \pi_{j,t}(a \mid H_{t-1}^j, x)P_X(\,\mathrm{d}x).$$

Now condition further on all round-t complete pairs and their matched components. A component with common pair (x, a) receives a fresh draw from $Q_a(\cdot \mid x)$, independent of the coupling randomness. Thus, for every node,

$$\mathcal{L}(R_t^j \mid \mathcal{F}_{t-1}, Z_t^j = (x, a)) = Q_a(\cdot \mid x). \tag{24}$$

Sharing the draw across a component changes cross-policy dependence but not any node's conditional law. Equations (23)–(24) are exactly the standalone transition (2). Starting from empty histories and iterating proves equality of each complete marginal path law by induction, equivalently by uniqueness in the Ionescu–Tulcea theorem. □

Proof of Corollary 5.1. Theorem 5.1 gives the standalone distribution, hence expectation, of every $S_j(T)$. Linearity gives the baseline-contrast identities. □

Proof of Theorem 5.2. Removing the mismatched edges from a tree creates exactly one more component per removed edge. Because `TCAB` queries one reward per component, (9) holds. Conditional on $\mathcal{F}_{t-1}$, every selected edge is a maximal coupling, so

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

Taking conditional expectation in the pathwise identity, then summing and taking ordinary expectations, proves (11).

For the lower bound, fix $\mathcal{F}_{t-1}$ and consider any conditionally exact edge-local design on the same selected tree. On edge $e = \{i, j\}$, conditional exactness fixes the two node marginals as $\nu_{i,t}^{H_{t-1}^i}$ and $\nu_{j,t}^{H_{t-1}^j}$. The coupling inequality therefore gives

$$\mathbb{P}(Z_t^i \neq Z_t^j \mid \mathcal{F}_{t-1}) \geq \delta_{e,t}.$$

Edge locality forces a new reward lineage whenever a child mismatches its parent, so its conditional expected cost is at least

$$1 + \sum_{e \in E_t} \mathbb{P}(Z_t^i \neq Z_t^j \mid \mathcal{F}_{t-1}) \geq 1 + \sum_{e \in E_t} \delta_{e,t}.$$

Tree gluing shows that all edgewise lower bounds can coexist, and `TCAB` attains them. □

Proof of Corollary 5.2. Apply Theorem 5.2 to the edges $\{r, j\}$, $j \neq r$. The lower-bound class specializes to designs in which every alternative can inherit only from the baseline. □

Proof of Corollary 5.3. For any selected tree, Theorem 5.2 gives minimal conditional tree-local cost $1 + \sum_{e \in E} \delta_{e,t}$. Minimizing this sum over spanning trees is exactly the minimum-spanning-tree problem. There are finitely many labeled trees; a deterministic lexicographic rule among minimizers makes the choice measurable in the edge weights and hence predictable. Applying the tree coupling preserves every node marginal. The argument is pointwise in the current histories and therefore establishes only the stated current-round result. □

D Proofs of regret-based efficiency

Proof of Theorem 5.3. For history h_j, define the conditional nonoracle probability

$$q_{j,t}(h_j) := \int_{\mathcal{X}} [1 - \pi_{j,t}(a^*(x) \mid h_j, x)] \, P_X(\mathrm{d}x).$$

For an edge $\{i, j\}$ and fixed context x, write p_i, p_j for the two current action distributions. Their overlap includes at least the smaller mass assigned to the common oracle action, so

$$\begin{aligned} \mathsf{TV}(p_i, p_j) &= 1 - \sum_a \min\{p_i(a), p_j(a)\} \\ &\leq 1 - \min\{p_i(a^*(x)), p_j(a^*(x))\} \\ &\leq [1 - p_i(a^*(x))] + [1 - p_j(a^*(x))]. \end{aligned}$$

Integrating over P_X gives

$$\delta_{ij,t} \le q_{i,t}(H^i_{t-1}) + q_{j,t}(H^j_{t-1}).$$

Summing over the round-t tree and then over time yields

$$\mathbb{E}[N(T)] - T \le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j\in\mathcal{I}} d_{j,t}\mathbb{1}\{A^j_t \neq a^*(X^j_t)\}\right]. \tag{25}$$

Here we used conditional exactness of `TCAB` to identify $q_{j,t}$ with the conditional nonoracle probability of policy j.

For every $\varepsilon > 0$,

$$d_{j,t}\mathbb{1}\{A^j_t \neq a^*(X^j_t)\} \le d_{j,t}\mathbb{1}\{\Delta(X^j_t) \le \varepsilon\} + \frac{d_{j,t}}{\varepsilon}\Delta_{A^j_t}(X^j_t).$$

Because $d_{j,t}$ is $\mathcal{F}_{t-1}$-measurable and X^j_t has conditional marginal P_X,

$$\sum_{j\in\mathcal{I}} \mathbb{E}\Big[d_{j,t}\mathbb{1}\{\Delta(X^j_t) \le \varepsilon\}\Big] = F_\Delta(\varepsilon)\mathbb{E}\left[\sum_j d_{j,t}\right] = 2(J-1)F_\Delta(\varepsilon).$$

Summing over t and using (15) in (25) proves (17).

For fixed J, (16) and $\mathcal{R}_j(T) = o(T)$ imply $r_T := \mathcal{R}_{\deg}(T)/T \to 0$. If $r_T > 0$, choose $\varepsilon_T = \sqrt{r_T}$; otherwise choose any deterministic $\varepsilon_T \downarrow 0$. Almost-sure uniqueness and finite K imply $\Delta(X) > 0$ almost surely and hence $F_\Delta(\varepsilon_T) \to 0$. Divide (17) by T to obtain (18). □

Proof of Corollary 5.4. Under the margin condition, (17) is at most

$$2C(J-1)T\varepsilon^\beta + \frac{\mathcal{R}_{\deg}(T)}{\varepsilon}.$$

Balancing the two terms gives

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

which proves (19). Under a uniform gap, every nonoracle action incurs loss at least $\Delta_{\min}$, so the right side of (25) is at most $\mathcal{R}_{\deg}(T)/\Delta_{\min}$. □

E Proofs of the variance results

Proof of Theorem 5.4. Fix i, j and define

$$U^{ij}_t := \mathbb{1}\{\text{at least one edge in } P_{ij,t} \text{ mismatches}\}.$$

Then

$$U^{ij}_t \le \sum_{e\in P_{ij,t}} D_{e,t}. \tag{26}$$

If $U^{ij}_t = 0$, equality propagates along the matched path, so $Z^i_t = Z^j_t$ and $R^i_t = R^j_t$.

Define

$$C^{ij}_t := \mu_*(X^j_t) - \mu_*(X^i_t).$$

Both endpoint contexts have conditional marginal P_X, hence $\mathbb{E}[C_t^{ij} \mid \mathcal{F}_{t-1}] = 0$. Also $C_t^{ij} = 0$ when $U_t^{ij} = 0$ and $|C_t^{ij}| \le 1$. Therefore

$$\mathrm{Var}\left(\sum_{t=1}^{T} C_t^{ij}\right) \le \mathbb{E}\left[\sum_{t=1}^{T} U_t^{ij}\right]. \tag{27}$$

Let $\mathcal{G}_t$ enlarge $\mathcal{F}_{t-1}$ by all round-t complete pairs, the tree, and the component partition, but not the reward draws. Define

$$E_t^{ij} := \{R_t^j - \mu_{A_t^j}(X_t^j)\} - \{R_t^i - \mu_{A_t^i}(X_t^i)\}.$$

Conditional on $\mathcal{G}_t$, each residual has mean zero. Thus $\mathbb{E}[E_t^{ij} \mid \mathcal{G}_t] = 0$. Moreover, $E_t^{ij} = 0$ when $U_t^{ij} = 0$ and $|E_t^{ij}| \le 2U_t^{ij}$, giving

$$\mathrm{Var}\left(\sum_{t=1}^{T} E_t^{ij}\right) \le 4\mathbb{E}\left[\sum_{t=1}^{T} U_t^{ij}\right]. \tag{28}$$

Writing $\ell_t^j := \Delta_{A_t^j}(X_t^j)$,

$$R_t^j - R_t^i = C_t^{ij} + \ell_t^i - \ell_t^j + E_t^{ij}.$$

After summing over t, use

$$\mathrm{Var}(U+V+W) \le 3\{\mathrm{Var}(U) + \mathrm{Var}(V) + \mathrm{Var}(W)\}, \qquad \mathrm{Var}(L_i - L_j) \le 2\,\mathrm{Var}(L_i) + 2\,\mathrm{Var}(L_j),$$

together with (26)–(28). This yields

$$\mathrm{Var}(S_j - S_i) \le 15 \sum_{t=1}^{T} \mathbb{E}\left[\sum_{e \in P_{ij,t}} D_{e,t}\right] + 6\,\mathrm{Var}(L_i) + 6\,\mathrm{Var}(L_j),$$

which is (22). □

E.1 General fixed zero-sum contrasts

The pairwise result immediately extends to comparative linear functionals. Let

$$S(T) := (S_0(T), \ldots, S_{J-1}(T))^\top, \qquad c \in \mathbb{R}^J, \qquad \mathbf{1}^\top c = 0,$$

and define the corresponding target by

$$\theta_c(T) := c^\top V(T).$$

By Theorem 5.1, $c^\top S(T)$ is unbiased for $\theta_c(T)$ whenever rewards have finite first moments.

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

$$\mathrm{Var}\Big(c^\top S(T)\Big) \le \left(\sum_{j \ne r} |c_j| \sqrt{B_{jr}(T)}\right)^2. \tag{29}$$

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

Corollary E.1 (Simple second-moment condition)**.** *Let*

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

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 of Theorem E.1. Because $\sum_j c_j = 0$,

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

Center both sides and apply Minkowski's inequality in L^2:

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

Theorem 5.4 gives (29). Moreover,

$$\sum_{t=1}^{T} \mathbb{E}\left[\sum_{e \in P_{jr,t}} D_{e,t}\right] \leq \mathbb{E}[N(T)] - T.$$

Thus every $B_{jr}(T) = o(T)$ under the stated assumptions, and a fixed finite sum of $o(\sqrt{T})$ terms has square $o(T)$. □

Proof of Corollary E.1. Since rewards lie in $[0, 1]$, every gap lies in $[0, 1]$ and

$$0 \leq L_j(T) \leq M_j(T).$$

Hence

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

Apply Theorem E.1. □