UNCERTAINTY QUANTIFICATION OF RIEMANNIAN LOW-RANK TENSOR BANDITS WITH ROTATED ELIMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank tensor bandits model multi-dimensional online decision problems, such as joint user segment–product–channel selection in online advertising. Existing tensor bandit algorithms focus on reward maximization but do not adequately address uncertainty quantification. To fill this gap, we study an inferential tensor bandit framework with a low-rank reward tensor and adaptively generated, batched online data. Two challenges arise. First, low-regret policies may prematurely stop sampling actions relevant to the inferential target, leaving insufficient information in its direction. Second, the low-rank constraint introduces nonlinear estimation bias, whereas policy-dependent sampling distributions invalidate conventional inference methods. We propose Rotated Inferential Tensor Elimination (RITE), an online, batched non-contextual bandit algorithm with a fixed rotated-elimination rule for decisions and an evolving low-rank nuisance estimator for quantification. RITE combines spectral rotation, maximum-confidence-width exploration, and active-set elimination, and uses Riemannian capped inverse-propensity updates for low-rank estimation. For inference, it constructs a stabilized one-step debiasing estimator that simultaneously handles low-rank estimation error and adaptive sampling. Theoretically, we establish a nonasymptotic error bound for Riemannian low-rank estimation, prove asymptotic normality for the efficient inference, and derive a finite-time regret bound that preserves the typical order of tensor elimination while adding only the cost of target exploration. Simulations and an online advertising experiment show that RITE yields confidence intervals with near-nominal coverage while maintaining regret close to tensor elimination.

1 INTRODUCTION

Multi-dimensional online decision problems arise in online advertising, recommender systems, and personalized marketing. An advertising platform may jointly select a user segment, product type, and delivery channel; a recommendation action may likewise be determined by a user group, item category, and display position. Such actions have a natural multi-way combinatorial structure, and arranging the mean rewards of all action combinations yields a higher-order reward tensor. Tensor decompositions provide a classical tool for representing multi-way interactions (Kolda & Bader, 2009) and have been used in recommender systems (Frolov & Oseledets, 2017), marketing analytics (Shan et al., 2016), and high-dimensional multimodal data modeling (Bi et al., 2018). A low-rank tensor structure means that a small number of latent factors may govern the reward pattern even when the combinatorial action space is large.

Bandit algorithms provide the basic framework for balancing exploration and exploitation in sequential decisions (Lattimore & Szepesvári, 2020). Linear and contextual bandits exploit feature structure to reduce the cost of exploration (Abbasi-Yadkori et al., 2011; Chu et al., 2011), while low-rank bandit methods use lower-dimensional structure to improve regret bounds in high-dimensional action spaces (Jun et al., 2019; Lu et al., 2021). Recent low-rank tensor bandits extend this principle to multi-way combinatorial actions (Shi et al., 2023; Zhou et al., 2024). Tensor elimination, in particular, estimates mode-wise factor spaces, transforms the original tensor bandit into a lower-dimensional linear bandit through spectral rotation, and then performs confidence-width exploration

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

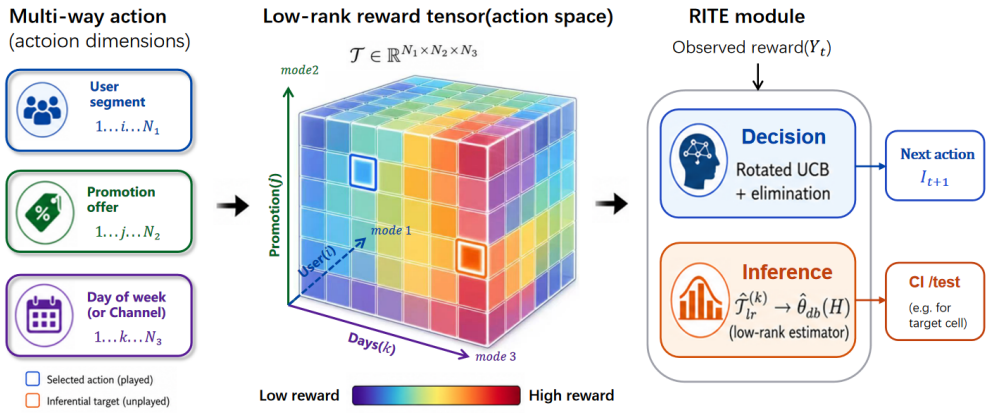

Figure 1: Overview of RITE. A multi-way choice is encoded as a one-hot tensor action $\mathcal{A}_t$ and yields $Y_t = \langle \mathcal{T}^\star, \mathcal{A}_t \rangle + \xi_t$. Proposal observations update the rotated-UCB elimination branch, whereas all observations enter the low-rank estimation and debiasing branch, which constructs confidence intervals and tests for the prespecified target $\theta^\star(H)$.

and active-set elimination. These methods show that low-rank tensor structure can substantially improve regret relative to naive vectorized UCB and matricized bandit methods. Figure 1 illustrates the multi-way action structure and the separate decision and inference branches of RITE.

Low regret, however, does not automatically yield valid inference. Existing low-rank tensor bandit algorithms primarily maximize reward by identifying high-reward actions and controlling cumulative regret. Many applications also require statistical answers. What is the true mean reward of a particular action? Is the reward difference between two candidate actions significant? An advertising platform may wish to compare the true returns of email and social-media channels for the same product, while a recommender system may test whether a user group responds differently to two recommendation policies. These questions are linear functionals of the low-rank reward tensor.

Low-rank matrix inference generally requires correction of rank-induced bias (Chen et al., 2019). Offline low-rank tensor inference has been developed for regression, completion, and heteroskedastic-noise models (Xia et al., 2022; Ma & Xia, 2024; Agterberg & Zhang, 2024). Bandit inference is further complicated by the interaction between low-rank structure and adaptive sampling. Rank constraints and nonlinear projection give the usual low-rank estimator a first-order bias, so confidence intervals based directly on a plug-in estimator may have distorted coverage (Ma & Xia, 2024). At the same time, elimination policies gradually stop sampling actions that appear suboptimal. This behavior reduces regret but can leave insufficient information in directions relevant to the inferential target. Related inference problems have been studied for contextual bandits and matrix contextual bandits (Chen et al., 2021; Han et al., 2025), but those methods do not directly address low-rank tensor rewards, spectral rotation, multi-way structure, and the nonconvex estimation induced by tensor elimination. Recent work also studies online low-rank tensor inference (Wen et al., 2026); its observations, however, are generally not selected by an adaptive reward-maximizing policy and therefore do not directly resolve inference under tensor bandit sampling.

This paper considers inference for tensor entries and action differences. RITE retains the warm-up spectral rotation and active-set elimination rule of tensor elimination (Zhou et al., 2024), and adds information in the target direction through low-probability target-specific exploration. Riemannian capped-IPW regression updates the low-rank nuisance estimator and supplies a batch-start tangent space. The score defined by a fixed reference distribution q_H is stabilized by $q_H(I_t)/\pi_t(I_t)$ and therefore forms an exact martingale difference. This construction draws primarily on inference for low-rank tensor functionals (Ma & Xia, 2024), online tensor inference (Wen et al., 2026), and inference under adaptive decision-making (Han et al., 2025).

We establish a nonasymptotic low-rank estimation bound, asymptotic normality of the doubly debiased estimator, and a finite-time regret bound. When $\varepsilon_t \asymp t^{-1/2}$, the regret introduced by the target branch is of order $\sqrt{T}$, while confidence intervals for the fixed target are asymptotically valid.

2 TENSOR BANDIT MODEL AND NOTATION

2.1 NOTATION AND TENSOR ALGEBRA

For a positive integer p, let $[p] = \{1, \dots, p\}$ and define the multi-dimensional action index set $\mathcal{I} = [p_1] \times \cdots \times [p_d]$. For $i = (i_1, \dots, i_d) \in \mathcal{I}$, the corresponding one-hot tensor is $\mathcal{A}(i) = e_{i_1} \circ \cdots \circ e_{i_d}$, where $e_{i_j} \in \mathbb{R}^{p_j}$ is a standard basis vector in mode j and $\circ$ denotes the outer product. For tensors $\mathcal{X}$ and $\mathcal{Y}$ of the same size, $\langle \mathcal{X}, \mathcal{Y} \rangle$ is the tensor inner product, $\|\mathcal{X}\|_F$ is the Frobenius norm, and $\mathrm{vec}(\mathcal{X})$ is ordinary vectorization. We use $\times_j$ for mode-j multiplication, $\mathrm{Mat}_j(\mathcal{X})$ for mode-j unfolding, and $M^\dagger$ for the Moore–Penrose pseudoinverse of a matrix M.

The true mean reward tensor has Tucker rank $\mathbf{r} = (r_1, \dots, r_d)$ and can be written as

$$\mathcal{T}^\star = \mathcal{G}^\star \times_1 U_1^\star \times_2 \cdots \times_d U_d^\star, \tag{1}$$

where $U_j^\star \in \mathbb{R}^{p_j \times r_j}$ has orthonormal columns and $\mathcal{G}^\star \in \mathbb{R}^{r_1 \times \cdots \times r_d}$ is the core tensor. The effective number of free parameters is $D_{\text{eff}} = \prod_{j=1}^d r_j + \sum_{j=1}^d (p_j - r_j) r_j$, which is much smaller than the ambient dimension $p_1 \cdots p_d$ when the mode ranks are small. We use $\widetilde{O}(\cdot)$ to suppress logarithmic factors in p and T, and state rates for the balanced case $p_j \asymp p$.

Let $\mathcal{M}_{\mathbf{r}} = \{\mathcal{T} : \mathrm{rank}_{\mathrm{Tuck}}(\mathcal{T}) = \mathbf{r}\}$ be the manifold of tensors with Tucker rank $\mathbf{r}$. Write $\mathbb{T}^\star = T_{\mathcal{T}^\star}\mathcal{M}_{\mathbf{r}}$ for its true tangent space at $\mathcal{T}^\star$, and $\widehat{\mathbb{T}}^{(k)} = T_{\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)}}\mathcal{M}_{\mathbf{r}}$ for the estimated tangent space at the low-rank estimator after batch k. Their orthogonal projections are $P_{\mathbb{T}^\star}$ and $P_{\widehat{\mathbb{T}}^{(k)}}$, respectively, and Retr denotes a retraction from a tangent space to $\mathcal{M}_{\mathbf{r}}$.

Given a fixed orthonormal basis of $\widehat{\mathbb{T}}^{(k)}$, $\mathrm{vec}_k(\mathcal{X})$ denotes the coordinate vector of $P_{\widehat{\mathbb{T}}^{(k)}}(\mathcal{X})$ in that basis. This notation is used only for local tangent-space coordinates and does not denote an inverse of ordinary vectorization. If $\widehat{U}_j^{(k)}$ is the estimated mode-j factor matrix after batch k, then $\widehat{U}_{j,\perp}^{(k)}$ denotes an orthogonal complement of its column space.

2.2 BANDIT MODEL AND INFERENTIAL TARGETS

At time t, the learner selects an index $I_t = (i_{1,t}, \dots, i_{d,t}) \in \mathcal{I}$, with corresponding one-hot action tensor $\mathcal{A}_t = \mathcal{A}(I_t) = e_{i_{1,t}} \circ \cdots \circ e_{i_{d,t}}$. It then observes

$$Y_t = \langle \mathcal{T}^\star, \mathcal{A}_t \rangle + \xi_t, \tag{2}$$

where ξ_t is conditionally mean-zero sub-Gaussian noise. Since $\mathcal{A}_t$ is one-hot, $\langle \mathcal{T}^\star, \mathcal{A}_t \rangle = \mathcal{T}^\star_{I_t}$. Actions are selected adaptively from the observed history, while the low-rank Tucker structure links their mean rewards.

Let $i^\star \in \arg\max_{i \in \mathcal{I}} \langle \mathcal{T}^\star, \mathcal{A}(i) \rangle$ and $\mathcal{A}^\star = \mathcal{A}(i^\star)$. The standard decision objective is to control cumulative regret

$$R_T = \sum_{t=1}^T \langle \mathcal{T}^\star, \mathcal{A}^\star - \mathcal{A}_t \rangle. \tag{3}$$

Direct vectorization yields a standard bandit problem with $\prod_{j=1}^d p_j$ actions. This representation discards the multi-way low-rank structure and can incur a large exploration cost in high dimensions, as also observed in related structured tensor problems (Xia et al., 2021; Bi et al., 2018). Tensor elimination uses spectral rotation and active-set elimination to exploit this structure and obtain lower finite-time regret (Zhou et al., 2024); our decision module adopts the same principle.

The inferential target is a linear functional $\theta^\star(H) = \langle \mathcal{T}^\star, H \rangle$. We focus on a single-arm mean reward and a two-arm reward gap:

$$H = \mathcal{A}(i_0) \quad \text{or} \quad H = \mathcal{A}(i_0) - \mathcal{A}(j_0), \tag{4}$$

where $i_0, j_0 \in \mathcal{I}$ are fixed before the online procedure starts. The first target is the mean reward of one action, and the second is the reward difference between two actions. After a warm-up of length s_0, set $t_1 = s_0 + 1$ and let t_k and $t_{k+1} - 1$ be the first and last times of batch k. Write $\mathcal{E}_k = \{t_k, \dots, t_{k+1} - 1\}$ and $n_k = t_{k+1} - t_k$, and let $k(t)$ denote the batch containing time t. Set $K_T = \max\{k : t_k \le T\}$. The Riemannian low-rank estimator after batch k is $\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)}$.

3 ROTATED INFERENTIAL TENSOR ELIMINATION

Tensor elimination estimates low-rank factor spaces during warm-up and then performs confidence-width exploration and active-set elimination in fixed spectral coordinates (Zhou et al., 2024). This construction avoids the high-dimensional cost of direct vectorization and substantially improves regret. RITE adopts the same decision structure. For inference, however, the difficulties described in the Introduction remain. Continuously preserving information for every action would require extensive exploration unrelated to reward maximization and would weaken the regret advantage of elimination.

Therefore, we introduce target-specific exploration only for the prespecified inferential target. RITE uses dynamic batches to align the progressive contraction of confidence intervals with phase-wise gap elimination. A Riemannian capped inverse-probability-weighted update maintains the Tucker low-rank reward tensor while mitigating the effect of adaptive nonuniform sampling on low-rank estimation. Together, these components provide uncertainty quantification for a prespecified target while preserving the leading regret order of tensor elimination. The complete RITE procedure is given in Supplementary Algorithm S1.

3.1 SPECTRAL ROTATION AND ACTIVE-SET ELIMINATION

RITE begins with an initialization period of length s_0. During this period, actions are drawn independently from a prespecified full-support distribution q_0 on $\mathcal{I}$, where $q_0(i) > 0$ for every $i \in \mathcal{I}$; the uniform distribution is the default choice. The corresponding rewards are then observed. From these samples, low-rank tensor completion yields the initial low-rank estimator $\widehat{\mathcal{T}}_{\mathrm{lr}}^{(0)}$ and the mode-wise factor-space estimates $\widehat{U}_1^{(0)}, \ldots, \widehat{U}_d^{(0)}$ in (1). The warm-up length s_0 is chosen so that the initializer enters the local stability neighborhood; see Supplementary Section S.3.1. RITE then follows the spectral rotation used in tensor elimination (Zhou et al., 2024) and fixes the rotated feature

$$a_i = \mathrm{vec}\left([\widehat{U}_1^{(0)}, \widehat{U}_{1,\perp}^{(0)}]^\top e_{i_1} \circ \cdots \circ [\widehat{U}_d^{(0)}, \widehat{U}_{d,\perp}^{(0)}]^\top e_{i_d}\right). \tag{5}$$

for each $i = (i_1, \ldots, i_d)$. The mean reward can be written exactly as $\langle \mathcal{T}^\star, \mathcal{A}(i)\rangle = a_i^\top \beta^\star$, where $\beta^\star$ is the vectorization of $\mathcal{T}^\star$ in these fixed rotated coordinates. The rotation isolates the coordinates orthogonal to every estimated factor space. The number of principal vectorized coordinates is $q = \prod_{j=1}^d p_j - \prod_{j=1}^d (p_j - r_j)$. Let P_{main} project onto this block and initialize the UCB design at each batch by $\Lambda_{\mathrm{ucb}} = \lambda_1 P_{\mathrm{main}} + \lambda_2 (I - P_{\mathrm{main}})$.

Batch k is also elimination phase k and has proposal budget $m_k = 2^{k-1}$. After c executed proposals, define

$$V_{k,c}^{\mathrm{ucb}} = \Lambda_{\mathrm{ucb}} + \sum_{s \in \mathcal{D}_k(c)} a_{I_s} a_{I_s}^\top, u_{k,c}(i) \quad = \{a_i^\top (V_{k,c}^{\mathrm{ucb}})^{-1} a_i\}^{1/2}. \tag{6}$$

Here $\mathcal{D}_k(c)$ contains the first c proposal pulls in batch k; target draws are excluded. The next proposal is $I_t^{\mathrm{TE}} \in \arg\max_{i \in \mathcal{C}_k} u_{k,c}(i)$. After m_k executed proposals, compute $\widehat{b}_k = (V_{k,m_k}^{\mathrm{ucb}})^{-1} \sum_{s \in \mathcal{D}_k} a_{I_s} Y_s$ and estimate the mean reward of action i by $\widehat{\mu}_k(i) = a_i^\top \widehat{b}_k$.

Starting from $\mathcal{C}_1 = \mathcal{I}$, update

$$\mathcal{C}_{k+1} = \left\{ i \in \mathcal{C}_k : \widehat{\mu}_k(i) + \beta_{\mathrm{ucb}} u_{k,m_k}(i) \geq \max_{j \in \mathcal{C}_k} [\widehat{\mu}_k(j) - \beta_{\mathrm{ucb}} u_{k,m_k}(j)] \right\}. \tag{7}$$

Here β_{ucb} is the confidence radius. Thus $\mathcal{C}_{k+1} \subseteq \mathcal{C}_k$. A target draw enters estimation and inference but does not change $\mathcal{D}_k$, $V_{k,c}^{\mathrm{ucb}}$, or c.

3.2 INFERENCE-PRESERVING EXPLORATION

Let $q_H = (1 - \eta) q_H^{\mathrm{tar}} + \eta q_{\mathrm{unif}}$, where $0 < \eta \leq 1$ and q_{unif} is uniform on $\mathcal{I}$; thus η controls the uniform support of the target-reference design. For a single-entry target, $q_H^{\mathrm{tar}} = \delta_{i_0}$; for an entry-gap target, $q_H^{\mathrm{tar}} = (\delta_{i_0} + \delta_{j_0})/2$. The target H and mixture fraction η are fixed by the decision maker before the online process begins. For $t \in \mathcal{E}_k$, the realized sampling distribution is

$$I_t \sim (1 - \varepsilon_t)\delta_{I_t^{\mathrm{TE}}} + \varepsilon_t q_H, \quad \pi_t(i) \quad = (1 - \varepsilon_t)\mathbf{1}\{i = I_t^{\mathrm{TE}}\} + \varepsilon_t q_H(i). \tag{8}$$

Target-specific observations enter the low-rank update and one-step inference, but do not update $V_{k,c}^{\mathrm{ucb}}$, $\mathcal{D}_k(c)$, or $\mathcal{C}_k$. Since $\pi_t(i) \geq \varepsilon_t q_H(i)$, the target-specific inverse-propensity weight satisfies

$$\rho_t^H = \frac{q_H(I_t)}{\pi_t(I_t)} \leq \varepsilon_t^{-1}. \tag{9}$$

This weight transports information under the realized policy π_t to the target-reference distribution q_H. For the batch-start tangent coordinates defined below, $\mathbb{E}\Big[\rho_t^H s_{I_t}^{(k)} s_{I_t}^{(k)\top} \mid \mathcal{F}_{t-1}\Big] = \sum_{i\in\mathcal{I}} \pi_t(i) \frac{q_H(i)}{\pi_t(i)} s_i^{(k)} s_i^{(k)\top} = \sum_{i\in\mathcal{I}} q_H(i) s_i^{(k)} s_i^{(k)\top} = J_{H,k}$. We take $\varepsilon_t = \min\{1/2, c_\varepsilon (t \vee s_0)^{-1/2}\}$. Hence $\sum_{t=s_0+1}^{T} \varepsilon_t \leq 2c_\varepsilon\sqrt{T}$. The expected number of target draws is therefore $O(\sqrt{T})$: RITE continues to accumulate information in the target direction while keeping the additional exploration regret at order $\sqrt{T}$. The resulting regret term is treated separately in Section 4.3.

3.3 Riemannian Capped-IPW Updating and Debiasing

The low-rank estimator uses the capped-IPW weight $\omega_t^{\mathrm{lr}} = \{\pi_t(I_t) \vee \pi_{\min}\}^{-1}$. The cap $\pi_{\min}$ bounds each nuisance-update weight by $\pi_{\min}^{-1}$ and controls fluctuations in the Riemannian gradient; it is not applied to the inferential weight ρ_t^H. If the conditional-mean model is correct and the weighted design has sufficient curvature, $\mathcal{T}^\star$ remains a minimizer of the corresponding population risk. The subsequent error bound is therefore expressed in terms of the effective information under the realized capped design.

At the end of batch k, the algorithm minimizes the cumulative loss on the Tucker manifold $\mathcal{M}_{\mathbf{r}}$, $L_k(\mathcal{T}) = \frac{1}{2}\sum_{t\leq t_{k+1}-1} \omega_t^{\mathrm{lr}} \{Y_t - \langle \mathcal{T}, \mathcal{A}_t\rangle\}^2, \qquad \mathcal{T} \in \mathcal{M}_{\mathbf{r}}$. Starting from $\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)}$, RITE performs J_k Riemannian gradient–retraction steps on $\mathcal{M}_{\mathbf{r}}$ using the capped-IPW loss above. The resulting finite-step optimization error, denoted by $\mathrm{Rem}_{\mathrm{opt}}(k)$, is controlled in Supplementary Section S.5.4.

We next construct a target-specific one-step correction. To avoid using the current batch rewards twice to determine the inferential direction, we fix the local geometry at the beginning of batch k. Let $\widehat{\mathbb{T}}^{(k-1)}$ be the tangent space of $\mathcal{M}_{\mathbf{r}}$ at $\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)}$, and project the actions and target onto this space:

$$s_i^{(k)} = \mathrm{vec}_{k-1}\big\{P_{\widehat{\mathbb{T}}^{(k-1)}}(\mathcal{A}(i))\big\}, \quad h_{k-1}(H) = \mathrm{vec}_{k-1}\big\{P_{\widehat{\mathbb{T}}^{(k-1)}}(H)\big\}. \tag{10}$$

Here vec_{k-1} gives coordinates in an orthonormal basis of the batch-start tangent space. The projection reduces full-dimensional tensor inference to a first-order problem in the local tangent space while retaining the identifiable principal perturbation directions of the low-rank model.

The information matrix induced by q_H in these local coordinates is $J_{H,k} = \sum_{i\in\mathcal{I}} q_H(i) s_i^{(k)} s_i^{(k)\top}, \quad g_{H,k} = J_{H,k}^{\dagger} h_{k-1}(H)$. Under the target-identifiability condition $h_{k-1}(H) \in \mathrm{Range}(J_{H,k})$, we have $J_{H,k} g_{H,k} = h_{k-1}(H)$, so $g_{H,k}$ provides the first-order correction direction for H. Both $J_{H,k}$ and $g_{H,k}$ are fixed before the rewards in batch k arrive.

For $t = t_k, \ldots, t_{k+1} - 1$, define the batch-start residual

$$\widehat{r}_t^{(k)} = Y_t - \langle \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)}, \mathcal{A}_t\rangle. \tag{11}$$

The batch-wise one-step estimator is $\widehat{\theta}_k(H) = \langle \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)}, H\rangle + \frac{1}{n_k}\sum_{t=t_k}^{t_{k+1}-1} \rho_t^H \{g_{H,k}^\top s_{I_t}^{(k)}\} \widehat{r}_t^{(k)}$. The low-rank plug-in term provides the local reference value, and the second term gives a first-order correction along the target tangent direction. The weight ρ_t^H transports the realized policy to the reference design.

Inference begins at a prespecified stable batch k_0. If $K_T \geq k_0$, let $N_T^{\mathrm{inf}} = T - t_{k_0} + 1, \alpha_k = \frac{n_k}{N_T^{\mathrm{inf}}}, k = k_0, \ldots, K_T$, and define

$$\widehat{\theta}_{\mathrm{db}}(H) = \sum_{k=k_0}^{K_T} \alpha_k \widehat{\theta}_k(H). \tag{12}$$

Thus $\alpha_k/n_k = (N_T^{\inf})^{-1}$ is fixed when batch k_0 begins. The last truncated batch also enters the aggregate. Its boundary depends only on the branch variables and does not use unobserved rewards from that batch. Section 4.2 therefore constructs a stopped martingale linear expansion. If $K_T < k_0$, RITE does not report an inferential result.

4 Theoretical Analysis

4.1 Low-Rank Estimation

We first state a nonasymptotic estimation bound and then give the inferential result for the target-specific one-step estimator. Let $p = \max_{j\in[d]} p_j$, $r = \max_{j\in[d]} r_j$, and let $D_{\text{eff}} = \prod_{j=1}^d r_j + \sum_{j=1}^d (p_j - r_j)r_j$ be the effective degrees of freedom. Define $\mathfrak{d}_T = \pi_{\min}^{-1} D_{\text{eff}} \log(pT)$. When the rank is fixed and the mode dimensions are comparable, D_{eff} is much smaller than the ambient dimension $p_1 \cdots p_d$. This reduction is the source of the estimation and inferential gains from low-rank tensor modeling (Kolda & Bader, 2009; Xia et al., 2021).

Assumption 1 (Low-Rank Estimation Regularity)**.** *The true reward tensor $\mathcal{T}^\star$ has Tucker rank $\mathbf{r} = (r_1, \ldots, r_d)$ with incoherent and well-conditioned factor matrices. Conditional on $\mathcal{F}_{t-1/2} = \sigma(\mathcal{F}_{t-1}, B_t, I_t)$, the noise ξ_t is mean-zero and σ-sub-Gaussian; for inference, its conditional variance is uniformly bounded away from zero and infinity and its conditional fourth moment is uniformly bounded. The policy is predictable, the sampling branch is chosen before Y_t, and batch endpoints use no future rewards. The warm-up and subsequent Riemannian iterates remain in the local regularity region where the weighted population risk has quadratic growth; finite-step optimization error is denoted by $\mathrm{Rem}_{\text{opt}}(k)$. Precise regularity and initialization conditions are given in Supplementary Section S.3.1.*

Assumption 2 (Policy-Induced Effective Information)**.** *The capped-IPW design satisfies the local restricted-curvature and upper-envelope conditions on rank-$2\mathbf{r}$ secant directions. Define $N_k^{\text{eff}} = \inf_{\substack{\Delta\in\mathbb{T}^\star \\ \|\Delta\|_F=1}} \sum_{t<t_{k+1}} \omega_t^{\text{lr}} \langle \Delta, \mathcal{A}_t\rangle^2$, and assume $N_k^{\text{eff}} \gtrsim \mathfrak{d}_T$ for every started batch. Predictable sufficient conditions and their transfer to the realized capped design are given in Supplementary Section S.3.2 and Lemma S.5.2.*

Theorem 1 (Low-Rank Estimation Error)**.** *Under Assumptions 1–2, there exist constants $c_1, c_2, C > 0$ such that, with probability at least $1 - c_1(pT)^{-c_2}$, simultaneously for every started batch $k = 1, \ldots, K_T$,*

$$\left\|\widehat{\mathcal{T}}_{\text{lr}}^{(k)} - \mathcal{T}^\star\right\|_F \le C\sigma \left\{ \sqrt{\frac{\mathfrak{d}_T}{N_k^{\text{eff}}}} + \frac{\mathfrak{d}_T}{N_k^{\text{eff}}} \right\} + C\mathrm{Rem}_{\text{opt}}(k). \tag{13}$$

Here N_k^{eff} measures the effective information in the capped-IPW design on the true tangent space through the end of batch k. The term $\mathrm{Rem}_{\text{opt}}(k)$ is the optimization error from finitely many Riemannian iterations. If it is dominated by the leading statistical term in (13)*, then* $\left\|\widehat{\mathcal{T}}_{\text{lr}}^{(k)} - \mathcal{T}^\star\right\|_F = O_p\left[\sigma\left\{\sqrt{\mathfrak{d}_T/N_k^{\text{eff}}} + \mathfrak{d}_T/N_k^{\text{eff}}\right\}\right]$.

Theorem 1 is design-adaptive: the rate is governed by the low-rank complexity D_{eff} and the effective information N_k^{eff} induced by adaptive sampling (Xia et al., 2021; Wen et al., 2026).

4.2 Doubly Debiased Inference

For a prespecified target H, let k_0 denote the burn-in batch at which inference begins. Assume that the low-rank estimator enters a locally stable region from batch k_0 onward. Define $N_T^{\inf} = \sum_{k=k_0}^{K_T} n_k, \qquad \alpha_k = \frac{n_k}{N_T^{\inf}}$. RITE estimates $\theta^\star(H) = \langle \mathcal{T}^\star, H\rangle$ using (12). The inverse-propensity ratio $\rho_t^H = q_H(I_t)/\pi_t(I_t)$ corrects for adaptive sampling. The one-step correction on the batch-start tangent space corrects the first-order bias of the low-rank plug-in estimator. Together, these two corrections yield the doubly debiased construction.

The direction $g_{H,k}$ is determined by the tangent space at the start of batch k and the fixed reference distribution q_H. Since $\mathbb{E}\Big[\rho_t^H s_{I_t}^{(k)} s_{I_t}^{(k)\top} \mid \mathcal{F}_{t-1}\Big] = J_{H,k}$ and $J_{H,k} g_{H,k} = h_{k-1}(H)$, the weighted score is conditionally aligned with the target tangent direction.

The corresponding empirical self-normalized variance is

$$\widehat{V}_T(H) = \sum_{k=k_0}^{K_T} \sum_{t \in \mathcal{E}_k} \left[\frac{1}{N_T^{\inf}} \rho_t^H \{ g_{H,k}^\top s_{I_t}^{(k)} \} \widehat{r}_t^{(k)} \right]^2 . \tag{14}$$

This variance estimator accumulates the batch-wise scores used by the algorithm and does not require a global homoscedastic noise parameter.

Assumption 3 (Target-Reference Information)**.** *For every $k = k_0, \ldots, K_T$,*

$$h_{k-1}(H) \in \mathrm{Range}(J_{H,k}), \qquad \left\| J_{H,k}^\dagger h_{k-1}(H) \right\| \le C_{g,H},$$
$$0 < c_{H,T} \le h_{k-1}(H)^\top J_{H,k}^\dagger h_{k-1}(H) \le C_{v,H}.$$

Here $c_{H,T}$ may depend on the dimension. These conditions ensure target identifiability and a non-degenerate, bounded reference variance; sufficient conditions for the single-entry and entry-gap designs are given in Supplementary Section S.4.1.

Under Assumptions 1–3, the one-step estimator admits the expansion $\widehat{\theta}_{\mathrm{db}}(H) - \theta^\star(H) = \sum_{k=k_0}^{K_T} \sum_{t \in \mathcal{E}_k} \psi_{k,t}(H) + \mathrm{Rem}_{\mathrm{db}}(H)$, where $\psi_{k,t}(H) := \rho_t^H \{ g_{H,k}^\top s_{I_t}^{(k)} \} \xi_t / N_T^{\inf}$ is a martingale difference. Supplementary Section S.6.2 gives a complete decomposition of the remainder.

Assumption 4 (Inferential Regularity)**.** *Let $V_T(H) := \sum_{k=k_0}^{K_T} \sum_{t \in \mathcal{E}_k} \mathbb{E}[\psi_{k,t}(H)^2 \mid \mathcal{F}_{t-1}]$. For some deterministic $v_T(H) > 0$, assume*

$$\frac{V_T(H)}{v_T(H)} \to_p 1, \qquad \frac{\mathrm{Rem}_{\mathrm{db}}(H)}{v_T(H)^{1/2}} \to_p 0, \qquad \frac{\widehat{V}_T(H)}{V_T(H)} \to_p 1,$$

and that the leading martingale-difference array satisfies the conditional Lindeberg condition. Verifiable sufficient conditions are given in Supplementary Sections S.6.3–S.6.5.

Theorem 2 (Asymptotic Normality of Doubly Debiased Inference)**.** *Under Assumptions 1–3 and Assumption 4,*

$$\frac{\widehat{\theta}_{\mathrm{db}}(H) - \theta^\star(H)}{\widehat{V}_T(H)^{1/2}} \Rightarrow N(0,1). \tag{15}$$

Consequently, an asymptotic $1-\alpha$ confidence interval is

$$\mathrm{CI}_{1-\alpha}(H) = \left[\widehat{\theta}_{\mathrm{db}}(H) \pm z_{1-\alpha/2} \widehat{V}_T(H)^{1/2} \right], \tag{16}$$

and $\Pr\{\theta^\star(H) \in \mathrm{CI}_{1-\alpha}(H)\} \longrightarrow 1-\alpha$.

When $\varepsilon_t \asymp t^{-1/2}$ and target actions are visited mainly through target-specific exploration, the effective sample size in the target direction is typically of order $\sqrt{T}$. The standard error is therefore of order $T^{-1/4}$. This rate captures the trade-off between low regret and inferential precision, unlike root-T inference based on T independent target observations.

4.3 Regret Analysis

Target-specific exploration rounds do not update the tensor-elimination design matrix, candidate set, or proposal count. After removing these rounds, the remaining decision subsequence is identical to a tensor-elimination process run for M_T rounds. Let $N_T^{\mathrm{tar}} = \sum_{t=s_0+1}^{T} B_t$, $M_T = T - s_0 - N_T^{\mathrm{tar}}$, $S_T^{\varepsilon} = \sum_{t=s_0+1}^{T} \varepsilon_t$, where N_T^{tar} is the realized number of target-specific exploration rounds. The quantity M_T is the number of tensor-elimination proposals executed after warm-up. Also define $\Delta_{\max} = \max_{i \in \mathcal{I}} \langle \mathcal{T}^\star, \mathcal{A}^\star - \mathcal{A}(i) \rangle$. The cumulative regret of RITE then decomposes as

$$R_T^{\mathrm{RITE}} \le s_0 \Delta_{\max} + R_{M_T}^{\mathrm{TE}} + \Delta_{\max} N_T^{\mathrm{tar}}, \tag{17}$$

where the three terms arise from warm-up, the tensor-elimination decision subsequence, and target-specific exploration, respectively. This decomposition shows that the additional regret introduced by the inference module depends only on the number of target-exploration rounds.

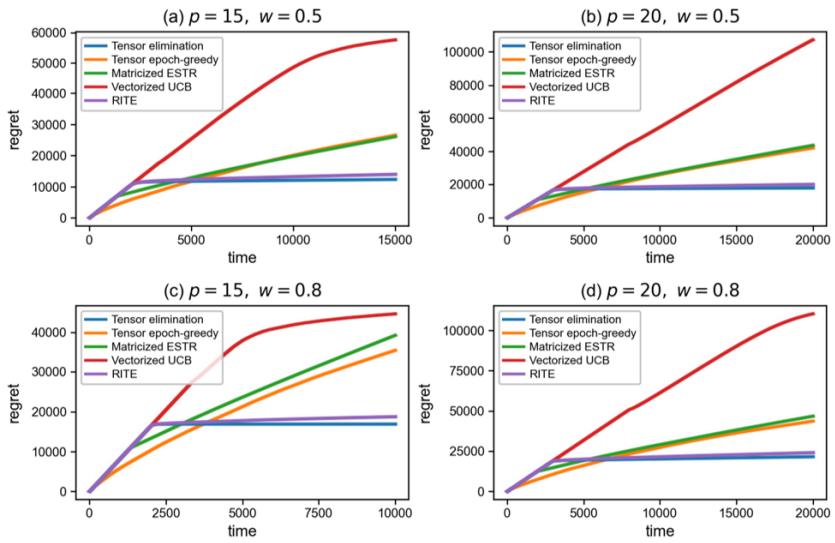

Figure 2: Mean cumulative regret under varying dimensions and signal strengths over 30 replications.

Theorem 3 (Regret Bound for RITE)**.** *Under conditionally sub-Gaussian noise, bounded reward gaps, and the warm-up and spectral-rotation conditions required by tensor elimination, the following holds for every* $\delta \in (0,1)$*. With probability at least* $1-\delta$*,*

$$R_T^{\mathrm{RITE}} \le s_0 \Delta_{\max} + \widetilde{O}_\delta\left(p^{(d-1)/2}\sqrt{M_T}\right) \quad + \Delta_{\max}\left[S_T^{\varepsilon} + 2\sqrt{S_T^{\varepsilon}\log(2/\delta)} + \frac{2}{3}\log(2/\delta)\right].$$

When the Tucker ranks are fixed and the mode dimensions are of the same order, i.e., $p_1 \asymp \cdots \asymp p_d \asymp p$, and the initialization requirement $s_0 = \widetilde{O}(p^{d/2})$, the regret bound in Theorem 3 can reduce to $R_T^{\mathrm{RITE}} = \widetilde{O}_\delta\left(p^{d/2} + p^{(d-1)/2}\sqrt{T} + \sqrt{T}\right)$. The first term comes from initialization, the second term is the tensor-elimination regret, and the additional $\widetilde{O}(\sqrt{T})$ term enables target-specific inference. This matches the typical regret order of tensor elimination (Zhou et al., 2024); the only additional contribution is the $\widetilde{O}(\sqrt{T})$ target-specific exploration cost required for inference. We know that the target-specific exploration supplies inferential information without changing the leading regret order of tensor elimination. Supplementary Section S.7 gives the required regularization parameters and confidence radii, and Supplementary Section S.7.3 proves the proposal-subsequence coupling.

5 Simulations

We conduct several preliminary experiments to examine two roles of RITE: regret control in online decision-making and uncertainty quantification for fixed targets. We first compare RITE with tensor elimination and tensor epoch-greedy (Zhou et al., 2024), matricized ESTR based on a mode unfolding (Jun et al., 2019), and vectorized UCB (Auer et al., 2002).

The regret experiment uses a synthetic third-order low-rank tensor of size $p \times p \times p$ as the true reward tensor, with Tucker rank $(2,2,2)$. The mode-wise factors are generated from random orthogonal matrices, and the signal strength is controlled by w. At each round, the observation satisfies $Y_t = \mathcal{T}^{\star}_{I_t} + \xi_t$, where ξ_t is Gaussian noise. Further details on data generation, horizons, and the pairing protocol are given in Supplementary Section S.1.1. The regret and inference experiments use fixed configurations specified in Supplementary Section S.1.3; their trade-off is reported in Section S.1.6.

Figure 2 reports cumulative regret for $(p,w) \in \{15,20\} \times \{0.5,0.8\}$ over 30 independent replications. Tensor elimination has the lowest final mean cumulative regret in all settings; RITE remains close and outperforms the other baselines over longer horizons. The paired design attributes their

modest gap to target exploration, consistent with Theorem 3: fixed-target information does not change tensor elimination's leading regret order.

We next examine the empirical coverage of nominal 95% confidence intervals. For $p \in \{10, 20, 30\}$, we consider the single-entry target $H_{\text{ent}} = \mathcal{A}(i_0)$ and the entry-gap target $H_{\text{gap}} = \mathcal{A}(i_0) - \mathcal{A}(j_0)$. Both targets are fixed before the online procedure starts. The true tensor is generated under the corresponding exact null constraints. Complete constructions of the targets and tensors are given in Supplementary Section S.1.1.

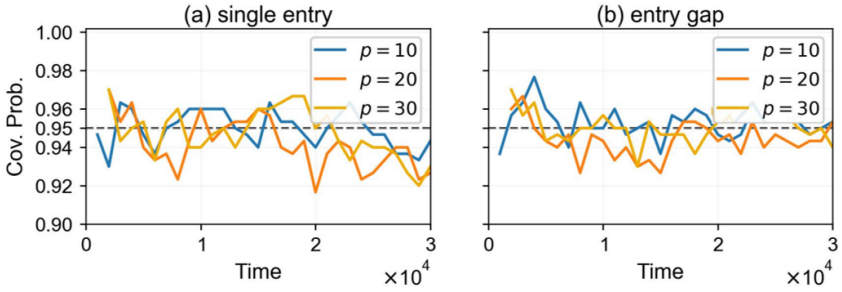

Figure 3: Empirical coverage of nominal 95% confidence intervals. Left: single-entry target. Right: entry-gap target.

Figure 3 reports empirical coverage over 300 independent online experiments. For both targets and all three dimensions, coverage fluctuates around the nominal 95% level, consistent with Theorem 2. Additional diagnostics and ablations are reported in Supplementary Sections S.1.2 and S.1.4–S.1.6.

6 Applications to Online Advertising

To examine RITE under a real multi-way reward structure, we use the non-contextual online advertising data of Zhou et al. (2024). Click counts form a $20 \times 10 \times 7$ tensor indexed by advertisement, publisher, and day of week. Each action selects one combination, and the objective is to maximize expected clicks. Supplementary Section S.2.1 gives the tensor-construction details.

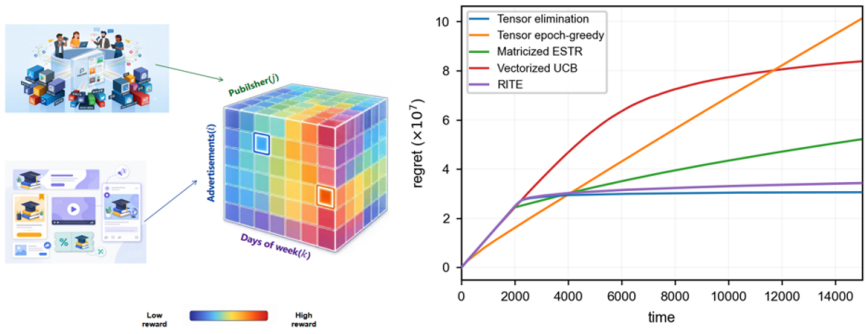

Figure 4: Online advertising benchmark. Left: three-way advertisement–publisher–day-of-week reward tensor. Right: mean cumulative regret over 30 runs.

All methods share the empirical tensor as the fixed mean reward and receive noisy feedback on its scale; Supplementary Section S.2.2 gives the protocol and tuning. Figure 4 shows RITE close to tensor elimination in final mean regret, with both outperforming the other baselines over longer horizons. Thus, RITE preserves target information at little empirical regret cost.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for manuscript translation, language polishing, and code debugging. We have not used generative AI tools for the development of the methodological framework or theoretical results, manuscript drafting, or the generation of figures and tables. Additionally, we used generative AI tools for formatting and style checks. We have reviewed and verified all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

Joshua Agterberg and Anru R. Zhang. Statistical inference for low-rank tensors: Heteroskedasticity, subgaussianity, and applications, 2024.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Xuan Bi, Annie Qu, and Xiaotong Shen. Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics*, 46(6B):3308–3333, 2018.

Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision-making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019. doi: 10.1073/pnas.1910053116.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214. PMLR, 2011.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

Evgeny Frolov and Ivan Oseledets. Tensor methods and recommender systems. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pp. 359–360, 2017.

Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, 1980.

Qiyu Han, Will Wei Sun, and Yichen Zhang. Online statistical inference for matrix contextual bandit, 2025.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pp. 3163–3172, 2019.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51 (3):455–500, 2009.

Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014. doi: 10.1007/s10543-013-0455-z.

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

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 460–468. PMLR, 2021.

Wanteng Ma and Dong Xia. Statistical inference in tensor completion: Optimal uncertainty quantification and statistical-to-computational gaps, 2024.

Lili Shan, Lei Lin, Chengjie Sun, and Xiaolong Wang. Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization. *Electronic Commerce Research and Applications*, 16:30–42, 2016. doi: 10.1016/j.elerap.2016.01.004.

Chengshuai Shi, Cong Shen, and Nicholas D. Sidiropoulos. On high-dimensional and low-rank tensor bandits. In *2023 IEEE International Symposium on Information Theory*, pp. 2279–2284, 2023. doi: 10.1109/ISIT54713.2023.10206552.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. doi: 10.1214/ECP.v16-1624.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12(1): 99–111, 1972. doi: 10.1007/BF01932678.

Xin Wen, Will Wei Sun, and Yichen Zhang. Online tensor inference, 2026.

Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *The Annals of Statistics*, 49(1):76–99, 2021.

Dong Xia, Anru R. Zhang, and Yuchen Zhou. Inference for low-rank tensors–no need to debias. *The Annals of Statistics*, 50(2):1220–1245, 2022. doi: 10.1214/21-AOS2146.

Jie Zhou, Botao Hao, Zheng Wen, Jingfei Zhang, and Will Wei Sun. Stochastic low-rank tensor bandits for multi-dimensional online decision making. *Journal of the American Statistical Association*, 2024. doi: 10.1080/01621459.2024.2311364.

SUPPLEMENTARY MATERIAL

This supplementary material first reports additional experiments, including synthetic diagnostics and the advertising benchmark. It then gives the notation and the full Rotated Inferential Tensor Elimination (RITE) procedure. The remaining sections prove the three main theoretical statements, establish supplementary factor-space and inferential-information results, state the general tools used in the analysis, and collect detailed proofs of the RITE-specific lemmas.

S.1 ADDITIONAL SIMULATION STUDIES

The main paper reports the primary regret and coverage results. Here we give the full data-generating mechanisms, tuning protocol, and additional diagnostics. The experiments address several questions. We ask whether RITE preserves the regret behavior of rotated elimination, whether the studentized estimator is calibrated for fixed targets, how the low-rank estimate behaves as dimension, rank, and noise change, whether the main algorithmic components are necessary, how target-specific exploration mediates the trade-off between regret and inferential precision, and whether dynamic batching and inference incur manageable computational overhead.

S.1.1 SYNTHETIC DESIGNS

The primary synthetic experiment uses third-order tensors of size $p \times p \times p$ and CP rank $r = 2$, following the construction of Zhou et al. (2024). For each mode $j \in [3]$, we draw a Gaussian matrix and retain the first r columns of its QR factor. Denoting these columns by $u_{j,\ell}$, we set

$$\mathcal{T}^\star = \sum_{\ell=1}^{r} w p^{3/2} u_{1,\ell} \circ u_{2,\ell} \circ u_{3,\ell}. \tag{18}$$

Table S1: Synthetic experiments and reported Monte Carlo sizes.

Experiment	Settings	Replications
Regret	$(p, w, T) = (15, 0.5, 15000)$, $(20, 0.5, 20000)$, $(15, 0.8, 10000)$, $(20, 0.8, 20000)$	30
Coverage	$p \in \{10, 20, 30\}$, $r = 2$, $w = 0.8$, $\sigma = 0.1$, $T = 30000$	300
Rank normality	$p = 15$, $r \in \{1, 2, 3\}$, $w = 0.8$, $\sigma = 0.1$, $T = 30000$	1000
Noise normality	$p = 15$, $r = 2$, $w = 0.8$, $\sigma \in \{1, 2, 3\}$, $T = 30000$	300
Estimation error	$p \in \{10, 20, 30\}$, $r \in \{1, 2, 3\}$, $\sigma \in \{1, 2, 3\}$, $T = 30000$	20

The factor $p^{3/2}$ keeps the entrywise reward scale comparable as p changes. At round t, the selected one-hot tensor $\mathcal{A}_t$ produces

$$Y_t = \langle \mathcal{T}^\star, \mathcal{A}_t \rangle + \xi_t, \qquad \xi_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2). \tag{19}$$

The regret experiments use $\sigma = 1$ and the four (p, w, T) settings listed in Table S1. A common reward tensor is used across methods within each setting and replication, while policy randomization is method-specific.

The RITE target is fixed independently of the generated rewards. With one-based indices $i_0 = (1, 1, 1)$ and $j_0 = (2, 2, 2)$, we use $H = \mathcal{A}(i_0) - \mathcal{A}(j_0)$ in every trajectory. The four comparison methods use local implementations following the algorithmic designs of Zhou et al. (2024). In the main regret comparison, tensor elimination is implemented as a matched decision-only control of RITE: it uses the same warm-up rotation, proposal-count batches, phase-wise ridge resets, active-set screening, and paired random streams, with target-specific exploration switched off by setting $c_\varepsilon = 0$.

Coverage is evaluated under an exact null. Let $i_0 = (1, 1, 1)$ and $j_0 = (2, 2, 2)$, with one-based indices. We consider

$$H_1 = \mathcal{A}(i_0), \qquad H_2 = \mathcal{A}(i_0) - \mathcal{A}(j_0). \tag{20}$$

For each replication, orthonormal Tucker factors are generated first. The random core is then projected onto the joint null space induced by $\langle \mathcal{T}^\star, H_1 \rangle = \langle \mathcal{T}^\star, H_2 \rangle = 0$. After projection, we verify $\operatorname{rank}\{M_j(\mathcal{G}^\star)\} = r_j$ for every mode and regenerate the core otherwise. The two coverage calculations therefore use exact nulls with the prescribed Tucker rank.

S.1.2 Online Low-Rank Estimation Error

We next isolate low-rank estimation from action selection. Following the diagnostic design of Wen et al. (2026), observations in this experiment use independent Gaussian tensor covariates rather than one-hot bandit actions. This isolates the effect of structural dimension and noise. For each checkpoint, we report

$$\mathrm{RelErr}(t) = \|\widehat{\mathcal{T}}^{(t)} - \mathcal{T}^\star\|_F / \|\mathcal{T}^\star\|_F. \tag{21}$$

The true tensor is rescaled so that $\|\mathcal{T}^\star\|_F = 12\sqrt{r}$ in the dimension and rank panels. For the noise panel, we use $\|\mathcal{T}^\star\|_F = 5.4\sqrt{r}$ to make the deterioration with increasing noise visible.

Figure S5 therefore provides a controlled diagnostic of how the low-rank estimation error changes with dimension, rank, and noise, independently of the adaptive action-selection mechanism studied in the bandit experiments below.

S.1.3 Hyperparameter Selection

We use a two-stage tuning protocol. A small collection of pilot seeds is used to choose one configuration for each method. These seeds are excluded from every reported Monte Carlo average. The

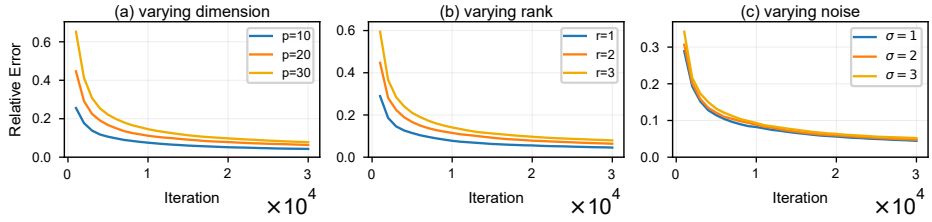

Figure S5: Full-information low-rank estimation error under varying tensor dimension, Tucker rank, and noise level. Bands show mean ± 1.96 Monte Carlo standard errors over 20 replications.

Table S2: Hyperparameters used in the synthetic regret experiment.

Method	Fixed configuration
Tensor elimination	$c_{\text{init}} = 24$, $\gamma = 0.75$, active-set multiplier 2.0, matched dynamic batching, $c_\varepsilon = 0$, and four Riemannian steps
RITE	$c_{\text{init}} = 24$, $\gamma = 0.75$, active-set multiplier 2.0, $c_\varepsilon = 2.0$, $\eta = 0.10$, $\pi_{\min} = 10^{-3}$, and four Riemannian steps
Tensor epoch-greedy	warm-up scale 1.0, exploration scale 1.0, exploitation scale 2.0, and block-style updates
Matricized ESTR	warm-up scale 16.0, active multiplier 80.0, UCB weight 0.45, and exploration rate 0.28
Vectorized UCB	UCB exploration coefficient $\beta = 2.0$

selected configuration is then frozen across all four (p, w) settings. This rule avoids panel-specific tuning and follows the sensitivity-based protocol used by Zhou et al. (2024).

For RITE, the warm-up length is

$$s_0 = \min\left\{\lceil c_{\text{init}} rp \log(p+1)\rceil, \lfloor T/4 \rfloor\right\}, \tag{22}$$

with a lower safeguard of 50 observations. The regret experiment uses $c_{\text{init}} = 24$, the rotated-width proposal, $\gamma = 0.75$, active-set multiplier 2.0, $c_\varepsilon = 2.0$, $\eta = 0.10$, $\pi_{\min} = 10^{-3}$, and four Riemannian steps. The exploration schedule is

$$\varepsilon_t = \min\left\{\frac{1}{2}, \frac{c_\varepsilon}{\sqrt{t+1}}\right\}, \qquad q_H = (1-\eta) q_H^{\text{tar}} + \eta q_{\text{unif}}, \tag{23}$$

where $\eta \in (0, 1]$ and q_H^{tar} is uniform over the entries appearing in H. Thus the added exploration follows the same $t^{-1/2}$ decay rate as assumed in the regret analysis.

Table S2 records the final regret configurations. The baselines use the local implementations distributed with our code, following the corresponding algorithmic designs in Zhou et al. (2024). Each method is fixed across all four panels and is not tuned on the reported replications.

The inference experiments use $c_{\text{init}} = 8$, $\gamma = 1.5$, active-set multiplier 3.0, $c_\varepsilon = 7.0$, $\eta = 0.01$, and $\pi_{\min} = 10^{-3}$. Their warm-up distribution is

$$q_0 = \frac{1}{2} q_H^{\text{tar}} + \frac{1}{2} q_{\text{unif}}.$$

Thus all warm-up draws are i.i.d. from a prespecified full-support distribution. Coverage uses the dimension-dependent rule in (22). The rank and noise normality diagnostics use the common length $s_0 = 1000$ so that every setting satisfies the same local-initialization check. All intervals use the uninflated self-normalized standard error.

Matricized ESTR uses a local ESTR-style implementation on the mode-1 unfolding, consisting of a rank-constrained matrix-completion warm-up followed by candidate-pool UCB. Vectorized UCB treats the p^3 entries as unrelated arms, while tensor epoch-greedy alternates random exploration with block-style exploitation. Every regret curve averages 30 independent replications.

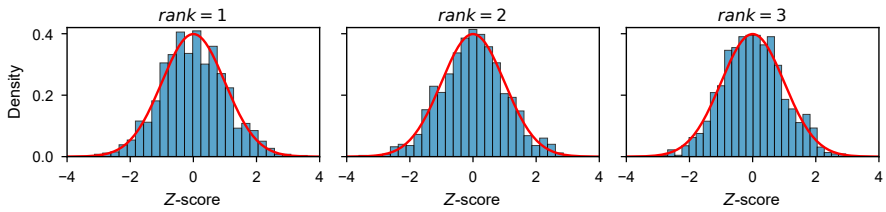

Figure S6: Studentized debiased estimates over 1000 replications for each Tucker rank. The red curve is the standard normal density.

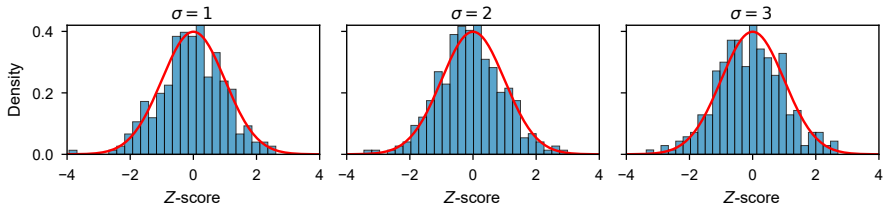

Figure S7: Studentized debiased estimates over 300 replications for each noise level. The red curve is the standard normal density.

The main-paper regret curves compare five methods under matched reward tensors and horizons. The four baselines use the local implementations described above. RITE uses the same fixed configuration in every panel; its target-specific draws advance calendar time but do not increase the internal tensor-elimination proposal count. The reported curves are unsmoothed averages over independent replications.

S.1.4 Coverage and Confidence Interval Length

At post-warm-up checkpoints in $\{1000, 2000, \ldots, 30000\}$, we compute the empirical coverage of

$$\widehat{\theta}_{\mathrm{db}}(H) \pm z_{1-\alpha/2}\widehat{V}_T(H)^{1/2},$$

with $\alpha = 0.05$. We require at least 250 post-warm-up observations, so the first eligible checkpoint is $t = 1000$ for $p = 10$ and $t = 2000$ for $p = 20, 30$. Each checkpoint reports pointwise coverage over independent online trajectories. The intervals are not simultaneous over time, and the plotted values are unsmoothed.

We also retain interval length, bias, and the studentized statistic for every replication and checkpoint. The plug-in target $\langle \widehat{\mathcal{T}}_{\mathrm{lr}}^{(K_T)}, H \rangle$ is recorded as a bias diagnostic. Formal coverage is computed from the batchwise aggregate in (48).

Figure S6 reports the empirical distribution of the studentized debiased statistic under three Tucker-rank settings. Each panel uses independent online trajectories, and the red curve is the standard normal density.

Figure S7 repeats the diagnostic for $\sigma = 1, 2, 3$. The comparison tests whether the studentized statistic remains close to the standard normal law as the signal-to-noise ratio decreases.

S.1.5 Component Ablation

We evaluate four variants at $p = 20$, Tucker rank $(2, 2, 2)$, $T = 30000$, $w = 0.8$, and $\sigma = 0.1$. The inferential target is the prespecified entry gap used in the coverage experiment. The 30 replications are paired across variants. The first ablation sets $c_\varepsilon = 0$ and removes post-warm-up target exploration. The second replaces $\omega_t^{\mathrm{lr}} = \{\pi_t(I_t) \vee \pi_{\min}\}^{-1}$ by unit weights. The third replaces each tangent-projected Riemannian update by a cumulative rank-truncated spectral update. All other parameters remain fixed.

Table S3: Component ablation over 30 paired replications. Reported values are replication means; coverage is the empirical coverage of the nominal 95% interval.

Variant	Final regret	Coverage	CI length	Rel. error
RITE	117181	0.967	0.0719	0.852
Without target exploration	104596	–	–	0.848
Without capped-IPW	114391	0.500	0.0644	1.670
Without Riemannian updating	111276	0.533	0.1150	2.385

Removing target exploration lowers mean regret but violates the target-support condition required for the reference-score correction. Consequently, target-specific coverage and interval length are not defined for this variant. Replacing capped-IPW or Riemannian updating leaves regret on the same scale but increases the relative tensor error and substantially deteriorates coverage.

S.1.6 Target-Exploration Trade-off

We vary the target-exploration multiplier over $c_\varepsilon \in \{0, 1, 2, 4, 7, 10\}$ in the first synthetic setting, while holding the remaining RITE parameters fixed. For each value, we report final regret, empirical coverage, mean 95% confidence-interval length, and the number of post-warm-up target draws. We use the first synthetic regret setting $(p, r, w, T, \sigma) = (15, 2, 0.5, 15000, 1)$. The vertical dashed line marks $c_\varepsilon = 2$, the value used in the main regret experiments; $c_\varepsilon = 7$, used in the main inference experiments, is included in the sensitivity grid.

Error bars in panels (a), (c), and (d) are 95% Monte Carlo confidence intervals, computed as the mean plus or minus 1.96 Monte Carlo standard errors. Panel (b) reports Wilson 95% binomial confidence intervals for empirical coverage. When $c_\varepsilon = 0$, no valid target-specific confidence interval is available, so coverage and confidence-interval length are omitted from panels (b) and (c).

S.1.7 Dynamic Batch Operation

Figure S9 records one representative RITE trajectory from the setting $p = 15$, $r = 2$, $w = 0.5$, and $T = 15000$. Panel (a) reports the active-set size, which changes only after a completed dynamic batch. Panel (b) separates the tensor-elimination proposal clock from the cumulative number of target-specific draws. Target draws advance calendar time but do not increase the proposal count. The vertical dashed lines indicate the realized batch boundaries.

S.1.8 Implementation Runtime and Memory

We compare the five implementations at $(p, r, w, T) = (15, 2, 0.5, 10000)$ using CPU-only isolated serial worker processes. Environment generation is excluded from the timed region. For RITE, elapsed time includes the final debiased estimate and confidence-interval calculation in addition to the online decision and low-rank updating steps. Peak host resident-set size includes the Python process and loaded libraries. Each entry reports the mean and standard deviation over ten independent runs.

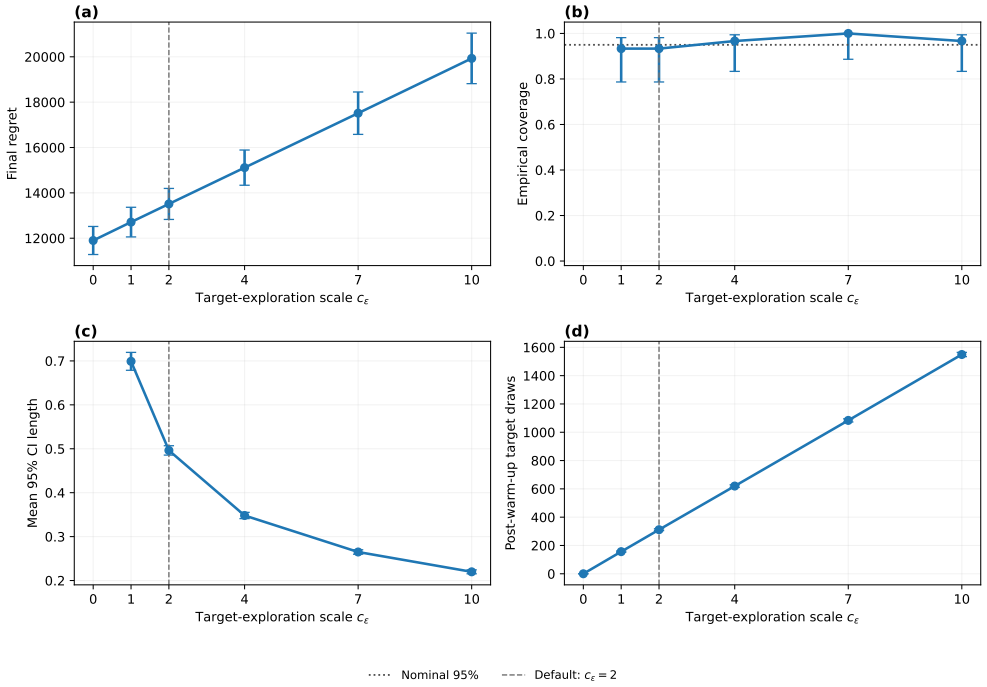

Figure S8: Decision–inference trade-off induced by the target-exploration scale c_ε. Panels (a)–(d) report final regret, empirical coverage, mean 95% confidence- interval length, and the number of post-warm-up target draws, respectively. The vertical dashed line marks $c_\varepsilon = 2$, the value used in the main regret experiments, and the dotted horizontal line in panel (b) denotes the nominal 95% coverage level. At $c_\varepsilon = 0$, no valid target-specific confidence interval is available; hence coverage and confidence-interval length are not reported in panels (b) and (c). Error bars in panels (a), (c), and (d) are mean ± 1.96 Monte Carlo standard errors over 30 replications, while panel (b) reports Wilson 95% binomial confidence intervals.

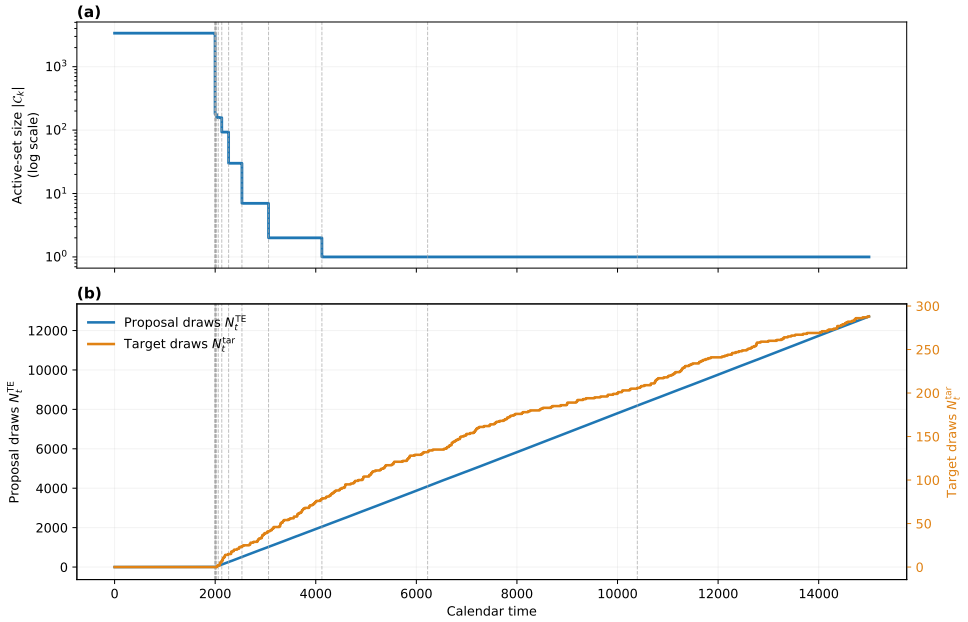

Figure S9: Dynamic batching and exploration behavior in one representative RITE trajectory. Panel (a) shows the active-set size on a logarithmic scale. Panel (b) shows the cumulative numbers of tensor-elimination proposals and target-specific draws against calendar time. Vertical dashed lines mark the dynamic batch boundaries.

Method	Time per 10^4 rounds (s)	Peak host RSS (MB)	Structural update	Inference
Tensor elimination	1.190 ± 0.301	580.4 ± 3.2	Warm-up spectral only	No
Tensor epoch-greedy	1.637 ± 0.038	568.3 ± 0.2	No	No
Matricized ESTR	0.324 ± 0.012	539.1 ± 1.5	Matrix	No
Vectorized UCB	0.134 ± 0.009	539.3 ± 0.5	No	No
RITE	1.852 ± 0.053	563.4 ± 2.4	Warm-up spectral + Riemannian	Yes (timed)

Figure S10: CPU runtime and peak host memory of RITE and the four competing methods. RITE incurs moderate additional computation from its Riemannian low-rank updates and inferential calculation, while remaining on the same computational scale as tensor elimination.

S.2 Additional Application Details

The advertising experiment uses the non-contextual benchmark tensor distributed with the supplementary code of Zhou et al. (2024). The supplied CSV contains one row for each publisher–day pair and 20 advertising-arm columns. We reshape it as

$$i = (\text{advertisement}, \text{publisher}, \text{day of week}) \in [20] \times [10] \times [7]. \tag{24}$$

The resulting tensor has 1400 actions. It is used as a fixed mean-reward environment. Because the experiment replays this tensor with simulated observation noise, we refer to it as a benchmark-based semi-synthetic application rather than a direct replay of logged propensities.

S.2.1 Tensor Construction

For each publisher and day-of-week pair, the 20 arm columns form the first mode of the tensor. No additional category aggregation or imputation is applied after loading the supplied CSV. The observed reward at round t is generated as

$$Y_t = \mathcal{T}^\star_{I_t} + \xi_t, \qquad \xi_t \sim N(0, 3000^2). \tag{25}$$

We use Tucker rank $r = 2$ and horizon $T = 15000$. The regret comparison uses 30 independent policy trajectories. The fixed tensor is shared across methods, while policy randomization is method-specific. The sequential confidence-interval diagnostic in Figure S11 is based on one representative RITE trajectory.

S.2.2 Replay Protocol

At each round, a method selects one of the 1400 entries and observes only its noisy reward. The four baselines use the fixed author configurations in Table S2. RITE uses 2000 warm-up rounds, $\gamma = 0.8$, active-set multiplier 3.0, and $c_\varepsilon = 2.0$ for the regret comparison. The separate target-interval diagnostic sets $c_\varepsilon = 7.0$, $\eta = 0.10$, $\pi_{\min} = 10^{-3}$, and uses four Riemannian steps per batch.

We fix two benchmark targets before every online trajectory. In one-based notation, let $i_0 = (1, 1, 1)$ and $j_0 = (2, 2, 2)$, and report

$$H_{\text{entry}} = \mathcal{A}(i_0), \qquad H_{\text{gap}} = \mathcal{A}(i_0) - \mathcal{A}(j_0). \tag{26}$$

The inferential distribution allocates its target component equally over the entries appearing in H_{gap}. Post-warm-up estimates are evaluated at checkpoints $t = 3000, 4000, \ldots, 15000$.

S.2.3 Reported Quantities

The main paper reports cumulative pseudo-regret relative to the largest entry of the fixed reward tensor. The supplement reports $\widehat{\theta}_{\text{db}}(H_{\text{entry}})$, $\widehat{\theta}_{\text{db}}(H_{\text{gap}})$, and their 95% confidence intervals. Capped weights are used only for the Riemannian nuisance update; the intervals use the target-reference ratio $q_H(I_t)/\pi_t(I_t)$ and the uninflated residual-variance estimator.

Figure S11 reports the two fixed-target estimates and their pointwise 95% confidence intervals along this representative trajectory. The solid reference line denotes the true value of H_{entry}, and the dashed reference line denotes the true value of H_{gap}.

S.3 Additional Notation and Technical Preliminaries

Let $p = \max_j p_j$, $\mathcal{I} = [p_1] \times \cdots \times [p_d]$, and $\mathcal{A}(i) = e_{i_1} \circ \cdots \circ e_{i_d}$ for $i = (i_1, \ldots, i_d) \in \mathcal{I}$. At time t, RITE selects I_t, sets $\mathcal{A}_t = \mathcal{A}(I_t)$, and observes

$$Y_t = \langle \mathcal{T}^\star, \mathcal{A}_t \rangle + \xi_t. \tag{27}$$

Let $\mathcal{F}_{t-1}$ contain all observations and algorithmic randomization revealed before round t. The proposal I_t^{TE} is $\mathcal{F}_{t-1}$-measurable. RITE draws $B_t \mid \mathcal{F}_{t-1} \sim \text{Bernoulli}(\varepsilon_t)$ and then chooses I_t. Define $\mathcal{F}_{t-1/2} = \sigma(\mathcal{F}_{t-1}, B_t, I_t)$. We assume

$$\mathbb{E}(\xi_t \mid \mathcal{F}_{t-1/2}) = 0, \qquad \mathbb{E}\{\exp(\lambda \xi_t) \mid \mathcal{F}_{t-1/2}\} \le \exp(\lambda^2 \sigma^2 / 2). \tag{28}$$

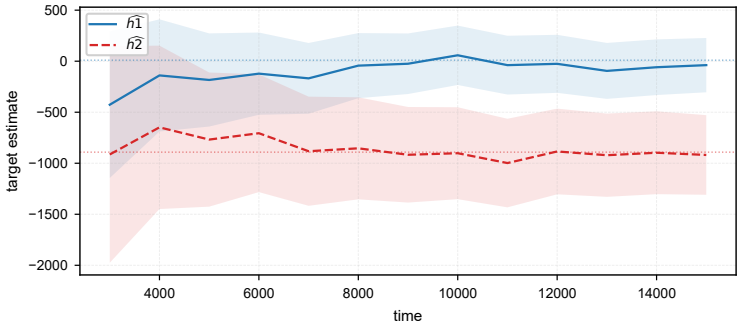

Figure S11: Estimated fixed entry and entry-gap advertising targets with 95% confidence intervals.

The action and the batch boundary are therefore determined before Y_t is observed.

Set $t_1 = s_0 + 1$. Batch k starts at t_k and ends when 2^{k-1} tensor-elimination proposals have been executed. Target draws advance calendar time but not this proposal count. If the horizon is reached first, set $t_{K_T+1} = T + 1$ and retain the last batch as truncated. Write

$$\mathcal{E}_k = \{t_k, \dots, t_{k+1} - 1\}, \qquad n_k = t_{k+1} - t_k, \qquad K_T = \max\{k : t_k \le T\}. \tag{29}$$

The t_k are stopping times for the pre-reward filtration, and $K_T \le 1 + \lfloor \log_2(T - s_0 + 1) \rfloor$.

For $t > s_0$, RITE uses

$$\varepsilon_t = \min\left\{\frac{1}{2}, c_\varepsilon (t \vee s_0)^{-1/2}\right\}, \qquad c_\varepsilon > 0, \tag{30}$$

and

$$\pi_t(i) = (1 - \varepsilon_t)\mathbf{1}\{i = I_t^{\mathrm{TE}}\} + \varepsilon_t q_H(i). \tag{31}$$

For batchwise bounds, define

$$\underline{\varepsilon}_k = \min_{t \in \mathcal{E}_k} \varepsilon_t = \varepsilon_{t_{k+1}-1}. \tag{32}$$

The cumulative low-rank loss uses

$$\omega_t^{\mathrm{lr}} = \{\pi_t(I_t) \vee \pi_{\min}\}^{-1}. \tag{33}$$

The one-step score uses

$$\rho_t^H = \frac{q_H(I_t)}{\pi_t(I_t)}. \tag{34}$$

Since $\pi_t(i) \ge \varepsilon_t q_H(i)$,

$$\rho_t^H \le \varepsilon_t^{-1} \le \underline{\varepsilon}_k^{-1}, \qquad t \in \mathcal{E}_k. \tag{35}$$

No cap is applied to ρ_t^H. During warm-up,

$$\pi_t(i) = q_0(i), \qquad \omega_t^{\mathrm{lr}} = \{q_0(I_t) \vee \pi_{\min}\}^{-1}, \qquad t \le s_0. \tag{36}$$

The Tucker model is

$$\mathcal{T}^\star = \mathcal{G}^\star \times_1 U_1^\star \times_2 \cdots \times_d U_d^\star.$$

Let

$$\mathcal{M}_{\mathbf{r}} = \{\mathcal{T} : \mathrm{rank}_{\mathrm{Tuck}}(\mathcal{T}) = \mathbf{r}\}, \qquad \mathbb{T}^\star = T_{\mathcal{T}^\star}\mathcal{M}_{\mathbf{r}}, \qquad \widehat{\mathbb{T}}^{(k)} = T_{\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)}}\mathcal{M}_{\mathbf{r}}.$$

The model dimension is

$$D_{\mathrm{eff}} = \prod_{j=1}^{d} r_j + \sum_{j=1}^{d} (p_j - r_j) r_j. \tag{37}$$

For later use, set

$$\mathfrak{d}_T = \pi_{\min}^{-1} D_{\mathrm{eff}} \log(pT). \tag{38}$$

The explicit structural complexity in Theorem S.5.1 is governed by D_{eff} rather than the ambient tensor dimension. The ambient action space still affects the rate through the logarithmic factor, the propensity cap, and the policy-induced information N_k^{eff}.

S.3.1 Regularity Conditions Used in the Main Text

The true tensor has Tucker rank $\mathbf{r}$. Its factor matrices are incoherent, and

$$\lambda_\star = \min_{j\in[d]} \sigma_{r_j}\{M_j(\mathcal{T}^\star)\} > 0.$$

The warm-up estimator satisfies $\left\|\widehat{\mathcal{T}}_{\mathrm{lr}}^{(0)} - \mathcal{T}^\star\right\|_F \le c_0\lambda_\star$ with probability at least $1 - c_{\mathrm{init}}(pT)^{-c'_{\mathrm{init}}}$. All cumulative Riemannian updates remain in this local basin. For the optimization transfer, the implemented iterate is required to satisfy (70), and the local objective is required to satisfy the quadratic-growth condition (79). These are optimization-accuracy conditions; the statistical theorem does not claim a solver-specific bound on J_k.

The capped design satisfies the predictable local-secant information conditions (40)–(41). The class is chosen before the rewards are observed and contains the local rank-$2\mathbf{r}$ secants needed to compare two feasible Tucker tensors. Lemma S.5.2 transfers this predictable condition to the realized capped design.

The target H, reference distribution q_H, and burn-in batch k_0 are fixed before online sampling. The batch-start system satisfies (44) for $k = k_0, \dots, K_T$. The remainder, variance-consistency, and Lindeberg conditions used for inference are proved in Sections S.6.3–S.6.5.

S.3.2 Effective Sample Size and Restricted Information

For every started batch, retain the realized tangent information used in the main text:

$$N_k^{\mathrm{eff}} = \inf_{\Delta\in\mathbb{T}^\star:\,\|\Delta\|_F=1} \sum_{t<t_{k+1}} \omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t\rangle^2. \tag{39}$$

Let $\mathcal{K}_{2\mathbf{r}}(r)$ be the normalized local rank-$2\mathbf{r}$ secant set

$$\left\{ \frac{\mathcal{T}_1 - \mathcal{T}_2}{\|\mathcal{T}_1 - \mathcal{T}_2\|_F} : \mathcal{T}_1, \mathcal{T}_2 \in \mathcal{M}_{\mathbf{r}} \cap B_F(\mathcal{T}^\star, r),\ \mathcal{T}_1 \neq \mathcal{T}_2 \right\}.$$

Let $\mathcal{K}_{\mathrm{env}}(r)$ be the union of $\mathcal{K}_{2\mathbf{r}}(r)$ and all normalized nonzero sums and differences of two of its elements. The envelope has Tucker rank at most $4\mathbf{r}$ and covering entropy of the same order D_{eff} up to rank-dependent constants. Its closure contains the unit sphere of $\mathbb{T}^\star$. Define the predictable lower information

$$\underline{N}_k = \inf_{\Delta\in\mathcal{K}_{2\mathbf{r}}(r)} \sum_{t<t_{k+1}} \mathbb{E}\left\{\omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t\rangle^2 \mid \mathcal{F}_{t-1}\right\}. \tag{40}$$

We assume that, almost surely and for fixed constants C_0, C_{ri},

$$\underline{N}_k \ge C_0 \mathfrak{d}_T, \qquad \sup_{\Delta\in\mathcal{K}_{\mathrm{env}}(r)} \sum_{t<t_{k+1}} \mathbb{E}\left\{\omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t\rangle^2 \mid \mathcal{F}_{t-1}\right\} \le C_{\mathrm{ri}} \underline{N}_k. \tag{41}$$

This is a condition on the predictable design induced by RITE, not on the random realized information. Lemma S.5.2 establishes the event

$$c_{\mathrm{ri}} \underline{N}_k \|\Delta\|_F^2 \le \sum_{t<t_{k+1}} \omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t\rangle^2 \le C_{\mathrm{ri}} \underline{N}_k \|\Delta\|_F^2 \tag{42}$$

uniformly over local rank-$2\mathbf{r}$ secants, together with $N_k^{\mathrm{eff}} \asymp \underline{N}_k$. Neither condition implies $N_k^{\mathrm{eff}} \asymp t_{k+1} - 1$ without an additional policy-coverage argument.

S.3.3 Batch-Start Target-Reference Directions

At time $t_k - 1$, choose an orthonormal coordinate basis for $\widehat{\mathbb{T}}^{(k-1)}$ and define

$$\begin{aligned} s_i^{(k)} &= \mathrm{vec}_{k-1}\{P_{\widehat{\mathbb{T}}^{(k-1)}}(\mathcal{A}(i))\}, \\ h_{k-1}(H) &= \mathrm{vec}_{k-1}\{P_{\widehat{\mathbb{T}}^{(k-1)}}(H)\}, \\ J_{H,k} &= \sum_{i\in\mathcal{I}} q_H(i) s_i^{(k)} s_i^{(k)\top}, \\ g_{H,k} &= J_{H,k}^\dagger h_{k-1}(H). \end{aligned} \tag{43}$$

These quantities are $\mathcal{F}_{t_k-1}$-measurable and remain fixed in batch k. Assume

$$\begin{aligned} h_{k-1}(H) &\in \mathrm{Range}(J_{H,k}), \\ \left\| J_{H,k}^{\dagger} h_{k-1}(H) \right\| \le C_{g,H}, &\qquad c_{H,T} \le h_{k-1}(H)^{\top} J_{H,k}^{\dagger} h_{k-1}(H) \le C_{v,H}. \end{aligned} \tag{44}$$

Here $c_{H,T} > 0$ may depend on the dimension. It is retained in the variance-rate conditions below and prevents the leading target variance from degenerating relative to the nonlinear remainder.

For $t \in \mathcal{E}_k$, set

$$\widehat{r}_t^{(k)} = Y_t - \langle \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)}, \mathcal{A}_t \rangle, \qquad \rho_t^H = \frac{q_H(I_t)}{\pi_t(I_t)}. \tag{45}$$

The batchwise estimator is

$$\widehat{\theta}_k(H) = \langle \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)}, H \rangle + \frac{1}{n_k} \sum_{t \in \mathcal{E}_k} \rho_t^H \{ g_{H,k}^{\top} s_{I_t}^{(k)} \} \widehat{r}_t^{(k)}. \tag{46}$$

For $K_T \ge k_0$, define

$$N_T^{\mathrm{inf}} = T - t_{k_0} + 1, \qquad \alpha_k = \frac{n_k}{N_T^{\mathrm{inf}}}, \tag{47}$$

and

$$\widehat{\theta}_{\mathrm{db}}(H) = \sum_{k=k_0}^{K_T} \alpha_k \widehat{\theta}_k(H). \tag{48}$$

The final truncated batch is included. The identity $\alpha_k / n_k = (N_T^{\mathrm{inf}})^{-1}$ is fixed when batch k_0 starts.

S.3.4 High-Probability Events

Let $\mathcal{E}_{\mathrm{init}}$ be the initialization event. Let $\mathcal{E}_{\mathrm{info}}(k)$ denote (42). Let $\mathcal{E}_{\mathrm{conc}}(k)$ be the event

$$\sup_{\Delta \in \mathcal{K}_{2\mathbf{r}}(r)} \left| \sum_{t < t_{k+1}} \omega_t^{\mathrm{lr}} \xi_t \langle \Delta, \mathcal{A}_t \rangle \right| \le C\sigma \left\{ \sqrt{\pi_{\min}^{-1} \underline{N}_k D_{\mathrm{eff}} \log(pT)} + \pi_{\min}^{-1} D_{\mathrm{eff}} \log(pT) \right\}.$$

Let $\mathcal{E}_{\mathrm{loc}}(k)$ state that all Riemannian iterates through batch k remain in the local basin. Define

$$\mathcal{E}_T = \mathcal{E}_{\mathrm{init}} \cap \bigcap_{k=1}^{K_T} \{ \mathcal{E}_{\mathrm{info}}(k) \cap \mathcal{E}_{\mathrm{conc}}(k) \cap \mathcal{E}_{\mathrm{loc}}(k) \}. \tag{49}$$

The concentration bounds below are uniform over deterministic endpoints $m \le T$. They therefore hold at $t_{k+1} - 1$. Since $K_T = O(\log T)$,

$$\mathbb{P}(\mathcal{E}_T) \ge 1 - c_1 (pT)^{-c_2}. \tag{50}$$

S.4 Full RITE Procedure

The warm-up factor spaces define one fixed decision rotation. Batch k is also tensor-elimination phase k. Its decision state advances only when the proposed action is executed.

Index a rotated coordinate by $\ell = (\ell_1, \ldots, \ell_d)$ and define

$$\mathcal{J}_{\perp} = \{ \ell : \ell_j > r_j \text{ for every } j \in [d] \}, \qquad |\mathcal{J}_{\perp}| = \prod_{j=1}^{d} (p_j - r_j). \tag{51}$$

Let

$$P = \prod_{j=1}^{d} p_j, \qquad q = P - \prod_{j=1}^{d} (p_j - r_j),$$

and order the rotated coordinates so that the q main coordinates appear first and the all-orthogonal coordinates appear last. Define

$$P_{\text{main}} = \begin{pmatrix} I_q & 0 \\ 0 & 0_{(P-q)\times(P-q)} \end{pmatrix}, \qquad I - P_{\text{main}} = \begin{pmatrix} 0_{q\times q} & 0 \\ 0 & I_{P-q} \end{pmatrix}. \tag{52}$$

Hence, for any rotated vector $x = (x_{\text{main}}^\top, x_\perp^\top)^\top$,

$$P_{\text{main}} x = \begin{pmatrix} x_{\text{main}} \\ 0 \end{pmatrix}, \qquad (I - P_{\text{main}})x = \begin{pmatrix} 0 \\ x_\perp \end{pmatrix}.$$

The decision regularizer is therefore

$$\Lambda_{\text{ucb}} = \lambda_1 P_{\text{main}} + \lambda_2 (I - P_{\text{main}}) = \begin{pmatrix} \lambda_1 I_q & 0 \\ 0 & \lambda_2 I_{P-q} \end{pmatrix}, \qquad P - q = \prod_{j=1}^{d} (p_j - r_j). \tag{53}$$

with the tail block omitted when $P = q$.

Let a_i be the fixed rotated feature. Within batch k, after c executed proposals, define

$$V_{k,c}^{\text{ucb}} = \Lambda_{\text{ucb}} + \sum_{s \in \mathcal{D}_k(c)} a_{I_s} a_{I_s}^\top, \qquad u_{k,c}(i) = \{a_i^\top (V_{k,c}^{\text{ucb}})^{-1} a_i\}^{1/2}. \tag{54}$$

After 2^{k-1} executed proposals, set

$$\begin{aligned} \widehat{b}_k &= (V_{k,2^{k-1}}^{\text{ucb}})^{-1} \sum_{s \in \mathcal{D}_k} a_{I_s} Y_s, \\ \mathcal{C}_{k+1} &= \left\{ i \in \mathcal{C}_k : a_i^\top \widehat{b}_k + \beta_{\text{ucb}} u_{k,2^{k-1}}(i) \geq \max_{j \in \mathcal{C}_k} [a_j^\top \widehat{b}_k - \beta_{\text{ucb}} u_{k,2^{k-1}}(j)] \right\}. \end{aligned} \tag{55}$$

Algorithm S1: Full RITE Procedure

Input: horizon T, warm-up length s_0, Tucker rank $\mathbf{r}$, fixed target H, prespecified burn-in batch k_0, q_0, q_H, $\{\varepsilon_t\}$, confidence radius β_{ucb}, λ_1, λ_2, $\pi_{\min}$, and iteration counts $\{J_k\}$.
Warm-up: for $t = 1, \ldots, s_0$, draw $I_t \sim q_0$, observe Y_t, and record $\pi_t(I_t) = q_0(I_t)$ and ω_t^{lr}.
Compute $\widehat{\mathcal{T}}_{\text{lr}}^{(0)}$ and $\{\widehat{U}_j^{(0)}\}_{j=1}^d$. Fix all rotated features $\{a_i\}_{i \in \mathcal{I}}$.
Set $\mathcal{C}_1 = \mathcal{I}$, $t_1 = s_0 + 1$, and $k = 1$.
while $t_k \leq T$ **do**
  Set $t \leftarrow t_k$, $c \leftarrow 0$, $\mathcal{D}_k \leftarrow \varnothing$, and $V_{k,0}^{\text{ucb}} \leftarrow \Lambda_{\text{ucb}}$.
  Compute $s_i^{(k)}$, $h_{k-1}(H)$, $J_{H,k}$, and $g_{H,k}$ from (43).
  while $c < 2^{k-1}$ and $t \leq T$ **do**
    Select $I_t^{\text{TE}} \in \arg\max_{i \in \mathcal{C}_k} u_{k,c}(i)$.
    Draw $B_t \sim \text{Bernoulli}(\varepsilon_t)$. Set $I_t = I_t^{\text{TE}}$ if $B_t = 0$; otherwise draw $I_t \sim q_H$.
    Record $\pi_t(I_t) = (1 - \varepsilon_t)\mathbf{1}\{I_t = I_t^{\text{TE}}\} + \varepsilon_t q_H(I_t)$, then observe Y_t.
    Store ω_t^{lr}, ρ_t^H, and the score contribution based on $\widehat{r}_t^{(k)}$.
    if $B_t = 0$ **then**
      Set $\mathcal{D}_k \leftarrow \mathcal{D}_k \cup \{t\}$, $c \leftarrow c + 1$, and $V_{k,c}^{\text{ucb}} \leftarrow V_{k,c-1}^{\text{ucb}} + a_{I_t} a_{I_t}^\top$.
    end if
    Set $t \leftarrow t + 1$.
  end while
  Set $t_{k+1} \leftarrow t$ and $n_k \leftarrow t_{k+1} - t_k$.
  if $c = 2^{k-1}$ **then**
    Compute $\widehat{b}_k$ and $\mathcal{C}_{k+1}$ from (55).
  else
    Set $\mathcal{C}_{k+1} \leftarrow \mathcal{C}_k$.
  end if
  Complete $\widehat{\theta}_k(H)$ from (46).
  Starting at $\widehat{\mathcal{T}}_{\text{lr}}^{(k-1)}$, perform J_k Riemannian gradient and retraction steps for the capped loss using all observations $t < t_{k+1}$; call the result $\widehat{\mathcal{T}}_{\text{lr}}^{(k)}$.

26: Set $k \leftarrow k+1$.
27: **end while**
28: If $K_T \geq k_0$, aggregate (48) and compute (120).
29: **Output:** $\widehat{\theta}_{\mathrm{db}}(H)$ and its confidence interval.

A truncated last batch does not update the active set. Its observations still enter the cumulative low-rank loss, the one-step estimator, and the sandwich variance.

S.4.1 Choice of the Inferential Distribution

Let q_{unif} be uniform on $\mathcal{I}$ and fix $\eta \in (0,1]$. For an entry target,

$$q_H = (1-\eta)\delta_{i_0} + \eta q_{\mathrm{unif}}.$$

For a gap target,

$$q_H = (1-\eta)\left\{\frac{1}{2}\delta_{i_0} + \frac{1}{2}\delta_{j_0}\right\} + \eta q_{\mathrm{unif}}.$$

The target and q_H are fixed before the online process. The batch-start tangent system changes with $\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)}$.

Lemma S.4.1 (Target Information for the Implemented Reference Distributions)**.** *Let $P = |\mathcal{I}|$. For every inference batch,*

$$\sum_{i\in\mathcal{I}} s_i^{(k)} s_i^{(k)\top} = I. \tag{56}$$

Consequently, both reference distributions above give

$$\frac{\eta}{P} I \preceq J_{H,k} \preceq I. \tag{57}$$

Hence $J_{H,k}$ is invertible and the range condition in (44) *holds automatically. For an entry target, writing $h = h_{k-1}(H)$,*

$$\|g_{H,k}\| = \frac{\|h\|}{\eta/P + (1-\eta)\|h\|^2},$$
$$h^\top J_{H,k}^{-1} h \leq \min\left\{\frac{P}{\eta}, \frac{1}{1-\eta}\right\}. \tag{58}$$

For a gap target,

$$\|g_{H,k}\| \leq \frac{\sqrt{2}P}{\eta},$$
$$h_{k-1}(H)^\top J_{H,k}^{-1} h_{k-1}(H) \leq \min\left\{\frac{2P}{\eta}, \frac{4}{1-\eta}\right\}. \tag{59}$$

The bounds containing $(1-\eta)^{-1}$ are interpreted as $+\infty$ when $\eta = 1$. Moreover,

$$h_{k-1}(H)^\top J_{H,k}^{-1} h_{k-1}(H) \geq \|h_{k-1}(H)\|^2. \tag{60}$$

Thus the lower bound in (44) *follows from the target-alignment condition*

$$\inf_{k_0 \leq k \leq K_T} \left\|P_{\widehat{\mathbb{T}}^{(k-1)}} H\right\|_F^2 \geq c_{H,T} > 0.$$

For an entry target, this condition also yields $\|g_{H,k}\| \leq \{(1-\eta)\sqrt{c_{H,T}}\}^{-1}$.

Proof. Let $\mathcal{U}_{k-1}$ be the isometry from $\widehat{\mathbb{T}}^{(k-1)}$ to its chosen coordinate space. For any coordinate vector x,

$$\sum_{i\in\mathcal{I}} (x^\top s_i^{(k)})^2 = \sum_{i\in\mathcal{I}} \langle \mathcal{U}_{k-1}^{-1} x, \mathcal{A}(i)\rangle^2 = \left\|\mathcal{U}_{k-1}^{-1} x\right\|_F^2 = \|x\|^2,$$

because the one-hot tensors form an orthonormal basis of the ambient tensor space. This proves (56). Since $\eta/P \leq q_H(i) \leq 1$ for every i, it also gives (57).

For an entry target, put $s = s_{i_0}^{(k)}$ and $a = \eta/P$. Then $J_{H,k} = aI + (1-\eta)ss^\top$ and $h_{k-1}(H) = s$. Since s is an eigenvector of $J_{H,k}$,

$$g_{H,k} = \frac{s}{a + (1-\eta)\|s\|^2}, \qquad s^\top J_{H,k}^{-1} s = \frac{\|s\|^2}{a + (1-\eta)\|s\|^2} \le \min\left\{\frac{P}{\eta}, \frac{1}{1-\eta}\right\}.$$

For a gap target, write $s_0 = s_{i_0}^{(k)}$ and $s_1 = s_{j_0}^{(k)}$. The target part of $J_{H,k}$ is $(1-\eta)(s_0 s_0^\top + s_1 s_1^\top)/2$, whereas $h_{k-1}(H) = s_0 - s_1$. For every x, let $a = s_0^\top x$ and $b = s_1^\top x$. Since

$$2h_{k-1}(H)^\top x - x^\top J_{H,k} x \le 2(a-b) - \frac{1-\eta}{2}(a^2+b^2) \le \frac{4}{1-\eta},$$

the variational identity $h^\top J^{-1} h = \sup_x\{2h^\top x - x^\top J x\}$ gives the required bound. The uniform component gives $h^\top J^{-1} h \le (P/\eta)\|h\|^2 \le 2P/\eta$ and the stated norm bounds. Finally, $J_{H,k} \preceq I$ implies $J_{H,k}^{-1} \succeq I$, which proves (60). For an entry target, $\|s\|^2 \ge c_{H,T}$ further gives $\|g_{H,k}\| \le \{(1-\eta)\sqrt{c_{H,T}}\}^{-1}$. □

Computationally, the Parseval identity in (56) gives

$$J_{H,k} = \frac{\eta}{P} I + (1-\eta) \sum_{i \in \mathrm{supp}(q_H^{\mathrm{tar}})} q_H^{\mathrm{tar}}(i) s_i^{(k)} s_i^{(k)\top}.$$

Thus, forming the reference information matrix does not require enumeration of all P actions. The entry-target direction has the closed form in (58). For a gap target, let $S = [s_{i_0}^{(k)}, s_{j_0}^{(k)}]$, $a = \eta/P$, and $b = (1-\eta)/2$. Then

$$g_{H,k} = S(aI_2 + bS^\top S)^{-1} \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

requiring only a 2×2 linear solve.

S.5 Proof of Low-Rank Estimation Error

Theorem S.5.1 (Restatement of the Low-Rank Estimation Bound)**.** *Suppose the conditions in Section S.3 hold, including* (41) *for every started batch. Then, for constants $c_1, c_2, C > 0$, with probability at least $1 - c_1(pT)^{-c_2}$,*

$$\left\|\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)} - \mathcal{T}^\star\right\|_F \le C\sigma\left\{\sqrt{\frac{\mathfrak{d}_T}{N_k^{\mathrm{eff}}}} + \frac{\mathfrak{d}_T}{N_k^{\mathrm{eff}}}\right\} + C\,\mathrm{Rem}_{\mathrm{opt}}(k), \qquad k = 1, \dots, K_T. \tag{61}$$

If the right-hand side tends to zero uniformly in k, the cumulative Riemannian estimators are uniformly consistent.

The proof has five steps. Step 0 fixes the local secant set and its information scales. Step 1 gives an exact basic inequality. Step 2 controls its curvature and score terms. Step 3 transfers the statistical solution to the finite Riemannian iterate. Step 4 makes the result uniform over the started batches.

S.5.1 Step 0: Local Coordinates and Information Events

Fix a batch k. For a candidate tensor $\mathcal{T}$ in the local basin of $\mathcal{T}^\star$, write

$$\Delta = \mathcal{T} - \mathcal{T}^\star. \tag{62}$$

The normalized displacement belongs to the previously defined rank-$2\mathbf{r}$ class:

$$\Delta / \|\Delta\|_F \in \mathcal{K}_{2\mathbf{r}}(r). \tag{63}$$

Lemma S.10.14 gives covering entropy $CD_{\mathrm{eff}} \log(Cr/\epsilon)$ for this class and for normalized sums and differences of its elements.

For any local displacement Δ, define the empirical quadratic form and the weighted noise score by

$$\mathcal{Q}_k(\Delta) = \sum_{t<t_{k+1}} \omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t \rangle^2, \tag{64}$$

$$\mathcal{M}_k(\Delta) = \sum_{t<t_{k+1}} \omega_t^{\mathrm{lr}} \xi_t \langle \Delta, \mathcal{A}_t \rangle. \tag{65}$$

For a fixed direction, the reward-half-step sub-Gaussian variance process is the realized squared-coefficient sum

$$\mathfrak{V}_{k,\mathrm{obs}}^{\mathrm{lr}}(\Delta) = \sum_{t<t_{k+1}} (\omega_t^{\mathrm{lr}})^2 \langle \Delta, \mathcal{A}_t \rangle^2. \tag{66}$$

On the realized curvature event, $(\omega_t^{\mathrm{lr}})^2 \le \pi_{\min}^{-1} \omega_t^{\mathrm{lr}}$ gives

$$\sup_{\Delta \in \mathcal{K}_{\mathrm{env}}(r)} \mathfrak{V}_{k,\mathrm{obs}}^{\mathrm{lr}}(\Delta) \le C \pi_{\min}^{-1} \underline{N}_k. \tag{67}$$

This form keeps the two filtration stages separate: the action and capped weight are revealed first, and the reward noise is revealed second.

We work on the event

$$\mathcal{E}_k = \mathcal{E}_{\mathrm{init}} \cap \mathcal{E}_{\mathrm{loc}}(k) \cap \mathcal{E}_{\mathrm{info}}(k) \cap \mathcal{E}_{\mathrm{conc}}(k). \tag{68}$$

Here $\mathcal{E}_{\mathrm{info}}(k)$ compares the empirical and predictable quadratic forms, while $\mathcal{E}_{\mathrm{conc}}(k)$ controls (65) uniformly on the local secant set. At this point the proof is conditional on $\mathcal{E}_k$; its probability is handled after the deterministic inequality has been solved.

S.5.2 Step 1: The Capped-IPW Basic Inequality

At batch k, consider the cumulative capped-IPW loss

$$L_k(\mathcal{T}) = \frac{1}{2} \sum_{t<t_{k+1}} \omega_t^{\mathrm{lr}} \{ Y_t - \langle \mathcal{T}, \mathcal{A}_t \rangle \}^2, \qquad \mathcal{T} \in \mathcal{M}_{\mathbf{r}}. \tag{69}$$

Let $\widehat{\mathcal{T}}_k^{\circ}$ minimize L_k over the intersection of $\mathcal{M}_{\mathbf{r}}$ and the local basin containing $\mathcal{T}^\star$. The implemented Riemannian iterate satisfies

$$L_k(\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)}) \le L_k(\widehat{\mathcal{T}}_k^{\circ}) + \epsilon_k^{\mathrm{opt}}. \tag{70}$$

Since $\mathcal{T}^\star$ is feasible, $L_k(\widehat{\mathcal{T}}_k^{\circ}) \le L_k(\mathcal{T}^\star)$; hence (70) gives

$$L_k(\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)}) - L_k(\mathcal{T}^\star) \le \epsilon_k^{\mathrm{opt}}. \tag{71}$$

Substituting $Y_t = \langle \mathcal{T}^\star, \mathcal{A}_t \rangle + \xi_t$ into (69), we obtain the exact local expansion

$$L_k(\mathcal{T}^\star + \Delta) - L_k(\mathcal{T}^\star) = \frac{1}{2} \mathcal{Q}_k(\Delta) - \mathcal{M}_k(\Delta). \tag{72}$$

Consequently, for $\Delta_k = \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)} - \mathcal{T}^\star$,

$$\frac{1}{2} \mathcal{Q}_k(\Delta_k) \le \mathcal{M}_k(\Delta_k) + \epsilon_k^{\mathrm{opt}}. \tag{73}$$

The cap does not create a nonzero conditional mean for the noise score. Conditional on $\mathcal{F}_{t-1/2}$, the weight ω_t^{lr} and the action tensor are fixed, whereas $\mathbb{E}(\xi_t \mid \mathcal{F}_{t-1/2}) = 0$. Capping changes the information geometry and bounds the score coefficients; both effects are already visible in (64) and (66).

Our task is now precise. The left-hand side of (73) must provide quadratic growth in $\|\Delta_k\|_F$, while the score on the right must remain linear in that norm. The next two lemmas give these bounds.

S.5.3 Step 2: Curvature and Weighted Score Bounds

Lemma S.5.2 (Restricted Capped-IPW Curvature)**.** *Suppose* (41) *holds. Then there exist constants* $c, C_1, C_2 > 0$ *such that, with probability at least* $1 - C_1(pT)^{-C_2}$*,*

$$c\underline{N}_k \|\Delta\|_F^2 \le \mathcal{Q}_k(\Delta) \le C\underline{N}_k \|\Delta\|_F^2 \tag{74}$$

uniformly over differences of two feasible tensors in the local basin. On the same event, $N_k^{\mathrm{eff}} \asymp \underline{N}_k$*.*

The proof is given in Section S.11.1. The result is a restricted matrix-martingale bound on a Tucker secant net. The uniform component of q_H is useful here because it prevents the predictable quadratic form from losing directions that are weakly visited by the elimination policy.

Lemma S.5.3 (Uniform Capped-IPW Score Bound)**.** *Under the conditional sub-Gaussian noise assumption and the conditions of Lemma S.5.2, there exist constants* $C, C_1, C_2 > 0$ *such that, with probability at least* $1 - C_1(pT)^{-C_2}$*,*

$$|\mathcal{M}_k(\Delta)| \le C\sigma \left\{ \sqrt{\pi_{\min}^{-1}\underline{N}_k D_{\mathrm{eff}} \log(pT)} + \pi_{\min}^{-1} D_{\mathrm{eff}} \log(pT) \right\} \|\Delta\|_F \tag{75}$$

uniformly over local displacements Δ*.*

The proof is given in Section S.11.2. Conditional on the action history, the summands in (65) are martingale differences. A net argument contributes D_{eff} rather than the ambient product $p_1 \cdots p_d$.

Combining (73), (74), and (75) yields

$$\begin{aligned} c\underline{N}_k \|\Delta_k\|_F^2 \le C\sigma \left\{ \sqrt{\pi_{\min}^{-1}\underline{N}_k D_{\mathrm{eff}} \log(pT)} + \pi_{\min}^{-1} D_{\mathrm{eff}} \log(pT) \right\} \|\Delta_k\|_F \\ + 2\epsilon_k^{\mathrm{opt}}. \end{aligned} \tag{76}$$

Adaptive sampling is already present in the predictable design and martingale score. It does not create an additional deterministic discrepancy in the ambient capped-IPW loss. The only remaining difference between the local minimizer and the implemented estimator is finite Riemannian optimization.

Solving (76) gives the statistical radius

$$\left\| \widehat{\mathcal{T}}_k^{\circ} - \mathcal{T}^{\star} \right\|_F \le C\sigma \left\{ \sqrt{\frac{\mathfrak{d}_T}{N_k^{\mathrm{eff}}}} + \frac{\mathfrak{d}_T}{N_k^{\mathrm{eff}}} \right\}. \tag{77}$$

The remaining issue is that RITE returns a finite Riemannian iterate rather than the exact local minimizer.

S.5.4 Step 3: From the Local Minimizer to the Riemannian Iterate

Let $\mathcal{T}^{(m)}$ be the inner iterate in batch k. Its update is

$$\mathcal{T}^{(m+1)} = R_{\mathcal{T}^{(m)}} \left\{ -\eta_m P_{T_{\mathcal{T}^{(m)}}\mathcal{M}_{\mathbf{r}}} \nabla L_k(\mathcal{T}^{(m)}) \right\}. \tag{78}$$

The local geometry in Lemma S.10.14 makes the retraction error second order. Restricted curvature then converts objective suboptimality into Frobenius error.

Lemma S.5.4 (Riemannian Optimization Transfer)**.** *On* $\mathcal{E}_{\mathrm{loc}}(k) \cap \mathcal{E}_{\mathrm{info}}(k)$*, suppose* (70) *holds and the local objective has quadratic growth:*

$$L_k(\mathcal{T}) - L_k(\widehat{\mathcal{T}}_k^{\circ}) \ge cN_k^{\mathrm{eff}} \left\| \mathcal{T} - \widehat{\mathcal{T}}_k^{\circ} \right\|_F^2 \tag{79}$$

for feasible $\mathcal{T}$ *in the local basin. Then*

$$\left\| \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)} - \widehat{\mathcal{T}}_k^{\circ} \right\|_F \le C \left(\frac{\epsilon_k^{\mathrm{opt}}}{N_k^{\mathrm{eff}}} \right)^{1/2} =: C\,\mathrm{Rem}_{\mathrm{opt}}(k). \tag{80}$$

The theorem treats (70) *as the required optimization accuracy. It does not assert that a particular stepsize or a particular number* J_k *of inner iterations attains this tolerance.*

The proof is given in Section S.11.3. Combining the lemma with (77) gives the batch-wise bound

$$\left\|\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)} - \mathcal{T}^\star\right\|_F \le C\sigma\left\{\sqrt{\frac{\mathfrak{d}_T}{N_k^{\mathrm{eff}}}} + \frac{\mathfrak{d}_T}{N_k^{\mathrm{eff}}}\right\} + C\,\mathrm{Rem}_{\mathrm{opt}}(k). \tag{81}$$

S.5.5 Step 4: Uniform Control over Batches

The initialization event has probability at least $1 - c_{\mathrm{init}}(pT)^{-c'_{\mathrm{init}}}$. Lemmas S.5.2 and S.5.3 each fail with probability at most $C_1(pT)^{-C_2}$ for a fixed batch. The number of batches is logarithmic under the geometric schedule. Choosing the constants in the confidence radii large enough and applying a union bound over $k = 1, \dots, K_T$ gives

$$\mathbb{P}\left(\bigcap_{k=1}^{K_T} \mathcal{E}_k\right) \ge 1 - c_1(pT)^{-c_2} \tag{82}$$

for constants $c_1, c_2 > 0$.

On this event, (81) holds simultaneously for all batches. This is the conclusion of Theorem S.5.1. The consistency requirement is $N_k^{\mathrm{eff}} \gg \pi_{\min}^{-1} D_{\mathrm{eff}} \log(pT)$. When $\pi_{\min}$ is fixed and a separate policy-coverage argument gives $N_k^{\mathrm{eff}} \asymp t_{k+1} - 1$, this reduces to $t_{k+1} - 1 \gg D_{\mathrm{eff}} \log(pT)$.

S.6 Proof of Doubly Debiased Inference

Theorem S.6.2 (Restatement of Debiased Normality)**.** *Fix H, q_H, and k_0 before online sampling. Suppose* (44)*,* all *conditions of Corollary S.6.1,* (88)*, and* (122) *hold. Assume also that, almost surely,*

$$0 < \nu_-^2 \le \mathbb{E}(\xi_t^2 \mid \mathcal{F}_{t-1/2}) \le \nu_+^2 < \infty$$

and that the conditional fourth moments are uniformly bounded. Then

$$\frac{\widehat{\theta}_{\mathrm{db}}(H) - \theta^\star(H)}{\widehat{V}_T(H)^{1/2}} \Rightarrow N(0,1). \tag{83}$$

Consequently, the Wald interval based on $\widehat{V}_T(H)$ has asymptotic coverage $1 - \alpha$.

The proof follows the decomposition used in online low-rank inference. Step 0 fixes the batch-start score. Step 1 derives an exact identity. Step 2 bounds its projection, design, and normal-space remainders. Steps 3 and 4 establish the martingale limit and variance consistency.

Our aim is to expose the cancellation within each batch before invoking a limit theorem. This order matters. The direction $g_{H,k}$ in (43) is computed from the fixed distribution q_H and the tangent space available at $t_k - 1$. It is therefore an actual predictable quantity, not an oracle replacement introduced only in the proof. Section S.6.1 records the corresponding martingale score. Section S.6.2 derives an exact batch-wise expansion. Section S.6.3 bounds its three remainders. We then establish the martingale normal approximation and variance consistency.

S.6.1 Step 0: Batch-Start Score System

For $k \ge k_0$, write

$$\Delta_{k-1} = \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)} - \mathcal{T}^\star, \qquad r_{k-1} = \|\Delta_{k-1}\|_F, \qquad \eta_{k-1} = \left\|P_{\widehat{\mathbb{T}}^{(k-1)}} - P_{\mathbb{T}^\star}\right\|_{\mathrm{op}}. \tag{84}$$

The tangent coordinates $s_i^{(k)}$, $h_{k-1}(H)$, information matrix $J_{H,k}$, and direction $g_{H,k}$ are defined in (43). Under (44),

$$J_{H,k} g_{H,k} = h_{k-1}(H), \qquad \|g_{H,k}\| \le C_{g,H}, \qquad h_{k-1}(H)^\top g_{H,k} \le C_{v,H}. \tag{85}$$

For $t \in \mathcal{E}_k$, define

$$\psi_{k,t}(H) = \frac{1}{N_T^{\mathrm{inf}}} \rho_t^H \{g_{H,k}^\top s_{I_t}^{(k)}\} \xi_t. \tag{86}$$

The conditional variance of the complete leading sum is

$$V_T(H) = \sum_{k=k_0}^{K_T} \sum_{t \in \mathcal{E}_k} \mathbb{E}\left[\psi_{k,t}(H)^2 \mid \mathcal{F}_{t-1}\right]. \tag{87}$$

We assume that there is a deterministic sequence $v_T(H) > 0$ such that

$$\frac{V_T(H)}{v_T(H)} \to_p 1. \tag{88}$$

This condition fixes the scale used by the triangular-array martingale CLT. It is not replaced by the random terminal quantity $V_T(H)$ inside the Lindeberg condition. Bounded conditional noise variances alone do not force this stabilization under an arbitrary adaptive policy, so it is stated explicitly. If $q_{H,\min} = \min_i q_H(i)$, then (44) and the conditional variance lower bound imply

$$V_T(H) \geq \frac{\nu_-^2 q_{H,\min} c_{H,T}}{N_T^{\inf}} \tag{89}$$

almost surely. Thus target alignment prevents exact degeneracy, while (88) records the sharper asymptotic scale needed to dominate the remainders. At the moment the reward is observed, $g_{H,k}$ is already fixed, while I_t, $s_{I_t}^{(k)}$, and ρ_t^H are $\mathcal{F}_{t-1/2}$-measurable. Hence

$$\mathbb{E}\{\psi_{k,t}(H) \mid \mathcal{F}_{t-1/2}\} = 0. \tag{90}$$

This is the exact martingale property used below. No full-sample empirical direction is substituted into (86).

S.6.2 Step 1: Exact Batch-Wise Expansion

Fix a batch k and abbreviate $P_{k-1} = P_{\widehat{\mathbb{T}}^{(k-1)}}$. Let

$$\delta_{k-1} = \mathrm{vec}_{k-1}\{P_{k-1}(\Delta_{k-1})\}. \tag{91}$$

For an action i, decompose its interaction with the nuisance error as

$$\langle \Delta_{k-1}, \mathcal{A}(i) \rangle = s_i^{(k)\top} \delta_{k-1} + b_{i,k}, \tag{92}$$

where

$$b_{i,k} = \langle (I - P_{k-1})\Delta_{k-1}, \mathcal{A}(i) \rangle.$$

The target admits the analogous decomposition

$$\langle \Delta_{k-1}, H \rangle = h_{k-1}(H)^\top \delta_{k-1} + R_{\mathrm{proj},k}(H), \tag{93}$$

with

$$R_{\mathrm{proj},k}(H) = \langle (I - P_{k-1})\Delta_{k-1}, H \rangle. \tag{94}$$

The residual in (45) is

$$\widehat{r}_t^{(k)} = \xi_t - s_{I_t}^{(k)\top} \delta_{k-1} - b_{I_t,k}. \tag{95}$$

Substituting this identity into (46) gives

$$\begin{aligned} \widehat{\theta}_k(H) - \theta^\star(H) = {} & h_{k-1}(H)^\top \delta_{k-1} + R_{\mathrm{proj},k}(H) \\ & + \frac{1}{n_k} \sum_{t \in \mathcal{E}_k} \rho_t^H \{g_{H,k}^\top s_{I_t}^{(k)}\} \xi_t \\ & - g_{H,k}^\top \left\{ \frac{1}{n_k} \sum_{t \in \mathcal{E}_k} \rho_t^H s_{I_t}^{(k)} s_{I_t}^{(k)\top} \right\} \delta_{k-1} \\ & - \frac{1}{n_k} \sum_{t \in \mathcal{E}_k} \rho_t^H \{g_{H,k}^\top s_{I_t}^{(k)}\} b_{I_t,k}. \end{aligned} \tag{96}$$

Now define the realized target-reference design

$$\widehat{J}_{H,k} = \frac{1}{n_k} \sum_{t \in \mathcal{E}_k} \rho_t^H s_{I_t}^{(k)} s_{I_t}^{(k)\top}. \tag{97}$$

The reason for introducing q_H now becomes visible. Conditional on the history before action t,

$$\begin{aligned} \mathbb{E}\left[\rho_t^H s_{I_t}^{(k)} s_{I_t}^{(k)\top} \mid \mathcal{F}_{t-1}\right] &= \sum_{i \in \mathcal{I}} \pi_t(i) \frac{q_H(i)}{\pi_t(i)} s_i^{(k)} s_i^{(k)\top} \\ &= J_{H,k}. \end{aligned} \tag{98}$$

Using (85), the tangent bias in (96) becomes

$$h_{k-1}(H)^\top \delta_{k-1} - g_{H,k}^\top \widehat{J}_{H,k} \delta_{k-1} = g_{H,k}^\top (J_{H,k} - \widehat{J}_{H,k}) \delta_{k-1}. \tag{99}$$

We have therefore obtained the exact identity

$$\widehat{\theta}_k(H) - \theta^\star(H) = \frac{1}{\alpha_k} \sum_{t \in \mathcal{E}_k} \psi_{k,t}(H) + R_{\mathrm{proj},k}(H) + R_{\mathrm{design},k}(H) + R_{\mathrm{orth},k}(H), \tag{100}$$

where

$$R_{\mathrm{design},k}(H) = g_{H,k}^\top (J_{H,k} - \widehat{J}_{H,k}) \delta_{k-1}, \tag{101}$$

$$R_{\mathrm{orth},k}(H) = -\frac{1}{n_k} \sum_{t \in \mathcal{E}_k} \rho_t^H \{g_{H,k}^\top s_{I_t}^{(k)}\} b_{I_t,k}. \tag{102}$$

Multiplying (100) by α_k and summing over the inference batches yields

$$\widehat{\theta}_{\mathrm{db}}(H) - \theta^\star(H) = \sum_{k=k_0}^{K_T} \sum_{t \in \mathcal{E}_k} \psi_{k,t}(H) + \mathrm{Rem}_{\mathrm{db}}(H), \tag{103}$$

where

$$\mathrm{Rem}_{\mathrm{db}}(H) = \sum_{k=k_0}^{K_T} \alpha_k \left\{R_{\mathrm{proj},k}(H) + R_{\mathrm{design},k}(H) + R_{\mathrm{orth},k}(H)\right\}. \tag{104}$$

S.6.3 Step 2: Bounds for the Three Remainders

The expansion above has no batch-lag term. The correction in batch k is attached to the same baseline $\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k-1)}$ that produced its residual. It also has no score-direction replacement term, because $g_{H,k}$ is part of the implemented estimator and is predictable.

Lemma S.6.5 (Batch-Wise Debiasing Remainder)**.** *Suppose the local Tucker geometry and target-direction conditions hold. Let*

$$\zeta_{H,k} = \left\| \widehat{J}_{H,k} - J_{H,k} \right\|_{\mathrm{op}}. \tag{105}$$

Then, on the factor-stability event,

$$|R_{\mathrm{proj},k}(H)| \le C \|H\|_F \left(r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star} \right), \tag{106}$$

$$|R_{\mathrm{design},k}(H)| \le C\, C_{g,H} \zeta_{H,k} r_{k-1}, \tag{107}$$

$$|R_{\mathrm{orth},k}(H)| \le C\, C_{g,H} \left(r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star} \right) \overline{\rho}_{H,k}, \tag{108}$$

where

$$\overline{\rho}_{H,k} = \frac{1}{n_k} \sum_{t \in \mathcal{E}_k} \rho_t^H \left\| s_{I_t}^{(k)} \right\|.$$

Moreover, for every $x > 0$, matrix martingale concentration gives

$$\mathbb{P}\left\{ \zeta_{H,k} > C \left(\sqrt{\frac{v_{H,k}\{D_{\mathrm{eff}} + x\}}{n_k}} + \frac{D_{\mathrm{eff}} + x}{n_k \underline{\varepsilon}_k} \right) \middle| \mathcal{F}_{t_k - 1} \right\} \le 2e^{-x}, \tag{109}$$

where $v_{H,k}$ is the conditional matrix-variance scale of the stabilized reference score. Moreover, $v_{H,k} \le C\underline{\varepsilon}_k^{-1}$ because $\pi_t(i) \ge \varepsilon_t q_H(i)$ and $\left\|s_i^{(k)}\right\| \le 1$. The same argument applied to the scalar process gives

$$\mathbb{P}\left[\overline{\rho}_{H,k} > 1 + C\left\{\sqrt{\frac{x}{n_k\underline{\varepsilon}_k}} + \frac{x}{n_k\underline{\varepsilon}_k}\right\}\middle|\mathcal{F}_{t_k-1}\right] \le 2e^{-x}. \tag{110}$$

The proof is given in Section S.11.4. The first and third bounds use local Tucker curvature. The second uses only the normal equation $J_{H,k}g_{H,k} = h_{k-1}(H)$ and concentration of the known reference design.

Lemma S.6.6 (Debiased Linear Representation)**.** *If*

$$\begin{aligned}\sum_{k=k_0}^{K_T}\alpha_k\Bigg[\|H\|_F\left\{r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star}\right\} + C_{g,H}\zeta_{H,k}r_{k-1} \\ + C_{g,H}\left\{r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star}\right\}\overline{\rho}_{H,k}\Bigg] = o_p\{v_T(H)^{1/2}\}.\end{aligned} \tag{111}$$

then

$$\widehat{\theta}_{\mathrm{db}}(H) - \theta^\star(H) = \sum_{k=k_0}^{K_T}\sum_{t\in\mathcal{E}_k}\psi_{k,t}(H) + o_p\{v_T(H)^{1/2}\}. \tag{112}$$

Corollary S.6.1 (A Checkable Product-Rate Condition)**.** *Let*

$$\bar{r}_k = C\sigma\left\{\sqrt{\frac{\mathfrak{d}_T}{N_k^{\mathrm{eff}}}} + \frac{\mathfrak{d}_T}{N_k^{\mathrm{eff}}}\right\} + \mathrm{Rem}_{\mathrm{opt}}(k), \tag{113}$$

and suppose $r_k = O_p(\bar{r}_k)$ and $\eta_k = O_p(\bar{r}_k/\lambda_\star)$ uniformly over the inference batches. Let

$$L_T = D_{\mathrm{eff}} + \log(pT), \qquad \mathcal{K}_T^{\mathrm{conc}} = \{k_0 \le k \le K_T : n_k\underline{\varepsilon}_k \ge L_T\},$$

and let $\mathcal{K}_T^{\mathrm{short}}$ be the remaining inference batches. If

$$\begin{aligned}\sum_{k\in\mathcal{K}_T^{\mathrm{conc}}}\alpha_k\Bigg[\{\|H\|_F + C_{g,H}\}\frac{\bar{r}_{k-1}^2}{\lambda_\star} \\ + C_{g,H}\bar{r}_{k-1}\left\{\sqrt{\frac{L_T}{n_k\underline{\varepsilon}_k}} + \frac{L_T}{n_k\underline{\varepsilon}_k}\right\}\Bigg] = o_p\{v_T(H)^{1/2}\},\end{aligned} \tag{114}$$

Define

$$A_k = r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star}.$$

The short-batch contribution is required to obey

$$\begin{aligned}\sum_{k\in\mathcal{K}_T^{\mathrm{short}}}\alpha_k\Big[\|H\|_F A_k + C_{g,H}r_{k-1}(1+\overline{\rho}_{H,k}) \\ + C_{g,H}A_k\overline{\rho}_{H,k}\Big] = o_p\{v_T(H)^{1/2}\}.\end{aligned} \tag{115}$$

then Lemma S.6.6 applies. In particular, a terminal truncated batch with too few target-reference observations is controlled by (115)*, rather than by an invalid within-batch concentration claim.*

For the distributional step, define the predictable maximal leverage

$$\mathcal{L}_T(H) = \max_{\substack{k_0\le k\le K_T,\ t\in\mathcal{E}_k \\ i\in\mathcal{I}}}\frac{q_H(i)|g_{H,k}^\top s_i^{(k)}|}{\pi_t(i)N_T^{\mathrm{inf}}v_T(H)^{1/2}}. \tag{116}$$

If

$$\mathcal{L}_T(H) \to_p 0 \tag{117}$$

then the realized maximal score is also negligible and the conditional Lindeberg condition in Lemma S.6.7 follows from the bounded conditional fourth moment.

Proof. For $k \in \mathcal{K}_T^{\mathrm{conc}}$, (110) and a union bound give $\max_{k\in\mathcal{K}_T^{\mathrm{conc}}} \overline{\rho}_{H,k} = O_p(1)$. The low-rank error bound, factor perturbation, and (109), with $x = C\log(pT)$, bound the corresponding summands in (111) by those in (114). For $k \in \mathcal{K}_T^{\mathrm{short}}$, the deterministic bounds $\zeta_{H,k} \le 1 + \overline{\rho}_{H,k}$ and (106)–(108) are bounded by the left-hand side of (115). The two rate conditions therefore imply (111) and Lemma S.6.6. The finite Riemannian optimization error enters through $\bar{r}_k$ and is not counted again. Finally, the definition of $\mathcal{L}_T(H)$ bounds every realized coefficient by $\mathcal{L}_T(H)v_T(H)^{1/2}$; the fourth-moment calculation is given explicitly in the proof of Lemma S.6.7. □

At this point the low-rank part of the argument is complete. The remaining work concerns the distribution of the exact martingale sum in (112).

S.6.4 Step 3: Martingale Normal Approximation

The filtration $\mathcal{F}_{t-1} \subset \mathcal{F}_{t-1/2} \subset \mathcal{F}_t$ reveals I_t before Y_t. Equation (90) and iterated expectation give

$$\mathbb{E}\{\psi_{k,t}(H) \mid \mathcal{F}_{t-1}\} = 0,$$

so the leading array is an exact martingale difference with respect to $\{\mathcal{F}_t\}$.

Lemma S.6.7 (Martingale Normal Approximation)**.** *Suppose* (88) *holds and, for every* $\epsilon > 0$*,*

$$\frac{1}{v_T(H)} \sum_{k=k_0}^{K_T} \sum_{t\in\mathcal{E}_k} \mathbb{E}\left[\psi_{k,t}(H)^2 \mathbf{1}\left\{|\psi_{k,t}(H)| > \epsilon v_T(H)^{1/2}\right\} \middle| \mathcal{F}_{t-1}\right] \to_p 0. \tag{118}$$

Then

$$\frac{\sum_{k=k_0}^{K_T} \sum_{t\in\mathcal{E}_k} \psi_{k,t}(H)}{v_T(H)^{1/2}} \Rightarrow N(0,1). \tag{119}$$

The same conclusion holds with $V_T(H)^{1/2}$ *in the denominator.*

The proof is deferred to Section S.11.6. The useful bound is $\rho_t^H \le \underline{\varepsilon}_k^{-1}$. Unlike capped-IPW, this bound comes from the policy mixture itself and leaves the target-reference identity (98) exact.

S.6.5 Step 4: Variance Consistency

The empirical variance uses the same batch-start direction as the estimator:

$$\widehat{V}_T(H) = \sum_{k=k_0}^{K_T} \sum_{t\in\mathcal{E}_k} \left(\frac{1}{N_T^{\mathrm{inf}}} \rho_t^H g_{H,k}^\top s_{I_t}^{(k)} \widehat{r}_t^{(k)}\right)^2. \tag{120}$$

To compare (120) with (87), write

$$d_{t,k} = \langle \Delta_{k-1}, \mathcal{A}_t \rangle, \qquad \widehat{r}_t^{(k)} = \xi_t - d_{t,k}.$$

For

$$c_{k,t}(H) = \frac{1}{N_T^{\mathrm{inf}}} \rho_t^H g_{H,k}^\top s_{I_t}^{(k)}, \qquad \nu_t^2 = \mathbb{E}(\xi_t^2 \mid \mathcal{F}_{t-1/2}),$$

we have the exact decomposition

$$\begin{aligned}
\widehat{V}_T(H) - V_T(H) = &\underbrace{\sum_{k,t} \left[c_{k,t}(H)^2 \nu_t^2 - \mathbb{E}\{c_{k,t}(H)^2 \nu_t^2 \mid \mathcal{F}_{t-1}\}\right]}_{\mathcal{V}_{0T}(H)} \\
&+ \underbrace{\sum_{k,t} c_{k,t}(H)^2 \left[\xi_t^2 - \nu_t^2\right]}_{\mathcal{V}_{1T}(H)} \\
&\underbrace{- 2\sum_{k,t} c_{k,t}(H)^2 \xi_t d_{t,k}}_{\mathcal{V}_{2T}(H)} + \underbrace{\sum_{k,t} c_{k,t}(H)^2 d_{t,k}^2}_{\mathcal{V}_{3T}(H)}.
\end{aligned} \tag{121}$$

The notation $\sum_{k,t}$ in (121) means $\sum_{k=k_0}^{K_T}\sum_{t\in\mathcal{E}_k}$. The term $\mathcal{V}_{0T}(H)$ is the action-design fluctuation revealed when I_t is drawn. The term $\mathcal{V}_{1T}(H)$ is the reward-variance fluctuation revealed with Y_t. The remaining terms are a martingale cross term and a squared nuisance error.

Lemma S.6.8 (Self-Normalized Variance Consistency)**.** *Under the conditions of Lemmas S.6.6 and S.6.7, suppose*

$$\frac{\sum_{k,t} c_{k,t}(H)^2 r_{k-1}^2}{v_T(H)} \to_p 0 \tag{122}$$

and (117) *holds. Then*

$$\frac{\widehat{V}_T(H)}{V_T(H)} \to_p 1, \qquad \frac{\widehat{V}_T(H)}{v_T(H)} \to_p 1. \tag{123}$$

The proof is given in Section S.11.7. We do not estimate a separate global noise variance in the inferential direction. Residual squares enter only the sandwich variance in (120).

S.6.6 Completion of the Proof

Lemma S.6.6 gives the exact martingale representation up to an $o_p\{v_T(H)^{1/2}\}$ remainder. Lemma S.6.7 yields a standard normal limit after normalization by the deterministic scale $v_T(H)^{1/2}$. Lemma S.6.8 permits its replacement by $\widehat{V}_T(H)^{1/2}$. Slutsky's theorem proves (83).

S.7 Proof of Regret Bound

Let $P = \prod_{j=1}^d p_j$, $\overline{M} = T - s_0$, and let $\beta \in \mathbb{R}^P$ be the fixed-rotation parameter, so that $\langle \mathcal{T}^\star, \mathcal{A}(i)\rangle = a_i^\top \beta$. Partition $\beta = (\beta_{\text{main}}^\top, \beta_\perp^\top)^\top$ according to (51). Fix bounds $B_{\text{main}} \geq \|\beta_{\text{main}}\|_2$ and $B_\perp \geq \|\beta_\perp\|_2$. Choose

$$0 < \lambda_1 \leq p^{-d}, \qquad \lambda_2 = \frac{\overline{M}}{q\log(1+\overline{M}/\lambda_1)}, \tag{124}$$

set $\Lambda = \lambda_1 P_{\text{main}} + \lambda_2(I - P_{\text{main}})$, and define

$$\Gamma_M = q\log\left(1+\frac{M}{q\lambda_1}\right) + (P-q)\log\left(1+\frac{M}{(P-q)\lambda_2}\right), \tag{125}$$

$$\xi_M(\delta) = \sigma\sqrt{2\log\left\{\frac{2|\mathcal{I}|(1+\log_2 M)}{\delta}\right\}} + \sqrt{\lambda_1}B_{\text{main}} + \sqrt{\lambda_2}B_\perp. \tag{126}$$

The convention is $\log_2 M = 0$ when $M \leq 1$. The second term in Γ_M is omitted when $P = q$.

Theorem S.7.3 (Restatement of the RITE Regret Bound)**.** *Assume $\|a_i\|_2 \leq 1$, the reward gaps are bounded by $\Delta_{\max}$, and the noise obeys* (28)*. For every $\delta \in (0,1)$, set $\beta_{\text{ucb}} = \xi_{\overline{M}}(\delta/2)$ and*

$$M_T = T - s_0 - N_T^{\text{tar}}, \qquad N_T^{\text{tar}} = \sum_{t=s_0+1}^{T} B_t, \qquad S_T^\varepsilon = \sum_{t=s_0+1}^{T}\varepsilon_t.$$

Then, with probability at least $1-\delta$,

$$\begin{aligned} R_T^{\text{RITE}} &\leq s_0\Delta_{\max} + C\{\Delta_{\max} + \xi_{\overline{M}}(\delta/2)\}\sqrt{M_T\Gamma_{\overline{M}}} \\ &\quad + \Delta_{\max}\left[S_T^\varepsilon + 2\sqrt{S_T^\varepsilon\log(2/\delta)} + \frac{2}{3}\log(2/\delta)\right]. \end{aligned} \tag{127}$$

Under (30)*, $S_T^\varepsilon \leq 2c_\varepsilon\sqrt{T}$.*

For balanced dimensions and fixed ranks, $q = O(p^{d-1})$. If $s_0 = \widetilde{O}(p^{d/2})$ and

$$\sqrt{\lambda_1}B_{\text{main}} = \widetilde{O}(1), \qquad \sqrt{\lambda_2}B_\perp = \widetilde{O}(1), \tag{128}$$

then $\xi_{\overline{M}}(\delta) = \widetilde{O}_\delta(1)$ and

$$R_T^{\text{RITE}} = \widetilde{O}_\delta\left\{p^{d/2} + p^{(d-1)/2}\sqrt{T} + \sqrt{T}\right\}. \tag{129}$$

The last term is the target-branch contribution.

S.7.1 Regret Decomposition by Sampling Branch

For any action i, let $\Delta_i = \langle \mathcal{T}^\star, \mathcal{A}^\star - \mathcal{A}(i) \rangle$. For $t > s_0$,

$$\Delta_{I_t} = (1 - B_t)\Delta_{I_t^{\mathrm{TE}}} + B_t \Delta_{I_t}. \tag{130}$$

The first sum is indexed by executed proposals. The second is bounded by $\Delta_{\max} N_T^{\mathrm{tar}}$.

We first control the fixed-rotation linear problem.

Lemma S.7.9 (Fixed-Rotation Tail)**.** *For the warm-up rotation,*

$$\|\beta_\perp\|_2 \le \|\mathcal{G}^\star\|_F \prod_{j=1}^{d} \left\| \widehat{U}_{j,\perp}^{(0)\top} U_j^\star \right\|_{\mathrm{op}}. \tag{131}$$

Consequently,

$$\|\beta\|_\Lambda \le \sqrt{\lambda_1}\, \|\mathcal{T}^\star\|_F + \sqrt{\lambda_2}\, \|\mathcal{G}^\star\|_F \prod_{j=1}^{d} \left\| \widehat{U}_{j,\perp}^{(0)\top} U_j^\star \right\|_{\mathrm{op}}. \tag{132}$$

Proof. The tail block is obtained by multiplying $\mathcal{T}^\star$ by $\widehat{U}_{j,\perp}^{(0)\top}$ in every mode. Tucker multilinearity gives

$$\beta_\perp = \mathrm{vec}\left\{ \mathcal{G}^\star \times_{j=1}^{d} (\widehat{U}_{j,\perp}^{(0)\top} U_j^\star) \right\}.$$

Repeated use of $\|\mathcal{X} \times_j A\|_F \le \|A\|_{\mathrm{op}} \|\mathcal{X}\|_F$ proves (131). The main block has norm at most $\|\mathcal{T}^\star\|_F$. The definition of $\|\cdot\|_\Lambda$ then proves (132). □

Thus B_{main} may be any a priori Frobenius-norm envelope for $\mathcal{T}^\star$, while $B_\perp$ may be chosen as a high-probability envelope for the right-hand side of (131).

Lemma S.7.10 (Uniform Phase Confidence)**.** *Let $\widehat{\beta}_k$ be the ridge estimate formed from the executed proposals in a completed phase k, with design matrix V_k. With probability at least $1 - \delta$,*

$$|a_i^\top (\widehat{\beta}_k - \beta)| \le \xi_M(\delta)\, \|a_i\|_{V_k^{-1}} \tag{133}$$

simultaneously for all completed phases with at most M proposals and all $i \in \mathcal{I}$.

Proof. For a phase with proposal index set $\mathcal{D}_k$,

$$\widehat{\beta}_k - \beta = V_k^{-1} \sum_{s \in \mathcal{D}_k} a_{I_s} \xi_s - V_k^{-1} \Lambda \beta.$$

Within a phase, proposal choices depend on the active set and accumulated design, but not on rewards from that phase. Conditional on the phase-start history and its complete proposal-action sequence, V_k and all coefficients below are fixed. For a fixed arm, the normalized noise term is conditionally sub-Gaussian because

$$\sum_{s \in \mathcal{D}_k} \{a_i^\top V_k^{-1} a_{I_s}\}^2 \le a_i^\top V_k^{-1} a_i.$$

A scalar martingale bound and a union bound over actions and phase endpoints give

$$\left| a_i^\top V_k^{-1} \sum_{s \in \mathcal{D}_k} a_{I_s} \xi_s \right| \le \sigma \sqrt{2 \log\left\{ \frac{2|\mathcal{I}|(1 + \log_2 M)}{\delta} \right\}}\, \|a_i\|_{V_k^{-1}}.$$

For the ridge bias, Cauchy–Schwarz gives

$$|a_i^\top V_k^{-1} \Lambda \beta| \le \|a_i\|_{V_k^{-1}}\, \|\beta\|_\Lambda.$$

Finally, $\|\beta\|_\Lambda \le \sqrt{\lambda_1} B_{\mathrm{main}} + \sqrt{\lambda_2} B_\perp$. This proves (133). □

Lemma S.7.11 (Safe Elimination and Phase Regret)**.** *On* (133)*, the optimal action remains in every active set. If phase k is completed and*

$$w_k = \max_{i \in \mathcal{C}_k} \|a_i\|_{V_k^{-1}},$$

then every $i \in \mathcal{C}_{k+1}$ satisfies

$$\Delta_i \le 4\xi_M(\delta) w_k. \tag{134}$$

Moreover, if $m_k = 2^{k-1}$ proposals are executed in phase k,

$$m_k \min\{1, w_k^2\} \le 2\Gamma_M. \tag{135}$$

Proof. Let $i^\star$ be optimal. Its upper confidence bound is at least $\mu_{i^\star}$, while every lower confidence bound is at most its true mean. Thus $i^\star$ is never eliminated.

For $i \in \mathcal{C}_{k+1}$, the elimination rule gives

$$a_i^\top \widehat{\beta}_k + \xi_M u_k(i) \ge a_{i^\star}^\top \widehat{\beta}_k - \xi_M u_k(i^\star).$$

Combining this relation with the confidence bounds for i and $i^\star$ gives

$$\Delta_i \le 2\xi_M \{u_k(i) + u_k(i^\star)\} \le 4\xi_M w_k.$$

Within phase k, the maximum width is nonincreasing. The executed proposal has the largest current width. Therefore

$$m_k \min\{1, w_k^2\} \le \sum_{c=0}^{m_k - 1} \min\{1, \|a_{I_{k,c}}\|_{V_{k,c}^{-1}}^2\}.$$

The elliptical-potential identity bounds the last sum by $2\log\{\det(V_k)/\det(\Lambda)\}$. Fischer's inequality separates the main and tail principal blocks. The arithmetic–geometric mean inequality applied within each block gives

$$\log\{\det(V_k)/\det(\Lambda)\} \le \Gamma_M.$$

This proves (135). □

S.7.2 The Number of Inferential Draws

Lemma S.7.12 (Inferential-Exploration Count)**.** *For every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$N_T^{\mathrm{tar}} \le S_T^\varepsilon + 2\sqrt{S_T^\varepsilon \log(1/\delta)} + \frac{2}{3}\log(1/\delta). \tag{136}$$

Proof. The process

$$M_m = \sum_{t=s_0+1}^{m} (B_t - \varepsilon_t)$$

is a martingale. Its increments are bounded by one, and its predictable quadratic variation is at most $\sum_{t=s_0+1}^{m} \varepsilon_t \le S_T^\varepsilon$. Freedman's inequality gives

$$M_T \le \sqrt{2S_T^\varepsilon \log(1/\delta)} + \frac{1}{3}\log(1/\delta).$$

The displayed bound is a convenient relaxation. □

The deterministic schedule satisfies

$$S_T^\varepsilon = \sum_{t=s_0+1}^{T} \varepsilon_t \le c_\varepsilon \sum_{t=s_0+1}^{T} (t \vee s_0)^{-1/2} \le 2c_\varepsilon \sqrt{T}.$$

S.7.3 Proposal Rounds and Completion of the Proof

Enumerate the executed proposal rounds as $s_1 < \cdots < s_{M_T}$.

Lemma S.7.13 (Proposal-Subsequence Coupling)**.** *Conditional on the warm-up rotation and proposal rewards, the sequence $\{I^{\mathrm{TE}}_{s_m}\}_{m=1}^{M_T}$ is exactly the trajectory of the fixed-rotation elimination algorithm run for M_T rounds with the same regularizer, confidence radius, phase budgets, and tie-breaking rule.*

Proof. Before the first proposal, the two procedures have the same active set, design matrix, phase counter, and within-phase count. A target draw changes none of these variables. Suppose the states agree before proposal m. Both procedures maximize the same width over the same active set and therefore select the same action. They observe the coupled proposal reward and make the same design update. If the phase count reaches its budget, they compute the same ridge estimate and active set. The states agree before proposal $m+1$. Induction proves the claim. □

We now bound the proposal regret without invoking an external regret theorem. The first proposal phase contributes at most $\Delta_{\max}$. For every later phase, Lemma S.7.11 and (135) give

$$\min\{\Delta_{\max}, 4\xi_{\overline{M}}(\delta/2)w_k\} \le 4\{\Delta_{\max} + \xi_{\overline{M}}(\delta/2)\}\min\{1, w_k\},$$

and Cauchy–Schwarz gives

$$m_k \min\{1, w_k\} \le \{m_k^2 \min(1, w_k^2)\}^{1/2} \le \sqrt{2m_k\Gamma_{\overline{M}}}.$$

Since phase $k+1$ has at most $2m_k$ proposals,

$$\sum_{m \text{ in phase } k+1} \Delta_{I^{\mathrm{TE}}_{s_m}} \le C\{\Delta_{\max} + \xi_{\overline{M}}(\delta/2)\}\sqrt{m_k\Gamma_{\overline{M}}}.$$

Because $m_k = 2^{k-1}$,

$$\sum_{k:\, m_k \le M_T} \sqrt{m_k} \le C\sqrt{M_T}.$$

Hence, on the uniform confidence event,

$$\sum_{m=1}^{M_T} \Delta_{I^{\mathrm{TE}}_{s_m}} \le C\{\Delta_{\max} + \xi_{\overline{M}}(\delta/2)\}\sqrt{M_T\Gamma_{\overline{M}}}. \tag{137}$$

The same bound covers a truncated final phase because its actions belong to the active set inherited from the preceding completed phase.

Combining (130), (136), and (137), then adding the warm-up regret, proves (127).

Finally, the choice $\lambda_2 = \overline{M}/\{q\log(1+\overline{M}/\lambda_1)\}$ gives

$$\Gamma_{\overline{M}} \le 2q\log(1+\overline{M}/\lambda_1).$$

Lemma S.7.9 controls the tail term in $\xi_{\overline{M}}$. The warm-up rate alone does not imply that this term is bounded at the scale selected by λ_2; condition (128) is also required. Substitution of that condition, the balanced-dimension warm-up rate, and $q = O(p^{d-1})$ proves (129).

S.8 Supplementary Factor and Information Stability Results

The main paper states results for the reward tensor and a fixed linear target. Two intermediate quantities also matter throughout RITE: the estimated Tucker factor spaces used by spectral rotation, and the inferential information matrix used by the debiasing direction. We record their stability here. These results play the same supporting role as the factor-space and plug-in variance results that accompany online tensor inference, while the conclusions below are tailored to the adaptive RITE trajectory.

Recall

$$\lambda_\star = \min_{j\in[d]} \sigma_{r_j}\{M_j(\mathcal{T}^\star)\} \tag{138}$$

as the smallest nonzero multilinear singular value. Define the batch-k estimation radius

$$\mathfrak{r}_k = C\sigma \left\{ \sqrt{\frac{\mathfrak{d}_T}{N_k^{\text{eff}}}} + \frac{\mathfrak{d}_T}{N_k^{\text{eff}}} \right\} + C\,\mathrm{Rem}_{\text{opt}}(k). \tag{139}$$

Theorem S.8.4 (Uniform Factor- and Tangent-Space Stability)**.** *Suppose the conditions of Theorem S.5.1 hold and* $\max_{k\le K_T} \mathfrak{r}_k \le c\lambda_\star$ *for a sufficiently small constant c. Then, with probability at least* $1 - c_1(pT)^{-c_2}$*, simultaneously for all batches k and modes j,*

$$\left\|\sin\Theta(\widehat{U}_j^{(k)}, U_j^\star)\right\|_{\text{op}} \le C\frac{\mathfrak{r}_k}{\lambda_\star}, \tag{140}$$

$$\left\|P_{\widehat{\mathbb{T}}^{(k)}} - P_{\mathbb{T}^\star}\right\|_{\text{op}} \le C_d\frac{\mathfrak{r}_k}{\lambda_\star}. \tag{141}$$

Consequently, for any fixed tensor H,

$$\left\|(P_{\widehat{\mathbb{T}}^{(k)}} - P_{\mathbb{T}^\star})H\right\|_F \le C_d\frac{\mathfrak{r}_k}{\lambda_\star}\left\|H\right\|_F. \tag{142}$$

The proof is given in Section S.9. At $k = 0$, Equation (140) justifies the fixed warm-up rotation used by the elimination policy. Across later batches, (141) controls the evolving tangent systems used for target inference.

No operator-norm perturbation bound for the full pseudoinverse is required. The leading martingale array and the exact cancellation use the implemented direction $g_{H,k}$ itself. Its norm and target variance are controlled directly by (44); Supplementary Theorem S.8.4 is used only for the tangent-projection remainder. This target-direction formulation remains meaningful when the uniform component of q_H gives vanishing information in directions unrelated to H.

S.9 Proof of Supplementary Factor-Space Stability

Fix a batch k and a mode j. Write

$$M_j(\widehat{\mathcal{T}}_{\text{lr}}^{(k)}) = M_j(\mathcal{T}^\star) + M_j(\widehat{\mathcal{T}}_{\text{lr}}^{(k)} - \mathcal{T}^\star). \tag{143}$$

On the event of Theorem S.5.1, the perturbation satisfies

$$\left\|M_j(\widehat{\mathcal{T}}_{\text{lr}}^{(k)} - \mathcal{T}^\star)\right\|_{\text{op}} \le \left\|\widehat{\mathcal{T}}_{\text{lr}}^{(k)} - \mathcal{T}^\star\right\|_F \le \mathfrak{r}_k. \tag{144}$$

The singular gap between the r_jth and (r_j+1)st singular values of $M_j(\mathcal{T}^\star)$ is at least $\lambda_\star$. Since $\mathfrak{r}_k \le c\lambda_\star$, Wedin's perturbation theorem (Wedin, 1972) gives

$$\left\|\sin\Theta(\widehat{U}_j^{(k)}, U_j^\star)\right\|_{\text{op}} \le \frac{C\left\|M_j(\widehat{\mathcal{T}}_{\text{lr}}^{(k)} - \mathcal{T}^\star)\right\|_{\text{op}}}{\lambda_\star} \le C\frac{\mathfrak{r}_k}{\lambda_\star}. \tag{145}$$

This proves (140) for one (j,k). The estimation event is already uniform over the finitely many modes and the K batches, so no additional probability loss is needed beyond the union bound in (82).

It remains to control the tangent projector. The Tucker tangent space is a smooth function of the factor projectors and the core as long as the multilinear singular values stay separated. Applying the local Lipschitz statement in Lemma S.10.14 and then (145),

$$\left\|P_{\widehat{\mathbb{T}}^{(k)}} - P_{\mathbb{T}^\star}\right\|_{\text{op}} \le C_d\max_{j\in[d]}\left\|\sin\Theta(\widehat{U}_j^{(k)}, U_j^\star)\right\|_{\text{op}} \le C_d\frac{\mathfrak{r}_k}{\lambda_\star}. \tag{146}$$

Applying the projector difference to H gives (142). This completes the proof of Supplementary Theorem S.8.4.

S.10 General Technical Lemmas

We first isolate the geometric and martingale inputs used below. The RITE-specific claims—stopped-batch concentration, capped-design curvature, reference-design concentration, and remainder control—are proved in Section S.11.

Lemma S.10.14 (Local Tucker Manifold Geometry). *Suppose the low-rank regularity condition in the main paper holds and $\mathcal{T}$ lies in a sufficiently small Frobenius neighborhood of $\mathcal{T}^\star$ on $\mathcal{M}_{\mathbf{r}}$. Then the retraction $R_{\mathcal{T}}$ is locally second-order accurate, the tangent projection $P_{T_{\mathcal{T}}\mathcal{M}_{\mathbf{r}}}$ is Lipschitz in $\mathcal{T}$, and the local covering entropy of $\mathcal{M}_{\mathbf{r}}$ is controlled by D_{eff}:*

$$\log N\{\epsilon, \mathcal{M}_{\mathbf{r}} \cap B_F(\mathcal{T}^\star, r), \|\cdot\|_F\} \le C D_{\text{eff}} \log(Cr/\epsilon).$$

The fixed-rank Tucker manifold and its retractions are described by Kressner et al. (2014), within the general framework of Absil et al. (2008). The entropy bound follows from the low-rank tensor net constructions used in tensor estimation and completion (Xia et al., 2021; Ma & Xia, 2024).

Lemma S.10.15 (Factor-Space Perturbation). *If $\mathcal{T} \in \mathcal{M}_{\mathbf{r}}$ satisfies $\|\mathcal{T} - \mathcal{T}^\star\|_F \le c\sigma_{\min}(\mathcal{T}^\star)$, then for each mode j,*

$$\left\|\sin\Theta(\widehat{U}_j, U_j^\star)\right\|_{\text{op}} \le C \frac{\|\mathcal{T} - \mathcal{T}^\star\|_F}{\sigma_{\min}(\mathcal{T}^\star)}.$$

Consequently, the tangent projection error satisfies

$$\|(P_{T_{\mathcal{T}}\mathcal{M}_{\mathbf{r}}} - P_{\mathbb{T}^\star})H\|_F \le C \frac{\|\mathcal{T} - \mathcal{T}^\star\|_F}{\sigma_{\min}(\mathcal{T}^\star)} \|H\|_F$$

for any fixed tensor H.

The mode-wise factor bound is Wedin's singular-subspace perturbation theorem (Wedin, 1972); the tangent-projector statement then follows from local Tucker geometry.

Lemma S.10.16 (Capped-IPW Martingale Concentration). *Let Δ range over a local net of rank-$2\mathbf{r}$ secant directions. Under the two-stage noise condition, on the realized curvature event and with probability at least $1 - (pT)^{-c}$,*

$$\left|\sum_{t \le t_{k+1}-1} \omega_t^{\text{lr}} \xi_t \langle \Delta, \mathcal{A}_t\rangle\right| \le C\sigma \left\{\sqrt{\pi_{\min}^{-1} \underline{N}_k D_{\text{eff}} \log(pT)} + \pi_{\min}^{-1} D_{\text{eff}} \log(pT)\right\} \|\Delta\|_F .$$

The same bound holds uniformly over the local net after a union bound.

The proof uses the reward-half-step exponential supermartingale with the realized coefficient sum (66); it does not replace that sum by an $\mathcal{F}_{t-1}$ conditional variance.

Lemma S.10.17 (Martingale CLT and Self-Normalization). *Let $\{\psi_{k,t}(H) : k \ge k_0, t \in \mathcal{E}_k\}$ be the leading score array in Lemma S.6.6. If there is a deterministic $v_T(H) > 0$ such that $V_T(H)/v_T(H) \to_p 1$ and the conditional Lindeberg condition* (118) *holds, then*

$$\frac{\sum_{k=k_0}^{K_T} \sum_{t \in \mathcal{E}_k} \psi_{k,t}(H)}{v_T(H)^{1/2}} \Rightarrow N(0,1).$$

Moreover, if the residual replacement and nuisance-estimation errors are $o_p\{v_T(H)\}$ after squaring and summing, then $\widehat{V}_T(H)/v_T(H) \to_p 1$.

The deterministic normalization is essential: the triangular-array martingale CLT (Hall & Heyde, 1980) is applied only after the predictable quadratic variation has been shown to converge to one on that scale.

Lemma S.10.18 (Target-Reference Design Concentration). *Fix batch k and condition on $\mathcal{F}_{t_k-1}$. Let $\widehat{J}_{H,k}$ and $J_{H,k}$ be defined in* (97) *and* (43). *If $\left\|s_i^{(k)}\right\| \le 1$, then for every $x > 0$, with conditional probability at least $1 - 2e^{-x}$,*

$$\left\|\widehat{J}_{H,k} - J_{H,k}\right\|_{\text{op}} \le C \left\{\sqrt{\frac{v_{H,k}\{D_{\text{eff}} + x\}}{n_k}} + \frac{\underline{\varepsilon}_k^{-1}\{D_{\text{eff}} + x\}}{n_k}\right\}, \tag{147}$$

where $v_{H,k}$ is the conditional matrix-variance scale of $\rho_t^H s_{I_t}^{(k)} s_{I_t}^{(k)\top}$. The factor $\underline{\varepsilon}_k^{-1}$ follows from $\rho_t^H \le \underline{\varepsilon}_k^{-1}$; no propensity cap is used.

The result is Tropp's matrix Freedman inequality (Tropp, 2011) applied on the target-relevant tangent subspace. Its RITE specialization is proved in Section S.11.8.

S.11 Proofs of RITE-Specific Technical Lemmas

This section contains the detailed proofs deferred from the preceding sections. The proofs repeatedly use the local Tucker geometry and factor perturbation results stated in Section S.10. Those external tools are cited there; the arguments below establish the parts created by the adaptive RITE sampling rule.

S.11.1 Proof of Lemma S.5.2

Fix $\Delta \in \mathcal{K}_{2\mathbf{r}}(r)$ and define

$$X_t(\Delta) = \omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t \rangle^2 - \mathbb{E}\left\{ \omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t \rangle^2 \mid \mathcal{F}_{t-1} \right\}. \tag{148}$$

Because $\mathcal{A}_t$ is one-hot and $\|\Delta\|_F = 1$,

$$|\langle \Delta, \mathcal{A}_t \rangle| \le 1.$$

The capped weight satisfies $\omega_t^{\mathrm{lr}} \le \pi_{\min}^{-1}$, and hence

$$|X_t(\Delta)| \le 2\pi_{\min}^{-1}. \tag{149}$$

Moreover,

$$\begin{aligned} \sum_{t<t_{k+1}} \mathbb{E}\{X_t(\Delta)^2 \mid \mathcal{F}_{t-1}\} &\le \sum_{t<t_{k+1}} \mathbb{E}\left\{ (\omega_t^{\mathrm{lr}})^2 \langle \Delta, \mathcal{A}_t \rangle^4 \mid \mathcal{F}_{t-1} \right\} \\ &\le \pi_{\min}^{-1} \sum_{t<t_{k+1}} \mathbb{E}\left\{ \omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t \rangle^2 \mid \mathcal{F}_{t-1} \right\}. \end{aligned} \tag{150}$$

The last inequality uses $(\omega_t^{\mathrm{lr}})^2 \le \pi_{\min}^{-1} \omega_t^{\mathrm{lr}}$ and $\langle \Delta, \mathcal{A}_t \rangle^4 \le \langle \Delta, \mathcal{A}_t \rangle^2$.

Let

$$\overline{\mathcal{Q}}_k(\Delta) = \sum_{t<t_{k+1}} \mathbb{E}\left\{ \omega_t^{\mathrm{lr}} \langle \Delta, \mathcal{A}_t \rangle^2 \mid \mathcal{F}_{t-1} \right\}.$$

The maximal form of Freedman's inequality (Freedman, 1975) is applied uniformly over deterministic endpoints $m \le T$ and then evaluated at the bounded stopping time $t_{k+1} - 1$. For every $x > 0$,

$$\left| \mathcal{Q}_k(\Delta) - \overline{\mathcal{Q}}_k(\Delta) \right| \le C \left\{ \sqrt{\pi_{\min}^{-1} \overline{\mathcal{Q}}_k(\Delta) x} + \pi_{\min}^{-1} x \right\} \tag{151}$$

with probability at least $1 - 2e^{-x}$.

We now make the bound uniform. Besides $\mathcal{K}_{2\mathbf{r}}(r)$, consider the normalized nonzero sums and differences of two elements of $\mathcal{K}_{2\mathbf{r}}(r)$. These tensors have Tucker ranks bounded by a constant multiple of $\mathbf{r}$. Lemma S.10.14 therefore gives a common ϵ-net $\mathcal{N}_\epsilon$ for these three classes such that

$$\log |\mathcal{N}_\epsilon| \le C D_{\mathrm{eff}} \log(C/\epsilon). \tag{152}$$

Take $\epsilon = (pT)^{-3}$ and

$$x = C_0 D_{\mathrm{eff}} \log(pT)$$

for a sufficiently large C_0. A union bound in (151) gives the claimed deviation simultaneously on the net. Under (41), the right-hand side is at most a fixed fraction of $\underline{N}_k$.

For a general Δ in either the secant or envelope class, choose a net point Δ° with $\|\Delta - \Delta^\circ\|_F \le \epsilon$. The polarization identity gives

$$|\mathcal{Q}_k(\Delta) - \mathcal{Q}_k(\Delta^\circ)| \le \left\{ \mathcal{Q}_k(\Delta - \Delta^\circ) \right\}^{1/2} \left\{ \mathcal{Q}_k(\Delta + \Delta^\circ) \right\}^{1/2}. \tag{153}$$

Since there are at most T summands and $\omega_t^{\mathrm{lr}} \le \pi_{\min}^{-1}$,

$$\mathcal{Q}_k(\Delta - \Delta^\circ) \le T\pi_{\min}^{-1}\epsilon^2, \qquad \mathcal{Q}_k(\Delta + \Delta^\circ) \le 4T\pi_{\min}^{-1}.$$

The right-hand side of (153) is therefore at most $2T\pi_{\min}^{-1}\epsilon$. The same deterministic bounds apply to $\overline{\mathcal{Q}}_k$. With $\epsilon = (pT)^{-3}$, this term is $o(\underline{N}_k)$ under (41). Hence

$$c\underline{N}_k \le \inf_{\Delta \in \mathcal{K}_{2\mathbf{r}}(r)} \mathcal{Q}_k(\Delta) \le \sup_{\Delta \in \mathcal{K}_{2\mathbf{r}}(r)} \mathcal{Q}_k(\Delta) \le C\underline{N}_k.$$

Every difference of two feasible tensors in the local basin is a scalar multiple of an element of $\mathcal{K}_{2\mathbf{r}}(r)$. Unit tangent directions are limits of such secants, so continuity extends the same bounds to the unit sphere of $\mathbb{T}^\star$. Taking the empirical infimum over that sphere yields $N_k^{\mathrm{eff}} \asymp \underline{N}_k$. This proves (42) and (74) without a tangent-to-secant assumption.

S.11.2 Proof of Lemma S.5.3

For a fixed Δ, set

$$Z_t(\Delta) = \omega_t^{\mathrm{lr}} \xi_t \langle \Delta, \mathcal{A}_t \rangle.$$

Conditional on $\mathcal{F}_{t-1/2}$, the multiplier $\omega_t^{\mathrm{lr}}\langle \Delta, \mathcal{A}_t\rangle$ is fixed. The conditional sub-Gaussian assumption gives

$$\mathbb{E}\left[\exp\{\lambda Z_t(\Delta)\} \mid \mathcal{F}_{t-1/2}\right] \le \exp\left\{\frac{\lambda^2\sigma^2}{2}(\omega_t^{\mathrm{lr}})^2\langle \Delta, \mathcal{A}_t\rangle^2\right\}. \tag{154}$$

For a deterministic endpoint m, set

$$S_m(\Delta) = \sum_{t \le m} Z_t(\Delta), \qquad W_m(\Delta) = \sum_{t \le m} (\omega_t^{\mathrm{lr}})^2 \langle \Delta, \mathcal{A}_t\rangle^2.$$

Equation (154) implies that

$$\exp\left\{\lambda S_m(\Delta) - \frac{\lambda^2\sigma^2}{2} W_m(\Delta)\right\}$$

is a nonnegative supermartingale on the interleaved action–reward filtration. Ville's inequality (Ville, 1939) followed by dyadic peeling over $W_m(\Delta)$ gives, simultaneously over $m \le T$,

$$|S_m(\Delta)| \le C\sigma\left\{\sqrt{W_m(\Delta)x} + \pi_{\min}^{-1}x\right\} \tag{155}$$

with probability at least $1 - 2e^{-x}$. The harmless linear term makes the bound uniform over the smallest variance slices. Evaluate this inequality at the pre-reward stopping time $t_{k+1} - 1$. On (74),

$$W_{t_{k+1}-1}(\Delta) \le \pi_{\min}^{-1}\mathcal{Q}_k(\Delta) \le C\pi_{\min}^{-1}\underline{N}_k$$

for every unit $\Delta \in \mathcal{K}_{\mathrm{env}}(r)$.

Apply (155) to the common net $\mathcal{N}_\epsilon$ used in (152), with $x = C_0 D_{\mathrm{eff}} \log(pT)$. A union bound therefore gives

$$|\mathcal{M}_k(\Delta^\circ)| \le C\sigma\left\{\sqrt{\pi_{\min}^{-1}\underline{N}_k D_{\mathrm{eff}} \log(pT)} + \pi_{\min}^{-1} D_{\mathrm{eff}} \log(pT)\right\}$$

for every $\Delta^\circ \in \mathcal{N}_\epsilon$.

For a secant direction and its nearest net point, linearity and the one-hot design give

$$|\mathcal{M}_k(\Delta) - \mathcal{M}_k(\Delta^\circ)| \le T\pi_{\min}^{-1}\epsilon \max_{t \le T} |\xi_t|.$$

The conditional sub-Gaussian assumption and a union bound imply

$$\max_{t \le T} |\xi_t| = O_p\{\sigma\sqrt{\log(pT)}\}.$$

With $\epsilon = (pT)^{-3}$, this approximation term is smaller than the displayed net bound. Homogeneity yields (75).

S.11.3 Proof of Lemma S.5.4

Apply (79) to $\mathcal{T} = \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)}$ and use (70):

$$cN_k^{\mathrm{eff}} \left\| \widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)} - \widehat{\mathcal{T}}_k^{\circ} \right\|_F^2 \le L_k(\widehat{\mathcal{T}}_{\mathrm{lr}}^{(k)}) - L_k(\widehat{\mathcal{T}}_k^{\circ}) \le \epsilon_k^{\mathrm{opt}}.$$

Taking square roots proves (80).

This transfer is conditional on the achieved objective tolerance. Proving that a specified Riemannian solver reaches $\epsilon_k^{\mathrm{opt}}$ after a given J_k would require solver-specific smoothness, stepsize, and gradient-dominance assumptions; no such finite-iteration claim is used here.

S.11.4 Proof of Lemma S.6.5

Fix batch k and work on the factor-stability event. Local Tucker geometry gives

$$\|(I - P_{k-1})\Delta_{k-1}\|_F \le C \left(r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star} \right). \tag{156}$$

Applying Cauchy–Schwarz to (94) immediately yields (106).

For the design term, (85) implies

$$\|g_{H,k}\| \le C_{g,H}.$$

Moreover, $\|\delta_{k-1}\| \le \|\Delta_{k-1}\|_F = r_{k-1}$. Hence

$$\begin{aligned} |R_{\mathrm{design},k}(H)| &\le \|g_{H,k}\| \left\| J_{H,k} - \widehat{J}_{H,k} \right\|_{\mathrm{op}} \|\delta_{k-1}\| \\ &\le C_{g,H}\zeta_{H,k} r_{k-1}, \end{aligned} \tag{157}$$

which proves (107).

For the orthogonal term, a one-hot tensor has unit Frobenius norm, so

$$|b_{I_t,k}| \le \|(I - P_{k-1})\Delta_{k-1}\|_F .$$

Using this bound in (102), followed by (156), gives

$$\begin{aligned} |R_{\mathrm{orth},k}(H)| &\le \|g_{H,k}\| \left\{ \frac{1}{n_k} \sum_{t \in \mathcal{E}_k} \rho_t^H \left\| s_{I_t}^{(k)} \right\| \right\} \|(I - P_{k-1})\Delta_{k-1}\|_F \\ &\le C\, C_{g,H} \overline{\rho}_{H,k} \left(r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star} \right). \end{aligned} \tag{158}$$

This is (108).

Finally, Lemma S.10.18, applied conditionally on $\mathcal{F}_{t_{k-1}}$, gives (109). For the scalar average, let $W_t = \rho_t^H \left\| s_{I_t}^{(k)} \right\|$. The transport identity gives

$$\mathbb{E}(W_t \mid \mathcal{F}_{t-1}) = \sum_{i \in \mathcal{I}} q_H(i) \left\| s_i^{(k)} \right\| \le 1.$$

Moreover $0 \le W_t \le \underline{\varepsilon}_k^{-1}$ and its conditional variance is at most $\underline{\varepsilon}_k^{-1}$. Scalar Freedman applied up to the batch endpoint proves (110). These concentration steps concern only realized batch averages; the direction $g_{H,k}$ remains fixed.

S.11.5 Proof of Lemma S.6.6

Equation (103) is an exact algebraic identity. Lemma S.6.5 bounds each of its three remainder components. Therefore

$$\begin{aligned} |\operatorname{Rem}_{\mathrm{db}}(H)| &\le C \|H\|_F \sum_{k=k_0}^{K_T} \alpha_k \left[r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star} \right] \\ &\quad + C\, C_{g,H} \sum_{k=k_0}^{K_T} \alpha_k \left[\zeta_{H,k} r_{k-1} + \overline{\rho}_{H,k} \left(r_{k-1}\eta_{k-1} + \frac{r_{k-1}^2}{\lambda_\star} \right) \right]. \end{aligned} \tag{159}$$

Under (111), the right-hand side of (159) is $o_p\{v_T(H)^{1/2}\}$. Substitution into (103) proves (112).

S.11.6 Proof of Lemma S.6.7

Reveal the action before the reward. For $t \in \mathcal{E}_k$, define

$$c_{k,t}(H) = \frac{1}{N_T^{\inf}} \rho_t^H g_{H,k}^\top s_{I_t}^{(k)}.$$

At the beginning of batch k, both $g_{H,k}$ and the entire collection $\{s_i^{(k)} : i \in \mathcal{I}\}$ are $\mathcal{F}_{t_k-1}$-measurable. After drawing I_t, the coefficient $c_{k,t}(H)$ is $\mathcal{F}_{t-1/2}$-measurable. Hence

$$\mathbb{E}\{c_{k,t}(H)\xi_t \mid \mathcal{F}_{t-1/2}\} = 0.$$

Thus the ordered array $\{\psi_{k,t}(H)\}$ is a martingale-difference array. Flatten the pairs (k,t) in chronological order and set

$$X_{T,j} = \psi_{k,t}(H)/v_T(H)^{1/2}.$$

Its conditional variance sum is

$$\sum_j \mathbb{E}(X_{T,j}^2 \mid \mathcal{G}_{T,j-1}) = \frac{V_T(H)}{v_T(H)} \to_p 1$$

by (88), where $\mathcal{G}_{T,j-1}$ is the corresponding pre-round filtration. Equation (118) is exactly the conditional Lindeberg condition for $\{X_{T,j}\}$. The triangular-array martingale CLT therefore gives (119).

We also verify that the stated leverage condition is sufficient for Lindeberg. By (116), $|c_{k,t}(H)| \le \mathcal{L}_T(H) v_T(H)^{1/2}$ for every realized action. The conditional fourth-moment bound and conditional Markov inequality give

$$\begin{aligned}
&\frac{1}{v_T(H)} \sum_{k,t} \mathbb{E}\left[\psi_{k,t}(H)^2 \mathbf{1}\{|\psi_{k,t}(H)| > \epsilon v_T(H)^{1/2}\} \mid \mathcal{F}_{t-1}\right] \\
&\quad \le \frac{C}{\epsilon^2 v_T(H)^2} \sum_{k,t} \mathbb{E}\{c_{k,t}(H)^4 \mid \mathcal{F}_{t-1}\} \\
&\quad \le \frac{C\mathcal{L}_T(H)^2}{v_T(H)} \sum_{k,t} \mathbb{E}\{c_{k,t}(H)^2 \mid \mathcal{F}_{t-1}\} \\
&\quad \le C\mathcal{L}_T(H)^2 \frac{V_T(H)}{v_T(H)} \to_p 0.
\end{aligned}$$

The last inequality uses $V_T(H) \ge \nu_-^2 \sum_{k,t} \mathbb{E}\{c_{k,t}(H)^2 \mid \mathcal{F}_{t-1}\}$. Finally, (88) and Slutsky's theorem allow $v_T(H)^{1/2}$ to be replaced by $V_T(H)^{1/2}$.

S.11.7 Proof of Lemma S.6.8

Use the notation $c_{k,t}(H)$ and ν_t^2 above, and let $d_{t,k} = \langle \Delta_{k-1}, \mathcal{A}_t \rangle$. Expanding $(\xi_t - d_{t,k})^2$ yields the exact decomposition (121).

For $\mathcal{V}_{0T}(H)$, the summands

$$c_{k,t}(H)^2 \nu_t^2 - \mathbb{E}\{c_{k,t}(H)^2 \nu_t^2 \mid \mathcal{F}_{t-1}\}$$

form a martingale-difference array on the action-revelation half-steps. Since $c_{k,t}(H)^2 \le \mathcal{L}_T(H)^2 v_T(H)$, their conditional variance sum is bounded by

$$C\mathcal{L}_T(H)^2 v_T(H) \sum_{k,t} \mathbb{E}\{c_{k,t}(H)^2 \mid \mathcal{F}_{t-1}\} = o_p\{v_T(H)^2\}.$$

Martingale Chebyshev therefore gives

$$\mathcal{V}_{0T}(H) = o_p\{v_T(H)\}. \tag{160}$$

For $\mathcal{V}_{1T}(H)$, the summands

$$c_{k,t}(H)^2 \left[\xi_t^2 - \nu_t^2\right]$$

form a martingale-difference array on the reward-revelation half-steps. The same leverage bound and the conditional fourth-moment bound make their conditional variance sum $o_p\{v_T(H)^2\}$. Hence

$$\mathcal{V}_{1T}(H) = o_p\{v_T(H)\}. \tag{161}$$

For the cross term, conditional Cauchy–Schwarz gives

$$|\mathcal{V}_{2T}(H)| \le 2\left\{\sum_{k,t} c_{k,t}(H)^2 \xi_t^2\right\}^{1/2} \left\{\sum_{k,t} c_{k,t}(H)^2 d_{t,k}^2\right\}^{1/2}. \tag{162}$$

The first factor is $O_p\{v_T(H)^{1/2}\}$ because $\sum_{k,t} c_{k,t}(H)^2\xi_t^2 = V_T(H) + \mathcal{V}_{0T}(H) + \mathcal{V}_{1T}(H)$. Since $|d_{t,k}| \le r_{k-1}$ for one-hot actions, the second squared factor is bounded by

$$\sum_{k,t} c_{k,t}(H)^2 r_{k-1}^2 = o_p\{v_T(H)\}$$

under (122). Consequently, $\mathcal{V}_{2T}(H) = o_p\{v_T(H)\}$. The same condition directly gives

$$\mathcal{V}_{3T}(H) \le \sum_{k,t} c_{k,t}(H)^2 r_{k-1}^2 = o_p\{v_T(H)\}.$$

Combining the four terms gives $\{\widehat{V}_T(H) - V_T(H)\}/v_T(H) \to_p 0$. Together with (88), this proves both claims in (123).

S.11.8 Proof of Lemma S.10.18

Fix batch k and condition on $\mathcal{F}_{t_k-1}$. For $t \ge t_k$, while batch k is active, define

$$Z_t = \rho_t^H s_{I_t}^{(k)} s_{I_t}^{(k)\top} - J_{H,k}.$$

Equation (98) gives $\mathbb{E}(Z_t \mid \mathcal{F}_{t-1}) = 0$. Tangent projections are contractions, so

$$\|Z_t\|_{\mathrm{op}} \le 1 + \varepsilon_t^{-1}.$$

Let

$$\widetilde{v}_{H,k}(m) = \left\| \sum_{t=t_k}^{m} \mathbb{E}(Z_t^2 \mid \mathcal{F}_{t-1}) \right\|_{\mathrm{op}}.$$

Apply matrix Freedman's inequality uniformly over $m = t_k, \dots, T$. A dyadic peeling argument for the deterministic bounds $1+\varepsilon_t^{-1}$ gives, with conditional probability at least $1-2e^{-x}$, simultaneously over these endpoints,

$$\left\| \sum_{t=t_k}^{m} Z_t \right\|_{\mathrm{op}} \le C\left[\sqrt{\widetilde{v}_{H,k}(m)\{D_{\mathrm{eff}} + x\}} + \varepsilon_m^{-1}\{D_{\mathrm{eff}} + x\}\right].$$

The endpoint $t_{k+1} - 1$ is a bounded pre-reward stopping time. Substitution of this endpoint and division by the realized n_k yield

$$\left\| \widehat{J}_{H,k} - J_{H,k} \right\|_{\mathrm{op}} \le C\left[\sqrt{\frac{v_{H,k}\{D_{\mathrm{eff}} + x\}}{n_k}} + \frac{\varepsilon_k^{-1}\{D_{\mathrm{eff}} + x\}}{n_k}\right],$$

where $v_{H,k} = \widetilde{v}_{H,k}(t_{k+1} - 1)/n_k$. This is (147).