Representation-Guided Preference Learning for Multi-Episode Group Decision-Making

Anonymous authors
Paper under double-blind review

Abstract

We study representation-guided group decision-making with a fixed pool of participants over a sequence of episodes. During each episode, a group coordinator observes scalar utility feedback from each participant in response to text descriptions of options, and seeks to find a fair solution to the posed decision problem. We formalize the setting as a bilinear low-rank bandit and apply an approach of alternating minimization with LCB-minimax exploration. We prove a cumulative pseudoregret bound that scales with the spectral dimension of the representation k_h, rather than the ambient encoder dimension d_h. On a controlled multi-episode simulator, the algorithm reduces the total episode-best-plan gap G_{total} to 22.06 ± 1.47 vs 46.57–146; per-episode error also drops sharply: from 5.57 in episode 1 to 2.00 in episode 3, and thereafter remains roughly in the $1.5 - 3$ range. The same reductions carry over to real pretrained encoders (Qwen3-8B, Llama-3.1-8B-Instruct, gemma-3-4b-it, and all-MiniLM-L6-v2), where jointly learning the shared probe and per-user preferences reduces G_{total} by 1.9–2.5× versus random projections and 3.2–4.9× versus one-hot features.

1 Introduction

Collective decision-making with heterogeneous preferences is central to economics, multi-agent systems, and AI. Participants in such settings often hold private and conflicting utilities over combinatorial decision spaces, making it challenging to achieve fair outcomes that balance agreement with the diversity of perspectives. Formal models in social choice and negotiation provide principled aggregation rules, but often assume structured preference reports or tractable decision spaces.

Recent advances in large language models (LLMs) offer new opportunities to support group decision-making. LLMs can act as mediators that synthesize information, interpret feedback, and coordinate distributed reasoning (Rahwan et al., 2019). However, applying these models to consensus problems introduces new challenges: decision spaces are discrete and high-dimensional, optimization landscapes are non-convex, and participant preferences are latent, sparse, and noisy.

We formalize this setting as a minimax divergence problem, where the goal is to minimize the worst-case dissatisfaction across participants. The objective is non-smooth, the decision space is combinatorial, and the mediator must infer useful structure from limited interaction. Even when cast as an integer program, the problem is computationally difficult: worst-case minimax combinatorial optimization over a decision space of size $\prod_j K_j$ is NP-hard, and inferring the latent preferences ϕ_i from scalar utility feedback in the presence of an unknown probe W is a non-convex bilinear estimation problem (Jun et al., 2019; Jain et al., 2013).

We model this as a multi-episode coordination protocol: the same participants encounter different sets of options in each episode, but their latent preferences ϕ_i persist. The mediator observes only text descriptions of the current episode's options (from which a pretrained encoder f_θ produces hidden-state features $h(o) = f_\theta(o)$) and scalar feedback on plans played within an episode. Real-world examples include a housemate group choosing rules, a team choosing recurring processes, and a company board choosing policies. We show that this setting admits a clean bilinear low-rank bandit formulation, in which the encoder's role is to

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.

provide a structured feature basis whose second-moment matrix Σ_h has spectral dimension k_h (a count of Σ_h's significant eigenvalues) that controls the representation-dependent component of the interaction/regret complexity. The coordinator learns a linear probe W and the per-user preferences $\{\phi_i\}$ *jointly from scalar feedback* across episodes via alternating minimization.

We make the following contributions.

• *Multi-episode formulation.* We formalize representation-guided decision-making as a sequential-episode protocol with persistent per-user preferences $\phi_i \in \mathbb{R}^r$ (latent rank r), fresh episode-specific options $\{o_{jk}^{(e)}\}$ (episode e, decision j, option k), and scalar utility feedback $U_i^{(e)}(x)$ (utility of user i in episode e under plan x, where a plan selects one option per decision) across all rounds and episodes. Utilities factor as $U_i^{(e)}(x) = \phi_i^\top W^\star H^{(e)}(x)$, where $H^{(e)}(x) \in \mathbb{R}^{d_h}$ is the sum of pretrained encoder outputs $f_\theta(o_{jk}^{(e)})$ over the chosen options and $W^\star \in \mathbb{R}^{r \times d_h}$ is an unknown probe. The mediator's goal is to jointly learn $\{\phi_i\}$ and $W^\star$ from scalar feedback across episodes while the encoder's role is to provide a low-effective-rank feature basis.

• *Joint-learning coordinator.* We design a coordinator that maintains a persistent per-user Gaussian posterior on ϕ_i across episodes, refits W periodically by SVD-projected ridge on the accumulated data buffer, and selects plans by LCB-minimax with progressive-widening coord-descent. Under five conditions (A1)–(A5) (Appendix B.1), we prove (Theorem 1) that the theoretical variant $\mathcal{A}^{\mathrm{T}}$ (an exact-LCB idealization of Algorithm 1, defined in Appendix B.2) attains cumulative pseudoregret $\tilde{O}\big(\sigma_{\mathrm{obs}} H_{\mathrm{plan}} \sqrt{r\, ET \log N / \gamma_h^{\mathrm{play}}} + S H_{\mathrm{plan}}^2 \sigma_{\mathrm{obs}} \sqrt{r\, k_h\, ET/(N \gamma_\phi (\gamma_h^{\mathrm{play}})^2)}\big) + \tilde{O}\big[\varepsilon_{\mathrm{eff}} ET (H_{\mathrm{plan}}/\sqrt{\gamma_h^{\mathrm{play}}} + S H_{\mathrm{plan}}^2/(\gamma_h^{\mathrm{play}} \sqrt{\gamma_\phi}))\big] + R_{\mathrm{lo}}$ over E episodes of T rounds each with N participants, where S bounds $\|\phi_i\|$, H_{plan} bounds the per-plan feature norm, σ_{obs} is the observation-noise scale, $\gamma_h^{\mathrm{play}}, \gamma_\phi$ are the played plan-feature and participant excitation constants of (A3), $\varepsilon_{\mathrm{eff}}$ is the effective misspecification level (A1) plus the (A5) alignment tail, and R_{lo} collects lower-order ($\log ET$, $\sqrt{ET}$, and burn-in) terms. The proof combines per-user ridge concentration (Abbasi-Yadkori et al., 2011), low-rank matrix regression (Jun et al., 2019; Kang et al., 2022; Jang et al., 2024), and geometric-contraction alternating minimization (Jain et al., 2013; Yang et al., 2021). Algorithm 1 (practical variant) differs from $\mathcal{A}^{\mathrm{T}}$ in four ways: approximate plan optimization by coord-descent + progressive widening in place of exact-LCB, active-set evaluation of the individual-optimum lower bound, use of a practical confidence radius in place of the full theoretical radius, and a full $r \times d_h$ ridge refit in place of the top-k_h-projected refit that Theorem 1 analyzes.

• *Empirical validation.* On a controlled multi-episode benchmark, joint learning achieves total episode-best-plan divergence gap $G_{\mathrm{total}} := \sum_e G_e$ with $G_e := \min_t \Phi_e(x_t^{(e)}) - \Phi_e^\star$ (per-episode gap between the coordinator's best plan and the offline minimax optimum) of 22.06 ± 1.47 vs 46.57 ± 3.47 for a fixed-true-$W^\star$ diagnostic, 49.19 ± 4.14 for frozen pca-W, 100.01 ± 9.51 for random-W, and 145.56 ± 7.35 for one-hot options. Per-episode G_e drops from 5.57 in episode 1 to 2.00 in episode 3, and thereafter remains roughly in the $1.5 - 3$ range, while the one-hot baseline stays flat at 16–19. On a 720-option corpus with an encoder-independent utility target $U_i(o) = \phi_i^\top \psi(o)$ (latent ψ from the corpus generator) verified by identical one-hot performance across rows (Section 4, Tables 2–3), joint learning reduces G_{total} by 1.9–2.5× vs random-W and 3.2–4.9× vs one-hot across the four encoders (Qwen3-8B, Llama-3.1-8B-Instruct, gemma-3-4b-it, all-MiniLM-L6-v2), and gets within 1.44–1.75× of the supervised-W reference.

2 Related Work

A large literature on social choice (Arrow, 1951; Brandt et al., 2016; Boutilier et al., 2015), negotiation (Kraus, 2001; Fatima et al., 2014; Baarslag et al., 2017), and multi-agent systems (Wooldridge, 2009; Albrecht & Stone, 2018) studies collective decisions under conflicting

preferences. Classical methods rely on predefined aggregation rules and structured inputs (Brandt et al., 2016), while our setting studies interactive AI mediation with pretrained-representation-derived option features and sparse, noisy, or natural-language feedback.

A growing line of work positions language models as mediators or coordinators in multi-party settings, building on the broader machine-behavior agenda of Rahwan et al. (2019) and on emerging evidence that LLMs exhibit theory-of-mind-relevant capabilities (Kosinski, 2023). Notably, the Habermas Machine (Tessler et al., 2024) demonstrates human-subject-scale LLM-mediated common-ground generation across thousands of participants. Most empirical studies in this line focus on whether an LLM can produce a plausible compromise, summarize a discussion, or simulate participants. Our contribution is orthogonal: rather than asking what an LLM can articulate, we ask what its hidden representations expose, and we tie coordination tractability to the geometric question of whether those representations linearly align with a low-rank preference embedding.

Computational theory of mind formalizes inference over agents' mental states (Premack & Woodruff, 1978), and Machine Theory of Mind (Rabinowitz et al., 2018) learns to predict behavior from observations. Our framework adopts a per-participant Bayesian view: a pretrained encoder provides a shared representation of option semantics, while participant-specific latent preferences are inferred from interaction histories via a Gaussian posterior.

Bradley–Terry, Plackett–Luce, ideal-point, and matrix-factorization models treat noisy observations, whether ordinal comparisons (Bradley & Terry, 1952; Plackett, 1975; Luce, 1959) or cardinal ratings (Coombs, 1950; Koren et al., 2009), as reflections of low-dimensional latent utilities. Our setting differs in two respects: the low-rank subspace lives inside a pretrained encoder's second-moment matrix Σ_h rather than being estimated from a user-option rating matrix directly, and options never repeat across episodes, so identifiability rests on transfer through W rather than on repeated observation of the same options.

GP-UCB (Srinivas et al., 2010) and linear-UCB (Abbasi-Yadkori et al., 2011) establish no-regret guarantees for reward maximization by upper-confidence-bound (UCB) exploration. Our LCB-minimax rule adapts this template to a divergence-minimization objective with N participant-specific targets in a low-rank latent space: because we minimize divergence rather than maximize reward, optimism corresponds to a *lower* confidence bound (LCB) rather than an upper one. Bilinear bandits (Jun et al., 2019; Kang et al., 2022; Jang et al., 2024) and low-rank matrix regression provide the sample-complexity template underlying our probe-recovery analysis. Cross-user representation learning has been studied in meta-linear-bandit settings (Cella et al., 2020; Yang et al., 2021); our contribution is to couple probe estimation and per-user preference estimation through joint alternating minimization (Jain et al., 2013) in a multi-episode protocol.

3 Methods: Multi-Episode Joint-Learning Coordinator

3.1 Setting

We have N participants making decisions over E episodes, each episode e presenting $J^{(e)}$ decisions with $K_j^{(e)}$ options each; a plan $x^{(e)}$ is a per-decision one-hot selection over $\mathcal{K}_j$, feasible set $\mathcal{X}^{(e)}$. Each participant has a persistent latent preference $\phi_i \in \mathbb{R}^r$ (same across episodes). Utilities are bilinear:

$$U_i^{(e)}(x) = \phi_i^\top W^\star H^{(e)}(x) + \delta_i^{(e)}(x), \qquad H^{(e)}(x) = \sum_{jk} f_\theta(o_{jk}^{(e)})\, x_{jk}, \tag{1}$$

where f_θ is a pretrained encoder, $W^\star \in \mathbb{R}^{r \times d_h}$ is an unknown probe, and $|\delta_i^{(e)}(x)| \le \varepsilon$ is a bilinear-model misspecification bounded by ε (assumption (A1)). The mediator observes noisy scalar feedback $\tilde{U}_{i,t}^{(e)} = U_i^{(e)}(x_t) + \eta_{i,t}$ with $\eta_{i,t}$ sub-Gaussian at scale σ_{obs}. Let $D_i(x) = U_i^{\max} - U_i(x)$ denote user i's divergence and

$$\Phi_e(x) := \max_i D_i(x), \qquad \Phi_e^\star := \min_{x \in \mathcal{X}^{(e)}} \Phi_e(x) \tag{2}$$

the worst-participant divergence. The persistent per-user state is the natural-parameter Gaussian posterior (P_i, h_i) on ϕ_i, initialized at $e = 1$ from prior $\mathcal{N}(0, \lambda^{-1} I_r)$ and updated on every observation from every episode. This protocol enables three properties: *(P1) warm start*: later episodes inherit precision on ϕ_i from prior ones; *(P2) cross-episode preference transfer*: shared ϕ_i lets any encoder that reduces ET-observation cost beat one-hot; *(P3) feedback-only probe learning*: W is learnable jointly with $\{\phi_i\}$ from scalar feedback.

3.2 Coordinator

The coordinator selects plans by *LCB-minimax*: pick $x_t \in \arg\min_x \max_i \mathrm{LCB}_t(D_i)(x)$, the plan whose worst-user divergence, under the most optimistic (lowest) interpretation of the posterior, is smallest. This is the divergence-minimization analog of UCB: instead of "the arm whose reward could plausibly be highest," we pick "the plan whose worst-participant divergence could plausibly be smallest." We use a *separate-bound* form: an LCB on each user's individual optimum $U_i^{\max}$ (the best-case-for-that-user reference point) and a UCB on the plan's utility $U_i(x)$, then differ them. The separate-bound construction remains valid despite the inner maximization defining $U_i^{\max}$.

Algorithm 1 implements this rule via three interleaved components. *(i) Per-user posterior.* Each participant i maintains a natural-parameter Gaussian posterior (P_i, h_i) on ϕ_i that is updated after every observation from every episode. *(ii) Plan selection.* At each round the coordinator forms projected option features $\hat{\psi}_{jk}^{(e)} = \hat{W}\, h(o_{jk}^{(e)})$, restricts attention to a progressively widening top-$m(t)$ active-decision set $\mathcal{A}_t$ (decisions where the option features spread most), and runs coord-descent to minimize the separate-bound LCB-minimax score $S(c)$ over the active decisions. A short uniform-random cold-start phase ($t \le T_0$) ensures initial excitation before LCB takes over. *(iii) Cross-episode probe refit.* Every K_{refit} rounds, the coordinator solves an SVD-projected ridge for $\hat{W}$ over the pooled data buffer $\mathcal{D} = \{(i, e, \tau, H^{(e)}(x_t^{(e)}), \tilde{U}_i)\}$, then rebuilds each per-user posterior (P_i, h_i) from $\mathcal{D}$ under the new $\hat{W}$. Thus, both $\hat{W}$ and $\{\hat{\phi}_i\}$ are estimated from the same scalar-feedback buffer via alternating minimization.

The time complexity per episode round is $O(Nr^2 + rd_h\,|x|)$ for posterior updates and feature computation, and one coord-descent sweep over $|\mathcal{A}_t|$ decisions. The refit cost per K_{refit}-boundary is $O(|\mathcal{D}|\, rd_h + rd_h^2)$ for SVD-projected ridge and the buffer replay. Encoder forward passes cost $\sum_j K_j^{(e)}$ once per episode.

Theorem 1 below bounds the cumulative minimax-divergence pseudoregret of the theoretical variant $\mathcal{A}^{\mathrm{T}}$ of Algorithm 1 (both $\mathcal{A}^{\mathrm{T}}$ and proof appear in Appendix B.2).

Theorem 1. *Consider the exact-LCB coordinator (Appendix B.2) run on the bilinear low-rank preference model for E episodes of T rounds each. Assume the five conditions (A1)–(A5) of Appendix B.1: approximate bilinear utility with sub-Gaussian noise (A1); integer spectral dimension k_h of Σ_h (A2); an excitation pair $(\gamma_h^{\mathrm{play}}, \gamma_\phi)$ on the played design (A3); warm-start alignment plus per-refit alt-min contraction (A4); and top-k_h alignment of $W^\star$ with residual ε_h (A5). Then Equation* (3) *holds with probability $\ge 1 - \delta$ for all $E, T \ge 1$, with $\varepsilon_{\mathrm{eff}} := \varepsilon + c_{\mathrm{align}} S H_{\mathrm{plan}} \varepsilon_h$.*

$$\begin{aligned}\sum_{e,t} \big(\Phi_e(x_t^{(e)}) - \Phi_e^\star\big) \;\le\; & \tilde{O}\left(\sigma_{\mathrm{obs}} H_{\mathrm{plan}} \sqrt{\frac{rET \log N}{\gamma_h^{\mathrm{play}}}} + S H_{\mathrm{plan}}^2 \sigma_{\mathrm{obs}} \sqrt{\frac{r k_h ET}{N \gamma_\phi (\gamma_h^{\mathrm{play}})^2}}\right) \\ & + \tilde{O}\left[\varepsilon_{\mathrm{eff}}\, ET \left(\frac{H_{\mathrm{plan}}}{\sqrt{\gamma_h^{\mathrm{play}}}} + \frac{S H_{\mathrm{plan}}^2}{\gamma_h^{\mathrm{play}} \sqrt{\gamma_\phi}}\right)\right] + R_{\mathrm{lo}},\end{aligned} \tag{3}$$

where

$$R_{\mathrm{lo}} := \tilde{O}\left(\frac{\sigma_{\mathrm{obs}}^2 H_{\mathrm{plan}}^2 k_h \sqrt{r}}{(\gamma_h^{\mathrm{play}})^{3/2} \sqrt{N \gamma_\phi}}\right) + \tilde{O}\left(\frac{\varepsilon_{\mathrm{eff}}\, \sigma_{\mathrm{obs}} H_{\mathrm{plan}}^2 \sqrt{k_h ET}}{\gamma_h^{\mathrm{play}} \sqrt{\gamma_\phi \gamma_h^{\mathrm{play}}}}\right) + O\big(S H_{\mathrm{plan}} (T_0 + t_1)\big)$$

collects lower-order terms ($\log ET$, $\sqrt{ET}$, and a burn-in constant up to the basin-entry round t_1 of (A4)(ii)).

Algorithm 1 Multi-episode joint-learning coordinator

Inputs: N users; latent rank r; pretrained encoder f_θ with representation dimension d_h; ridge $\lambda > 0$; observation noise $\sigma_{\text{obs}} > 0$; exploration schedule $\{\beta_t\}$; progressive-widening schedule $m(t)$; W-refit period $K_{\text{refit}} \in \mathbb{N}$; cold-start horizon $T_0 \in \mathbb{N}$; W-ridge $\lambda_W > 0$.
Initialize: $P_i \leftarrow \lambda I_r$, $h_i \leftarrow \mathbf{0}$ for $i = 1, \dots, N$; data buffer $\mathcal{D} \leftarrow \emptyset$; global round counter $t \leftarrow 0$.
$\hat{W} \leftarrow$ top-r right singular vectors of an initial unlabeled hidden-state matrix (from a pilot batch of option texts).
for episode $e = 1, 2, \dots$ **do**
  Retrieve option texts $\{o_{jk}^{(e)}\}$; compute $h(o_{jk}^{(e)})$ (one encoder forward pass per option).
  $\hat{\psi}_{jk}^{(e)} \leftarrow \hat{W}\, h(o_{jk}^{(e)})$ for all (j, k).
  for within-episode round $\tau = 1, \dots, T_e$ **do**
    $t \leftarrow t + 1$; $\beta_t \leftarrow \sqrt{2\log(Nt+1) + r\log(1+t)}$.
    $\hat{\phi}_i \leftarrow P_i^{-1} h_i$; $\Sigma_i \leftarrow P_i^{-1}$ for all i.
    Active decision set: $\mathcal{A}_t \leftarrow \{$ top-$m(t)$ decisions j by $\max_{k,k'} \|\hat{\psi}_{jk'}^{(e)} - \hat{\psi}_{jk}^{(e)}\|\}$; feasible active-set plans $\mathcal{X}_t(c) := \{z \in \mathcal{X} : z_j = c_j \text{ for } j \notin \mathcal{A}_t\}$.
    if $t \le T_0$ **then**
      $x_t^{(e)} \leftarrow$ uniform-random plan *// cold-start exploration*
    else
      $c \leftarrow \arg\max_k \text{mean}_i\big(\hat{\phi}_i^\top \hat{\psi}_{jk}^{(e)}\big)$ for all j.
      Coord-descent on c over decisions in $\mathcal{A}_t$: accept any toggle that decreases the separate-bound LCB-minimax score

$$S(c) = \max_i \big[\underline{U}_{i,\mathcal{A}_t}^{\max}(t;c) - \overline{U}_i(t;c)\big], \quad \underline{U}_{i,\mathcal{A}_t}^{\max}(t;c) := \max_{z \in \mathcal{X}_t(c)} \big[\hat{\phi}_i^\top \hat{\Psi}^{(e)}(z) - \beta_t \|\hat{\Psi}^{(e)}(z)\|_{\Sigma_i}\big],$$

$$\overline{U}_i(t;c) := \hat{\phi}_i^\top \hat{\Psi}^{(e)}(c) + \beta_t \|\hat{\Psi}^{(e)}(c)\|_{\Sigma_i}, \quad \hat{\Psi}^{(e)}(z) := \sum_j \hat{\psi}_{jz_j}^{(e)}.$$

      $x_t^{(e)} \leftarrow$ one-hot of the resulting c.
    end if
    Play $x_t^{(e)}$; observe $\tilde{U}_i^{(e)}(x_t^{(e)})$ for all i.
    $\Psi_t \leftarrow \sum_{jk} \hat{\psi}_{jk}^{(e)}(x_t^{(e)})_{jk}$.
    $P_i \leftarrow P_i + \Psi_t \Psi_t^\top / \sigma_{\text{obs}}^2$; $h_i \leftarrow h_i + \Psi_t\, \tilde{U}_i^{(e)}(x_t^{(e)}) / \sigma_{\text{obs}}^2$.
    Append $\{(i, e, \tau, H^{(e)}(x_t^{(e)}), \tilde{U}_i^{(e)}(x_t^{(e)}))\}_{i=1}^N$ to $\mathcal{D}$, where $H^{(e)}(x) = \sum_{jk} h(o_{jk}^{(e)}) x_{jk}$.
    if $t > T_0$ and $t \bmod K_{\text{refit}} = 0$ **then**
      Solve $\widetilde{W} \leftarrow \arg\min_{W \in \mathbb{R}^{r \times d_h}} \sum_{(i,e',t',H,\tilde{u}) \in \mathcal{D}} (\tilde{u} - \hat{\phi}_i^\top W H)^2 + \lambda_W \|W\|_F^2$ (unconstrained ridge), then project $\hat{W} \leftarrow \arg\min_{W:\, WW^\top = I_r} \|W - \widetilde{W}\|_F$ by SVD (row-orthonormalization).
      Rebuild (P_i, h_i) from $\mathcal{D}$ under the new $\hat{W}$: for each buffer entry, $\Psi \leftarrow \hat{W} H$, then $P_i \leftarrow \lambda I_r + \sum \Psi \Psi^\top / \sigma_{\text{obs}}^2$, $h_i \leftarrow \sum \Psi\, \tilde{u} / \sigma_{\text{obs}}^2$.
      Recompute $\hat{\psi}_{jk}^{(e)}$ for the remaining episode-e options using the new $\hat{W}$.
    end if
  end for
end for
Output: Persistent (P_i, h_i); per-episode chosen plans $\{x_t^{(e)}\}$; final $\hat{W}$.

The theorem is conditional on (A3)(i) and (A4)(ii) holding along the coordinator's own trajectory (Table 5 maps each condition to precedents). The bound applies to the posterior Algorithm 1 actually maintains: Lemma 3 splits the data-dependent probe into a fixed part, on which self-normalized concentration applies, and a remainder controlled deterministically by a projection inequality (Lemma 1).

Algorithm 1 differs from the theoretical variant $\mathcal{A}^{\text{T}}$ (Appendix B.2) in the following ways: (a) *approximate plan optimization.* The $\arg\min_x \max_i \text{LCB}_t$ step of (T1) is replaced by a progressive-widening coord-descent restricted to the top-$m(t)$ decisions, incur-

ring an additive optimization-error term $\sum_{e,t} \xi_t \geq 0$ with $\xi_t := \max_i \mathrm{LCB}_t(D_i)(x_t^{(e)}) - \min_{x \in \mathcal{X}} \max_i \mathrm{LCB}_t(D_i)(x)$. (b) *Active-set individual-optimum evaluation.* The individual-optimum lower bound is computed over the active-set feasible plans $\mathcal{X}_t(c)$ rather than the full $\mathcal{X}$, giving an additional additive term $\sum_{e,t} \zeta_t \geq 0$ with $\zeta_t := \max_i \left[\underline{U}_i^{\max}(t) - \underline{U}_{i,\mathcal{A}_t}^{\max}(t; c)\right]$ (both ξ_t and ζ_t vanish once $\mathcal{A}_t$ covers the support of the true individual optima). (c) *Practical confidence radius.* Algorithm 1 uses $\beta_t = \sqrt{2\log(Nt+1) + r\log(1+t)}$ in place of the theoretical radius $\bar{\beta}_t^{\text{full}}$ of (T3), dropping the W-bias, misspecification, and cross-term corrections. (d) *Full vs. top-k_h-projected ridge.* $\mathcal{A}^{\mathrm{T}}$'s refit (T2) fits the ridge on top-k_h-projected features $P_{k_h}^\top H(x_s)$ (ambient dim rk_h); Algorithm 1 fits the full $r \times d_h$ ridge without projection. (A5) bounds only the true probe's tail, not estimation variance in the remaining $d_h - k_h$ directions.

4 Empirical Results

4.1 Simulator

This section reports empirical results for Algorithm 1 on a controlled multi-episode simulator. We evaluate on a synthetic multi-episode instance with $N = 20$ users, $r = 4$, $E = 8$ episodes, $T = 20$ rounds per episode, $J = 30$ decisions, $K = 3$ options per decision. LLM-style hidden states are simulated as $h(o) = A\psi(o) + b + \eta$ with $A \in \mathbb{R}^{d_h \times r}$ a random Gaussian projection, b a bias, and $\eta \sim \mathcal{N}(0, \sigma_h^2 I)$ nuisance noise.

We compare joint learning against five contrast conditions that isolate different sources of the structured-representation advantage in this synthetic simulator.

- *joint*: full alternating-min: W learned from scalar feedback, $\{\phi_i\}$ learned from scalar feedback with a persistent posterior across episodes, r-dim latent space.
- *joint-coldstart*: identical to *joint* but with the per-user ϕ_i posterior reset at every episode boundary; the cross-episode data buffer for W refits is retained. Isolates per-user preference persistence from W transfer.
- *random-W*: W frozen at a random Gaussian; $\{\phi_i\}$ learned. Tests whether performance follows from an arbitrary low-dimensional projection or from the structured low-rank projection.
- *pca-W*: W frozen at the top-r SVD of pooled unlabeled hidden states, no bandit refit; $\{\phi_i\}$ learned. Isolates structured low-rank feature geometry (from the synthetic $A\psi + b + \eta$ generator) from feedback-driven W refinement.
- *onehot*: no LLM: per-episode one-hot option features (feature dim $J^{(e)}K^{(e)}$), ϕ posterior reset per episode (no cross-episode transfer possible because option identities do not repeat). Isolates the value of transferable structured features relative to episode-specific identity features.
- *fixed-true-$W^\star$*: simulator-only diagnostic with $\hat{W} = W^\star$ held fixed (no W-refit); $\{\phi_i\}$ learned.

Let $\Phi_e^\star$ denote the ground-truth worst-case-divergence optimum for episode e (computed offline by 20-restart coord-descent on true features). Table 1 reports the *per-episode best-plan gap*

$$G_e := \min_{t \in [T]} \Phi_e\big(x_t^{(e)}\big) - \Phi_e^\star, \qquad G_{\text{total}} := \sum_{e=1}^{E} G_e,$$

i.e., G_e is the suboptimality of the best plan the coordinator produced in episode e, and G_{total} is the total suboptimality across the E episodes. This differs from the standard bandit cumulative regret $R_{\text{total}} := \sum_e \sum_t (\Phi_e(x_t^{(e)}) - \Phi_e^\star)$ that appears in (3): G_e scores the episode-final plan, which is the operationally relevant deliverable to participants, while R_{total} also penalizes within-episode exploration.

As Figure 1 shows, joint's episode-1 G_1 (5.57) is already below fixed-true-$W^\star$'s (6.91) and pca-W's (7.08). Note that G_1 is the best-plan gap *over all* T=20 rounds of episode 1, during

Table 1: Multi-episode joint-learning coordinator vs baselines. Two metrics: $G_{\text{total}} = \sum_e (\min_t \Phi_e(x_t^{(e)}) - \Phi_e^\star)$ scores the per-episode best-plan (final-deliverable) gap summed over episodes; $R_{\text{total}} = \sum_{e,t} (\Phi_e(x_t^{(e)}) - \Phi_e^\star)$ is the round-wise cumulative pseudoregret that Theorem 1 bounds. E=8 episodes with T=20 rounds each; N=20, r=4, J=30, K=3, d_h=64, σ_h=0.1; G numbers from 30 seeds, R numbers from 10 seeds. On G_{total} (final-plan quality) *joint-coldstart* matches *joint*, suggesting per-user preference persistence is not required. But on R_{total} (roundwise regret) *joint-coldstart* is nearly 2× higher, resetting per-user posteriors makes the coordinator pay exploration cost each episode. The stronger *joint-newusers* ablation, in which $\phi_i^\star$ itself is resampled each episode while the W-refit buffer is retained under the same user indices, collapses performance on both metrics.

Mode	G_{total}	R_{total}
joint (main)	$\mathbf{22.06 \pm 1.47}$	$\mathbf{877 \pm 64}$
joint-coldstart (ϕ reset per episode; *same users*; W across)	19.99 ± 1.48	1799 ± 58
joint-newusers (ϕ_i resampled per episode; W across)	61.74 ± 2.61	3323 ± 185
fixed-true-$W^\star$ (simulator-only diagnostic)	46.57 ± 3.47	1469 ± 140
pca-W (frozen PCA warm start)	49.19 ± 4.14	1570 ± 157
random-W (frozen random projection)	100.01 ± 9.51	2640 ± 296
onehot (no LLM, per-episode one-hot options)	145.56 ± 7.35	2723 ± 165

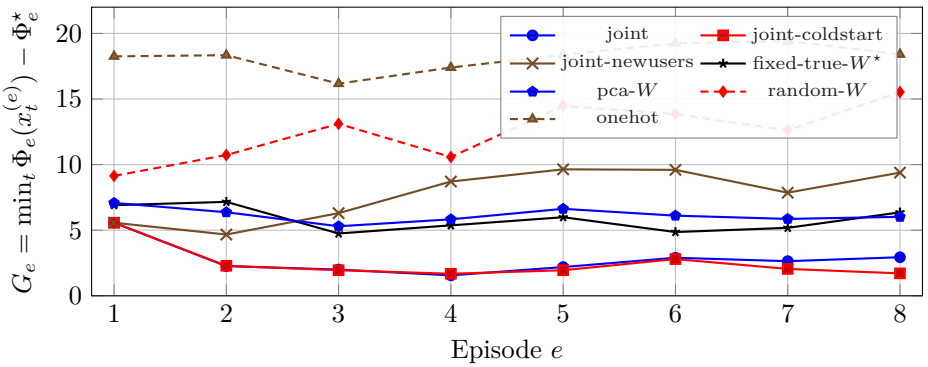

Figure 1: Per-episode best-plan gap $G_e := \min_{t \in [T]} \Phi_e(x_t^{(e)}) - \Phi_e^\star$ (suboptimality of the coordinator's best plan in episode e; per-episode summand of G_{total} in Table 1). Per-episode means; N=20, r=4, E=8, T=20, J=30, K=3, d_h=64. All seven modes of Table 1 shown; aggregate SEs on G_{total} and R_{total} appear in Table 1 (per-episode SEs not shown). joint and joint-coldstart drop sharply from episode 1 to episode 2–3 and plateau at $G_e \in [1.5, 3]$; joint-newusers rises after episode 1 as the W-refit buffer accumulates observations under indices whose latent $\phi_i^\star$ has been resampled; fixed-true-$W^\star$ and pca-W stay in the 5–7 band with no downward trend; random-W drifts upward without stabilizing; onehot stays flat at 16–19 (no cross-episode transfer, since option identities do not repeat).

which two W-refits fire (at t=10 and t=20) after the T_0=5 uniform cold-start; thus what joint's episode-1 gap actually reflects is the PCA warm-start plus the two within-episode refits, not the warm-start alone. Even so, joint beats fixed-true-$W^\star$ at episode 1 despite starting from a worse probe, consistent with the next observation: joint's W refits track the observed feedback and inject implicit exploration diversity, while fixed-true-$W^\star$'s frozen features let the ϕ_i posterior shrink fast and cause premature exploitation. *Second*, the cold-start variant (per-user ϕ_i posterior reset at every episode boundary but W refits retain the cross-episode data buffer) matches joint on final-plan quality G_{total}; both eventually converge to good plans within each episode. Finally, on the round-wise regret metric R_{total}, cold-start pays a nearly 2× penalty vs joint (1799 vs 877), and the stronger *new-users* ablation (in which $\phi_i^\star$ is resampled each episode while the W-refit buffer is retained) collapses performance on both metrics ($G_{\text{total}} = 61$, $R_{\text{total}} = 3323$, respectively 2.8× and 3.8× joint). The mechanism is therefore consistent with *shared users* × *cumulative* W *refinement*: the persistent participant pool provides an anchor on which the W refits' data-buffer $\hat{\phi}_i$ values

remain accurate, and per-user preference persistence within a shared pool matters most for round-wise exploration efficiency rather than for eventual plan quality.

4.2 Actual encoders

4.2.1 Error measurements

We repeated the experiment with real per-option hidden states extracted from three instruction-tuned LLMs (Qwen3-8B, Llama-3.1-8B-Instruct, gemma-3-4b-it) and one non-LLM sentence encoder (all-MiniLM-L6-v2), on a deterministic multi-episode corpus of team-process rules ($E = 8$, $J = 30$, $K = 3 = 720$ option texts). Latent ψ is synthetic ($r = 4$); for the three LLMs we use the last-layer state at the last token, and for all-MiniLM-L6-v2 we use its default mean-pooled sentence embedding. A ridge probe W_{fit} is trained on episodes 1–4 and evaluated on the held-out episodes 5–8 (an episode-held-out / unseen-option split); the reported $\hat{\varepsilon}$ is the held-out value. Table 2 reports empirical diagnostics related to the representation quantities motivating the theorem; Table 3 reports G_{total} and R_{total} across all six modes and all four encoders (10 seeds each).

Table 2: Encoder diagnostics on the actual-encoder corpus (720 option texts, $r = 4$). $\hat{\varepsilon}$: held-out bilinear probe residual $\|\psi - W_{\text{fit}}H\|_F/\|\psi\|_F$ (train on episodes 1–4, evaluate on 5–8). $r_H = \text{tr}(\Sigma_h)/\|\Sigma_h\|_{\text{op}}$: stable rank. k_h^p: smallest integer with cumulative eigen-mass $\geq p$. All four encoders have $k_h^p \ll d_h$ (right-most column).

Encoder	$\hat{\varepsilon}$	r_H	k_h^{50}	k_h^{90}	k_h^{95}	d_h
Qwen3-8B	0.110	3.6	4	22	45	4096
Llama-3.1-8B-Instruct	0.118	6.1	6	29	60	4096
gemma-3-4b-it	0.172	5.4	6	44	94	3072
all-MiniLM-L6-v2	**0.129**	6.1	**4**	**14**	**19**	384

Table 3: Actual-encoder multi-episode coordinator results across four encoders and six modes; 10 seeds per cell, $N = 20$, $r = 4$, $E = 8$, $T = 20$ rounds per episode. Utility target $U_i(o) = \phi_i^\top \psi(o)$ is encoder-independent: same ψ (from the corpus generator), ϕ_i, episodes, and paired seeds across all four rows; only $H(o)$ changes. Note that one-hot's G_{total} (28.75 ± 0.03) and R_{total} (680 ± 1) are identical to numerical precision across the four rows: since one-hot never touches $H(o)$, this is direct empirical evidence that the four rows really are comparisons on the same utility landscape. Across encoders, joint learning reduces G_{total} by 1.9–2.5× vs random-W and by 3.2–4.9× vs one-hot; under native feature scaling, the non-LLM sentence-transformer all-MiniLM-L6-v2 is the best non-oracle encoder on both metrics.

G_{total} (mean ± SE)						
Encoder	joint	joint-cs	random-W	pca-W	onehot	supervised-W
Qwen3-8B	8.93 ± 0.69	9.97 ± 0.35	16.85 ± 1.16	15.47 ± 0.94	28.75 ± 0.03	5.93 ± 0.29
Llama-3.1-8B-Instruct	8.11 ± 0.52	7.96 ± 0.42	16.83 ± 0.78	23.58 ± 0.93	28.75 ± 0.03	4.63 ± 0.36
gemma-3-4b-it	8.23 ± 0.47	9.13 ± 0.59	19.19 ± 1.71	23.37 ± 1.47	28.75 ± 0.03	5.47 ± 0.37
all-MiniLM-L6-v2	$\mathbf{5.91 \pm 0.42}$	6.38 ± 0.72	14.64 ± 0.91	21.82 ± 1.12	28.75 ± 0.03	4.11 ± 0.25
R_{total} (mean ± SE)						
Encoder	joint	joint-cs	random-W	pca-W	onehot	supervised-W
Qwen3-8B	447 ± 14	681 ± 17	718 ± 35	665 ± 26	680 ± 1	280 ± 9
Llama-3.1-8B-Instruct	449 ± 14	665 ± 9	738 ± 26	757 ± 22	680 ± 1	251 ± 5
gemma-3-4b-it	476 ± 10	694 ± 16	732 ± 30	842 ± 25	680 ± 1	309 ± 9
all-MiniLM-L6-v2	$\mathbf{369 \pm 16}$	563 ± 18	684 ± 38	827 ± 35	680 ± 1	235 ± 6

We summarize four findings from the experiments. *(F1) Real encoder hidden states exhibit substantial spectral compression on this corpus.* Every encoder has k_h^{95} substantially smaller than its ambient d_h (45–94 for the 8B LLMs vs $d_h \in \{3072, 4096\}$), and held-out probe residual $\hat{\varepsilon}_{\text{heldout}}$ small enough that a linear r-dim probe recovers most of the utility-carrying signal. Joint learning correspondingly reduces G_{total} by 1.9–2.5× vs random-W across encoders; the remaining 1.44–1.75× gap to *supervised-W* (the coordinator handed W_{fit} and skipping joint refit) shows the algorithm gets within a factor of 1.5–2 of a supervised reference.

(F2) Cold-start (per-user ϕ_i reset per episode, W buffer retained) matches joint on G_{total} across all four encoders but is $\sim 1.5\times$ worse on R_{total}, replicating the simulator's (G, R) decomposition on real encoders. *(F3) Data-driven W refit beats spectral truncation.* pca-W is worse than random-W on Llama-3.1-8B-Instruct (23.58 vs 16.83), on gemma-3-4b-it (23.37 vs 19.19), and on MiniLM (21.82 vs 14.64), i.e., the top-r subspace of Σ_h is *not* well aligned with the utility-carrying subspace for real encoder hidden states. *(F4) Under native feature scaling, a small non-LLM sentence encoder is the best non-oracle encoder.* all-MiniLM-L6-v2 achieves $G_{\text{total}} = 5.91$ and $R_{\text{total}} = 369$, materially better than every 8B LLM on both metrics; its smaller $k_h^{95} = 19$ is consistent with the theorem's prediction that lower representation complexity can improve joint learning. Sentence-transformer models are trained for fixed-dimensional sentence embeddings, which may favor linear recoverability in our setting.

4.2.2 Effective-rank (r_H) scaling on the simulator

Theorem 1 predicts that the W-side of the regret grows as $\sqrt{k_h}$. We test the analogous prediction for the observable effective rank $r_H = \text{tr}(\Sigma_h)/\|\Sigma_h\|_{\text{op}}$ by sweeping the nuisance-noise standard deviation σ_h in the synthetic hidden-state generator $h(o) = A\psi(o) + b + \eta$, which increases the effective rank of the observed Σ_h from $\approx r$ (at $\sigma_h = 0$) toward the ambient $d_h = 64$ (at $\sigma_h = 3$). Table 4 reports measured r_H, G_{total}, and R_{total} for the joint coordinator across five settings of σ_h (10 seeds each). Both performance metrics grow monotonically with r_H (roughly a $6\times$ growth from $r_H \approx 3$ to $r_H \approx 19$), consistent with the theorem's $\sqrt{k_h}$ dependence.

Table 4: Effective-rank sweep at fixed r=4, N=20, E=8, T=20, d_h=64. Measured $r_H = \text{tr}(\Sigma_h)/\|\Sigma_h\|_{\text{op}}$ on episode-1 hidden states; performance metrics are for the joint-learning coordinator (10 seeds).

σ_h	measured r_H	G_{total}	R_{total}
0.0	3.1	11.15 ± 1.32	468 ± 42
0.1	3.6	19.22 ± 1.75	877 ± 64
0.3	6.8	35.51 ± 3.51	1620 ± 83
1.0	18.2	60.52 ± 4.52	2751 ± 104
3.0	19.4	66.10 ± 8.06	3101 ± 193

5 Conclusions

We formulated representation-guided group decision-making as a multi-episode coordination protocol with a fixed participant pool and a bilinear low-rank utility model, and presented a joint-learning coordinator that recovers a linear probe W and per-user preferences $\{\phi_i\}$ from scalar feedback alone (Algorithm 1). Theorem 1 gives a cumulative pseudoregret guarantee for the exact-LCB idealization A^T, with the representation-dependent contribution controlled by the spectral dimension k_h rather than the ambient encoder dimension d_h, and with the analysis applying to the retroactively rebuilt posterior used by Algorithm 1.

On a controlled synthetic simulator, joint learning cuts G_{total} to less than half of any fixed-representation baseline; the (G_{total}, R_{total}) decomposition separates two mechanisms: cross-episode W refinement drives final-plan quality, while persistent per-user posteriors reduce interaction cost by $\sim 2\times$. Actual-encoder validation across four pretrained encoders (Qwen3-8B, Llama-3.1-8B-Instruct, gemma-3-4b-it, all-MiniLM-L6-v2) replicates both findings and quantitatively supports the theorem's approximate-realizability and small-k_h assumptions ($\hat{\varepsilon} \leq 0.17$, $k_h^{95} \leq 94 \ll d_h$ across all four encoders). Two further findings sharpen the empirical picture: data-driven W-refit beats spectral truncation of Σ_h (pca-W is worse than random-W on three of the four encoders, indicating the top-r subspace of Σ_h is *not* well aligned with the utility-carrying subspace), and, under native feature scaling, the non-LLM MiniLM encoder is the best non-oracle encoder on both metrics, indicating the method is representation-agnostic rather than LLM-specific.

AI Use Statement

The authors used large language models for coding assistance (drafting and refactoring simulator, extraction, and analysis scripts), writing and editing assistance, and review-style critical feedback on drafts. All model outputs were reviewed by the authors, who retain full responsibility for the scientific content, experimental design, empirical results, theoretical claims, and any errors.

Ethics Statement

This work is a theoretical and simulator-based study of AI-mediated group decision-making; no human-subjects data was collected. All experiments use synthetic personas generated from a rank-r latent model and a 720-option corpus of team-process rules generated by the authors; pretrained encoders (Qwen3-8B, Llama-3.1-8B-Instruct, gemma-3-4b-it, all-MiniLM-L6-v2) were used for feature extraction under their respective open-source licenses. Because the coordinator's role is to elicit and aggregate revealed preferences from participants, deployment in real group settings would raise several concerns we do not address here: (i) preference elicitation may reflect, amplify, or misrepresent existing power asymmetries within a group; (ii) participants may not consent to having their preferences modeled by a low-rank latent representation; (iii) the coordinator's plan recommendations could be used to legitimize decisions that a subset of participants oppose. Any real-world deployment should be accompanied by IRB review, informed consent, disclosure of the aggregation rule, and human oversight of recommendations, none of which our synthetic evaluation validates. The pretrained encoders we use inherit whatever biases are present in their training data; we make no claims about equitable representation across participant subgroups.

Reproducibility

Full details for reproducing all results are provided in the paper and appendix. The multi-episode simulator (Section 4) reports all instance parameters ($N = 20$, $r = 4$, $E = 8$, $T = 20$, $J = 30$, $K = 3$, $d_h = 64$, $\sigma_h = 0.1$, 30/10 seeds); Appendix A.1 lists coordinator hyper-parameters (λ, σ_{obs}, λ_W, K_{refit}, T_0, $m(t)$, β_t). The actual-encoder experiment (Section 4, Tables 2–3) uses a deterministic 720-option corpus of team-process rules with fixed per-model extraction protocols (last-layer last-token state for the three LLMs; the default mean-pooled sentence embedding for all-MiniLM-L6-v2) and a synthetic rank-4 latent ψ from the corpus generator, held-out probe residual $\hat{\varepsilon}$ computed on episodes 5–8 after fitting on episodes 1–4. Code, corpus generator, extraction scripts, coordinator, and result-aggregation utilities will be released with the camera-ready version.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

Kenneth J. Arrow. *Social Choice and Individual Values*. Wiley, 1951.

Tim Baarslag, Michael Kaisers, Enrico H. Gerding, Catholijn M. Jonker, and Jonathan Gratch. When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4684–4690, 2017.

Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D. Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. *Artificial Intelligence*, 227:190–213, 2015.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (eds.). *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In *Proc. International Conference on Machine Learning (ICML)*, 2020.

Clyde H. Coombs. Psychological scaling without a unit of measurement. *Psychological Review*, 57(3):145–158, 1950.

Shaheen Fatima, Sarit Kraus, and Michael Wooldridge. *Principles of Automated Negotiation*. Cambridge University Press, 2014.

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. In *Advances in Neural Information Processing Systems*, 2020.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. In *AAAI Conference on Artificial Intelligence*, 2017.

Moritz Hardt. Understanding alternating minimization for matrix completion. In *IEEE Symposium on Foundations of Computer Science*, 2014.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proc. ACM Symposium on Theory of Computing (STOC)*, 2013.

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Low-rank matrix bandit under bilinear contextual model. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *Proc. International Conference on Machine Learning (ICML)*, 2019.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Efficient frameworks for generalized low-rank matrix bandit problems. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Sarit Kraus. *Strategic Negotiation in Multiagent Environments*. MIT Press, 2001.

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

Tor Lattimore, Csaba Szepesvári, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning (ICML)*, 2020.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 24(2):193–202, 1975.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *Proc. International Conference on Machine Learning (ICML)*, pp. 4218–4227, 2018.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, pp. 1015–1022, 2010.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):ead3084, 2024.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *Advances in Neural Information Processing Systems*, 2020.

Michael Wooldridge. *An Introduction to MultiAgent Systems*. Wiley, 2nd edition, 2009.

Jiaqi Yang, Wei Hu, Jason D. Lee, and Simon S. Du. Impact of representation learning in linear bandits. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

A Notation Glossary

Symbol	Meaning	First used
N	number of participants (fixed pool across episodes)	§3
E	number of episodes	§3
T	rounds per episode	§3
$J^{(e)}, K_j^{(e)}$	number of decisions and options per decision in episode e	§3
r	latent preference dimension (rank of $W^\star$)	Eq. (1)
d_h	pretrained-encoder representation dimension	Eq. (1)
$\phi_i \in \mathbb{R}^r$	participant i's persistent latent preference vector	Eq. (1)
S	norm bound on preferences: $\|\phi_i^\star\| \le S$, (A1)	App. B.1
$W^\star \in \mathbb{R}^{r\times d_h}$	true probe (row-orthonormal, $W^\star W^{\star\top} = I_r$)	Eq. (1)
$h(o) = f_\theta(o)$	encoder output for option o	Eq. (1)
$H_{\max}$	per-option feature bound: $\sup_o \|h(o)\| \le H_{\max}$	App. B.1
$H^{(e)}(x)$	plan feature, $\sum_{jk} h(o_{jk}^{(e)}) x_{jk}$	Eq. (1)
H_{plan}	per-plan feature bound: $\sup_{x\in\mathcal{X}} \|H(x)\| \le H_{\text{plan}} \le JH_{\max}$	App. B.1
σ_{obs}	observation-noise scale (sub-Gaussian)	Eq. (1)
P_i, h_i	natural-parameter Gaussian posterior on ϕ_i; note $h_i \ne h(o)$	Alg. 1
$D_i(x), \Phi_e(x), \Phi_e^\star$	per-user divergence, worst-participant divergence, offline optimum	Eq. (2)
$G_e, G_{\text{total}}, R_{\text{total}}$	per-episode best-plan gap; total gap; cumulative pseudoregret	§4
δ	failure probability parameter (theorem holds w.p. $\ge 1-\delta$)	Thm. 1

ε	bilinear-realizability residual, (A1)	App. B.1
k_h, γ_h	integer spectral dim of Σ_h; top-k_h eigenvalue lower bound, (A2)	App. B.1
P_{k_h}	top-k_h eigenprojector of Σ_h, (A2), (A3)(i), (A5)	App. B.1
γ_h^{play}	played plan-level excitation on top-k_h subspace, (A3)(i)	App. B.1
ε_h	$W^\star$ top-k_h alignment residual, (A5)	App. B.1
γ_ϕ	participant-diversity excitation, (A3)(ii)	App. B.1
ρ_0	PCA warm-start basin radius, (A4)(i)	App. B.1
$\rho_\infty(m), \rho_\varepsilon$	probe-recovery statistical rate at m pooled observations; misspecification floor, (A4)(ii)	App. B.1
$e_k, \bar{e}_k$	joint estimation error after the k-th refit; its deterministic envelope, (A4)(ii)	App. B.1
$\bar{\theta}_t$	computable envelope of θ_t: $\sqrt{2}\,\bar{e}_k + \varepsilon_h^2$ on $[t_k, t_{k+1})$, (A4)(ii)	App. B.1
t_1	basin-entry refit round, (A4)(ii)	App. B.1
ε_{eff}	effective misspecification: $\varepsilon + c_{\text{align}} S H_{\text{plan}} \varepsilon_h$	Thm. 1
R_{lo}	lower-order remainder (log, $\sqrt{ET}$, and burn-in terms)	Thm. 1
$P := P_{k_h}$	top-k_h eigenprojector (proof notation)	App. B.2
$\tilde{H}_s := P^\top H(x_s)$	projected plan feature in $\mathbb{R}^{k_h}$	App. B.2
$B_0 := W^\star P$	projected true probe ($B_0 B_0^\top \succeq (1-\varepsilon_h^2) I_r$)	App. B.2
$B_t := \hat{W}_t \in \mathbb{R}^{r\times k_h}$	projected probe estimate in force at round t (constant between refits)	App. B.2
$\sin\Theta_t$	$\sin\Theta(\text{row}\, B_t, \text{row}\, B_0)$	Lemma 2
θ_t	operator-norm alignment error $\sqrt{2}\sin\Theta_t + \varepsilon_h^2$	Lemma 2
Q_t	rotation aligning B_0 to B_t; $\tilde{\phi}_i^{(t)} := Q_t \phi_i^\star$ is the truth in B_t's gauge	Lemma 2
Γ_t	projected-feature Gram $\sum_{s\le t} \tilde{H}_s \tilde{H}_s^\top$ (distinct from the per-episode gap G_e above)	App. B.2
V_t	shared posterior precision under B_t	App. B.2
$\sigma_i^{(t)}(z)$	posterior width in direction z	App. B.2
β_t^{full}, $\bar{\beta}_t^{\text{full}}$	full confidence radius $\beta_t^{\text{noise}} + \beta_t^W + \beta_t^\varepsilon + \beta_t^\times$ (Lemma 3); the same with $\theta_t \to \bar{\theta}_t$ (used by $\mathcal{A}^{\text{T}}$)	App. B.2
β_t (practical)	$\sqrt{2\log(Nt+1) + r\log(1+t)}$ used by Algorithm 1	Alg. 1

ξ_t, ζ_t	optimization gap (coord-descent) and active-set gap of Algorithm 1	§3
$c, c_1, \ldots, c_6, c', c''$	positive absolute constants (assumptions)	throughout
c_{align}	dimension-free constant absorbing the ε_h tail into ε_{eff}	throughout
$\lambda, \lambda_W, T_0, K_{\text{refit}}, m(t)$	Algorithm 1 hyperparameters (defaults in §A.1)	Alg. 1

A.1 Hyperparameter defaults

Reported simulator runs use $\lambda = 1$, $\sigma_{\text{obs}} = 0.1$, $\lambda_W = 10^{-3}$, $K_{\text{refit}} = 10$, $T_0 = 5$, $m(t) = \lceil 1.5\, t^{2/3} \rceil$ capped at $J^{(e)}$, and LCB coefficient $\beta_t = \sqrt{2\log(Nt+1) + r\log(1+t)}$. Ablation sensitivity to these choices is not reported.

B Theory: Assumptions and Proof of Theorem 1

B.1 Formal Assumptions

We work with bounded features: $\sup_o \|h(o)\| \le H_{\max}$ per option and $H_{\text{plan}} := \sup_{x \in \mathcal{X}} \|H(x)\|$ per plan (with $H_{\text{plan}} \le J H_{\max}$). The effective misspecification level $\varepsilon_{\text{eff}} := \varepsilon + c_{\text{align}} \cdot S\, H_{\text{plan}}\, \varepsilon_h$ combines the bilinear residual ε from (A1) with the $W^\star$-alignment tail ε_h from (A5), scaled by user and feature norms (as in Theorem 1). The second-moment matrix is $\Sigma_h := \mathbb{E}_o[h(o)h(o)^\top]$.

The theorem uses *five* conditions.

(A1) *Approximate bilinear model.* There exist $W^\star \in \mathbb{R}^{r \times d_h}$ (normalization convention: $W^\star W^{\star\top} = I_r$) and $\{\phi_i^\star\}$ with $\|\phi_i^\star\| \le S$ such that $U_i^{(e)}(x) = \phi_i^{\star\top} W^\star H^{(e)}(x) + \delta_i^{(e)}(x)$ with $|\delta_i^{(e)}(x)| \le \varepsilon$ (misspecification); the observed feedback $\tilde{U}_{i,t}^{(e)} = U_i^{(e)}(x_t) + \eta_{i,t}$ has $\eta_{i,t}$ sub-Gaussian at scale σ_{obs}. The $W^\star W^{\star\top} = I_r$ row-orthonormality is a normalization (fixing the joint scale of $W^\star$ and $\phi_i^\star$), not a restriction on the model class.

(A2) *Hidden-state spectral compression.* There exists an integer $k_h \ge 1$ with $\lambda_{k_h}(\Sigma_h) > \lambda_{k_h+1}(\Sigma_h)$ and $\lambda_{k_h}(\Sigma_h) \ge \gamma_h > 0$ (a strict spectral gap after the k_h-th eigenvalue); this makes the top-k_h eigenprojector P_{k_h} uniquely defined. Encoder-side condition; the representation-dependent term of Theorem 1 scales with k_h.

(A3) *Excitation pair on the played design.* Two excitation conditions, jointly required for the W-recovery lemma; part (i) alone also drives the ϕ-side concentration:

- (i) *Plan-level hidden-space excitation.* The coordinator's played plan-features satisfy $\lambda_{\min}(P_{k_h}^\top (\frac{1}{t}\sum_{s \le t} H(x_s^{(e_s)}) H(x_s^{(e_s)})^\top) P_{k_h}) \ge \gamma_h^{\text{play}} > 0$ for all $t \ge T_0$ (the cold-start horizon of Algorithm 1). Plan-level analog of (A2): observations are on plan sums $H(x) = \sum_{jk} h(o_{jk}) x_{jk}$ chosen adaptively, so option-level Σ_h alone does not guarantee excitation of the top-k_h subspace in the played design.
- (ii) *Participant excitation.* $N^{-1} \sum_{i=1}^N \phi_i^\star \phi_i^{\star\top} \succeq \gamma_\phi I_r$ with $\gamma_\phi > 0$.

Condition (i) concerns the plans the coordinator itself chooses and is not guaranteed by pure LCB selection; Theorem 1 is stated conditional on (A3)(i) and (A4)(ii) holding along $\mathcal{A}^{\text{T}}$'s own trajectory.

(A4) *Warm start plus alternating-minimization contraction.* Two sub-conditions packaging the "alt-min works" hypothesis:

- (i) *Warm-start alignment.* For a constant ρ_0 with $\rho_0\, H_{\text{plan}}^2 S^2 < c_1 \gamma_h \gamma_\phi$ (basin radius), the warm start satisfies $\sin\Theta(\text{row}(\hat{W}_0), \text{row}(W^\star)) \le \rho_0$ with probability $\ge 1 - \delta/4$.
- (ii) *Per-refit local contraction.* Assume $K_{\text{refit}} \ge T_0$. Let $t_k := k K_{\text{refit}}$ and let

$$e_k := \max\Big\{ \sin\Theta\big(\text{row}(\hat{W}_{t_k}), \text{row}(W^\star P_{k_h})\big),\ \max_i \|\hat{\phi}_{i,t_k} - Q_{t_k} \phi_i^\star\|/S \Big\}$$

be the joint estimation error after the k-th refit ($k \ge 1$), where the refit is the one-step map of Algorithm 1 (projected ridge for $\hat{W}$ given the current $\{\hat{\phi}_i\}$, then posterior rebuild under the new $\hat{W}$) and $Q_{t_k} \in O(r)$ is the rotation of Lemma 2; set $e_0 := \sin\Theta(\text{row}(\hat{W}_0), \text{row}(W^\star)) \le \rho_0$ by (A4)(i) (the initial $\{\hat{\phi}_i\}$ are not assumed accurate). Under (A1)–(A3)+(A4)(i), the refit map contracts:

$$e_k \ \le\ \tfrac{1}{2}\, e_{k-1} + c_5\big(\rho_\infty(N t_k) + \rho_\varepsilon\big), \qquad k \ge 1,$$

where $\rho_\infty(m) := c_3 \sigma_{\text{obs}} \sqrt{r k_h \log(d_h/\delta)/(m\, \gamma_\phi \gamma_h^{\text{play}})}$ and $\rho_\varepsilon := c_4 \varepsilon_{\text{eff}} / \sqrt{\gamma_\phi \gamma_h^{\text{play}}}$, with per-refit failure probability $\le \delta K_{\text{refit}}/(4ET)$ (total $\le \delta/4$ over all refits). These are the statistical rate and misspecification floor of the top-k_h-projected ridge estimator of (T2) with *known*

$\{\phi_i^\star\}$ from m pooled observations ($m \geq c_2 k_h S^2 H_{\text{plan}}^2 \log(d_h/\delta)/(\gamma_\phi \gamma_h^{\text{play}})$, fixed ridge $\lambda_W \asymp \sigma_{\text{obs}}^2$): the vectorized design $z_{i,s} = \phi_i^\star \otimes \tilde{H}_s \in \mathbb{R}^{rk_h}$ has pooled Gram $(N^{-1}\sum_i \phi_i^\star \phi_i^{\star\top}) \otimes (T_0^{-1}\sum_s \tilde{H}_s \tilde{H}_s^\top) \succeq \gamma_\phi \gamma_h^{\text{play}} I$ by (A3), the residual $\nu'_{i,s}$ (bilinear slack plus the (A5) tail $\phi_i^{\star\top} W^\star (I - P_{k_h} P_{k_h}^\top) H(x_s)$) satisfies $|\nu'_{i,s}| \leq \varepsilon_{\text{eff}}$, so Abbasi-Yadkori et al. (2011, Thm. 2) and Lemma 1 bound the Frobenius error by $c'\sigma_{\text{obs}}\sqrt{rk_h \log(d_h/\delta)/(m\gamma_\phi \gamma_h^{\text{play}})} + c''\varepsilon_{\text{eff}}/\sqrt{\gamma_\phi \gamma_h^{\text{play}}}$ (Ghosh et al., 2017), and Davis–Kahan (Yang et al., 2021, Prop. 2.5) converts it to $\sin\Theta$ after row-orthonormalization. The hypothesis is that the one-step refit with *estimated* $\{\hat{\phi}_i\}$ contracts toward this rate. Unrolling (using $\rho_\infty(Nt_j) \leq \sqrt{k/j}\,\rho_\infty(Nt_k)$ for $j \leq k$) gives the deterministic envelope

$$e_k \;\leq\; \bar{e}_k := 2^{-k}\rho_0 + c_6\big(\rho_\infty(Nt_k) + \rho_\varepsilon\big), \qquad \bar{\theta}_t := \sqrt{2}\,\bar{e}_k + \varepsilon_h^2 \;\text{ for } t \in [t_k, t_{k+1}),$$

which depends only on the problem constants $(\rho_0, \sigma_{\text{obs}}, \gamma_\phi, \gamma_h^{\text{play}}, \varepsilon_{\text{eff}}, \varepsilon_h)$ and is therefore available to $\mathcal{A}^{\text{T}}$ given conservative bounds on these constants (upper bounds on $\rho_0, \sigma_{\text{obs}}, \varepsilon_{\text{eff}}, \varepsilon_h$, lower bounds on $\gamma_\phi, \gamma_h^{\text{play}}$; any such bounds only enlarge $\bar{\theta}_t$); on the good event $\theta_t \leq \bar{\theta}_t$. We also assume $(\rho_0 + c_6(\rho_\infty(NT_0) + \rho_\varepsilon))S^2 \leq \gamma_\phi/4$, so that the estimated participant Gram stays $\succeq (\gamma_\phi/2) I_r$ at every refit. Let

$$t_1 := \min\big\{t_k :\; 2^{-k}\rho_0 \leq \rho_\infty(Nt_k) + \rho_\varepsilon \;\text{ and }\; \bar{\theta}_{t_k} \leq \gamma_h^{\text{play}}/(8H_{\text{plan}}^2)\big\}$$

be the *basin-entry* round, after which the warm-start error has been contracted below the statistical floor and the probe error is small relative to the played excitation (condition (6) of Lemma 3); we assume $\sqrt{2}(c_6+1)\rho_\varepsilon + \varepsilon_h^2 < \gamma_h^{\text{play}}/(8H_{\text{plan}}^2)$ so that $t_1 < \infty$. For $t \geq t_1$, $\bar{\theta}_t \leq \sqrt{2}(c_6+1)(\rho_\infty(Nt_k) + \rho_\varepsilon) + \varepsilon_h^2$.

(A5) *$W^\star$ top-k_h alignment.* The true probe row-space is approximately aligned with the top-k_h eigenspace of Σ_h: $\|W^\star(I - P_{k_h}P_{k_h}^\top)\|_{\text{op}} \leq \varepsilon_h$.

Table 5 maps each of (A1)–(A5) to the closest precedent in the bandit / matrix-recovery literature and states what, if anything, is specific to our setting. Only (A5) has no direct counterpart.

Table 5: Lineage of the five conditions used by Theorem 1.

Cond.	Content	Precedent(s)	Status
(A1)	Bilinear utility, sub-Gaussian noise, ε-slack	Jun et al. (2019); Kang et al. (2022); Jang et al. (2024); Abbasi-Yadkori et al. (2011); Ghosh et al. (2017); Foster et al. (2020)	standard
(A2)	Integer spectral dim k_h (spectral gap)	Spectral truncation / stable-rank feature analysis (Tripuraneni et al., 2020); tail folded into ε_{eff} via (A5)	adapted (scope-narrowing)
(A3)(i)	Played-plan excitation on top-k_h	"Diverse actions" / min-eigenvalue Gram condition in linear bandits (Rusmevichientong & Tsitsiklis, 2010; Lattimore & Szepesvári, 2020)	standard
(A3)(ii)	Participant diversity γ_ϕ	Task diversity in meta / multi-task linear bandits (Cella et al., 2020; Tripuraneni et al., 2020; Yang et al., 2021)	standard
(A4)(i)	Warm-start alignment into local basin; PCA is one possible initialization	Standard warm-start analysis for alt-min (Jain et al., 2013; Hardt, 2014; Yang et al., 2021)	adapted / assumed
(A4)(ii)	Local geometric contraction of alt-min	Same lineage; conclusion of the above analyses (adopted as hypothesis here)	adapted / assumed
(A5)	Alignment $\|W^\star(I - P_{k_h}P_{k_h}^\top)\|_{\text{op}} \leq \varepsilon_h$	Task/subspace compatibility in transfer learning (Tripuraneni et al., 2020), but no direct bilinear-bandit precedent	paper-specific (scope-narrowing)

B.2 Detailed Proofs

This appendix contains the proof of Theorem 1 and its supporting lemmas.

Notation. Write $P := P_{k_h} \in \mathbb{R}^{d_h \times k_h}$ and $\tilde{H}_s := P^\top H(x_s) \in \mathbb{R}^{k_h}$ for the projected plan feature ($\|\tilde{H}_s\| \leq H_{\text{plan}}$). Let

$$B_0 := W^\star P \in \mathbb{R}^{r \times k_h}$$

be the projected true probe; by (A5), $B_0 B_0^\top = I_r - W^\star(I - PP^\top)W^{\star\top} \succeq (1 - \varepsilon_h^2) I_r$, and we assume $\varepsilon_h^2 \leq 1/2$ throughout. Let $B_t := \hat{W}_t \in \mathbb{R}^{r \times k_h}$ (row-orthonormal) denote the coordinator's projected probe estimate in force at round t; B_t is constant on each inter-refit block $[t_k, t_{k+1})$ with $t_k := kK_{\text{refit}}$. Let $\Gamma_t := \sum_{s \leq t} \tilde{H}_s \tilde{H}_s^\top$ and

$$\sin\Theta_t := \sin\Theta\big(\text{row}(B_t), \text{row}(B_0)\big), \qquad \theta_t := \sqrt{2}\,\sin\Theta_t + \varepsilon_h^2.$$

(P is an isometry, so $\sin\Theta_t$ is the quantity (A4)(ii) controls.)

Idealized exact-LCB coordinator ($\mathcal{A}^{\mathrm{T}}$, theoretical variant). The coordinator to which the theorem applies is defined by three modifications relative to Algorithm 1:

(T1) *Exact-LCB plan selection.* Every round t plays $x_t^{(e)} \in \arg\min_{x\in\mathcal{X}} \max_i \mathrm{LCB}_t(D_i)(x)$, replacing coord-descent + progressive widening.

(T2) *Top-k_h-projected refit.* The refit step of $\mathcal{A}^{\mathrm{T}}$ fits the two-stage ridge of Algorithm 1 on the projected features $\tilde{H}_s = P^\top H(x_s)$ (ambient dimension rk_h): $\widetilde{W} \in \arg\min_{W\in\mathbb{R}^{r\times k_h}} \sum_{(i,s)} (\tilde{U}_{i,s} - \hat{\phi}_i^\top W \tilde{H}_s)^2 + \lambda_W \|W\|_F^2$, then $\hat{W}_t \in \arg\min_{WW^\top = I_r} \|W - \widetilde{W}\|_F$, with $\hat{W}_t^{\mathrm{full}} := \hat{W}_t P^\top \in \mathbb{R}^{r\times d_h}$ used for downstream plan scoring. Algorithm 1 instead fits the full $r \times d_h$ ridge without projection.

(T3) *Full confidence radius.* $\mathcal{A}^{\mathrm{T}}$ uses $\bar{\beta}_t^{\mathrm{full}}$, the radius of Lemma 3 with the true (unobservable, $W^\star$-dependent) subspace error θ_t replaced by the deterministic envelope $\bar{\theta}_t$ of (A4)(ii), in place of the practical $\beta_t = \sqrt{2\log(Nt+1) + r\log(1+t)}$.

The per-user posterior is the same as in Algorithm 1: at every round t it is the ridge estimator under the *current* probe B_t applied to *all* past observations (the retroactive rebuild of Algorithm 1; between refits the incremental update with $\Psi_s = B_{t_k}\tilde{H}_s$ coincides with this rebuild). Writing $V_t := \lambda I_r + \sigma_{\mathrm{obs}}^{-2} \sum_{s\le t} B_t \tilde{H}_s \tilde{H}_s^\top B_t^\top = \lambda I_r + \sigma_{\mathrm{obs}}^{-2} B_t \Gamma_t B_t^\top$ and $\hat{\phi}_i^{(t)} := V_t^{-1} \sigma_{\mathrm{obs}}^{-2} \sum_{s\le t} B_t \tilde{H}_s \tilde{U}_{i,s}$, note that V_t is identical for all users (under (T1) all N users observe the same played plan).

Lemma 1 (Projection inequality)**.** *Let $Z \in \mathbb{R}^{r\times t}$, $\lambda > 0$, $\sigma > 0$, and $V := \lambda I_r + \sigma^{-2} ZZ^\top$. Then for every $v \in \mathbb{R}^t$, $\|Zv\|_{V^{-1}} \le \sigma \|v\|_2$.*

Proof. Let $ZZ^\top = U\Lambda U^\top$ with $\Lambda \succ 0$ on $\mathrm{range}(Z)$. Since $Zv \in \mathrm{range}(Z)$ and V acts as $\lambda I + \sigma^{-2}\Lambda$ there,

$$\begin{aligned} \|Zv\|_{V^{-1}}^2 &= v^\top Z^\top U (\lambda I + \sigma^{-2}\Lambda)^{-1} U^\top Z v \le \sigma^2\, v^\top Z^\top U \Lambda^{-1} U^\top Z v \\ &= \sigma^2\, v^\top \big[Z^\top (ZZ^\top)^+ Z \big] v \le \sigma^2 \|v\|^2, \end{aligned}$$

because $Z^\top (ZZ^\top)^+ Z$ is the orthogonal projector onto $\mathrm{row}(Z)$. □

Lemma 2 (Optimal rotation)**.** *Let $B_1, B_2 \in \mathbb{R}^{r\times k}$ be row-orthonormal. There exists $R \in O(r)$ with*

$$\|RB_1 - B_2\|_{\mathrm{op}} \;\le\; \sqrt{2}\, \sin\Theta\big(\mathrm{row}(B_1), \mathrm{row}(B_2)\big). \tag{4}$$

Consequently, with $\bar{B}_0 := (B_0 B_0^\top)^{-1/2} B_0$ (row-orthonormal, $\mathrm{row}(\bar{B}_0) = \mathrm{row}(B_0)$) and $Q_t \in O(r)$ the rotation of (4) for the pair $(\bar{B}_0, B_t)$,

$$\|Q_t B_0 - B_t\|_{\mathrm{op}} \;\le\; \sqrt{2}\, \sin\Theta_t + \varepsilon_h^2 \;=\; \theta_t. \tag{5}$$

Proof. Let $B_2 B_1^\top = U\Sigma V^\top$ be an SVD; the singular values are $\cos\vartheta_1, \dots, \cos\vartheta_r$, the cosines of the principal angles between the two row-spaces. Take $R := UV^\top$. Then $RB_1 B_2^\top = U\Sigma U^\top$ and

$$(RB_1 - B_2)(RB_1 - B_2)^\top = 2I_r - RB_1 B_2^\top - B_2 B_1^\top R^\top = U(2I_r - 2\Sigma)U^\top,$$

whose eigenvalues are $2(1-\cos\vartheta_i) = 4\sin^2(\vartheta_i/2) \le 2\sin^2\vartheta_i$ for $\vartheta_i \in [0, \pi/2]$. This gives (4). For (5), write $B_0 B_0^\top = I_r - E$ with $0 \preceq E \preceq \varepsilon_h^2 I_r$ by (A5); for $\varepsilon_h^2 \le 1/2$, $\|(I-E)^{-1/2} - I\|_{\mathrm{op}} \le \varepsilon_h^2$, hence $\|\bar{B}_0 - B_0\|_{\mathrm{op}} = \|((I-E)^{-1/2} - I)B_0\|_{\mathrm{op}} \le \varepsilon_h^2$, and $\|Q_t B_0 - B_t\|_{\mathrm{op}} \le \|Q_t\bar{B}_0 - B_t\|_{\mathrm{op}} + \|Q_t(B_0 - \bar{B}_0)\|_{\mathrm{op}} \le \sqrt{2}\sin\Theta_t + \varepsilon_h^2$. □

Lemma 3 (Per-user ridge concentration under the current $\hat{W}_t$)**.** *Assume (A1), (A3)(i), (A5). Let $t \ge T_0$ and suppose the probe in force at round t satisfies the* basin condition

$$\theta_t \;\le\; \frac{\gamma_h^{\mathrm{play}}}{8\, H_{\mathrm{plan}}^2}. \tag{6}$$

Let Q_t be the rotation of Lemma 2 and $\tilde{\phi}_i^{(t)} := Q_t \phi_i^\star$ (the truth expressed in the gauge of B_t; utilities are gauge-invariant). Then:

(a) (Deterministic.) $V_t \succeq \frac{1}{2} V_t^\star$ with $V_t^\star := \lambda I_r + \sigma_{\mathrm{obs}}^{-2} Q_t B_0 \Gamma_t B_0^\top Q_t^\top \succeq \gamma_h^{\mathrm{play}} t/(2\sigma_{\mathrm{obs}}^2)\, I_r$; hence for every $z \in \mathcal{X}$, $\sigma_i^{(t)}(z) := \|B_t \tilde{H}(z)\|_{V_t^{-1}} \le 2\sigma_{\mathrm{obs}} H_{\mathrm{plan}} / \sqrt{\gamma_h^{\mathrm{play}} t}$.

(b) With probability at least $1-\delta/2$, simultaneously for all $i \in [N]$ and all $t \le ET$ at which (6) holds,

$$\big\|\hat{\phi}_i^{(t)} - \tilde{\phi}_i^{(t)}\big\|_{V_t} \;\le\; \beta_t^{\mathrm{full}} := \beta_t^{\mathrm{noise}} + \beta_t^{W} + \beta_t^{\varepsilon} + \beta_t^{\times},$$

where

$$\beta_t^{\mathrm{noise}} := \sqrt{2}\,\sqrt{r\log\!\big(1 + tH_{\mathrm{plan}}^2/(\lambda\sigma_{\mathrm{obs}}^2 r)\big) + 2\log(4N/\delta)} + \sqrt{\lambda}\,S,$$
$$\beta_t^{W} := SH_{\mathrm{plan}}\,\theta_t\sqrt{t}/\sigma_{\mathrm{obs}}, \qquad \beta_t^{\varepsilon} := \varepsilon_{\mathrm{eff}}\sqrt{t}/\sigma_{\mathrm{obs}},$$
$$\beta_t^{\times} := 2\sqrt{2}\,\theta_t H_{\mathrm{plan}}\sqrt{k_h\log(8k_hNET/\delta)/\gamma_h^{\mathrm{play}}},$$

the self-normalized width on the fixed features $B_0\tilde{H}_s$, the W-bias, the misspecification radius, and a noise×subspace-error cross term, respectively. Since β_t^{full} is increasing in θ_t, the bound also holds with θ_t replaced by any $\bar{\theta}_t \ge \theta_t$; we write $\bar{\beta}_t^{\mathrm{full}}$ for the radius so obtained from the envelope $\bar{\theta}_t$ of (A4)(ii).

Proof of Lemma 3. (a) Gram comparison. Write $\Delta_t := B_t - Q_tB_0$, so $\|\Delta_t\|_{\mathrm{op}} \le \theta_t$ by (5). Since $\|B_t\|_{\mathrm{op}} = 1$, $\|Q_tB_0\|_{\mathrm{op}} \le 1$ and $\|\Gamma_t\|_{\mathrm{op}} \le H_{\mathrm{plan}}^2 t$,

$$\|B_t\Gamma_tB_t^\top - Q_tB_0\Gamma_tB_0^\top Q_t^\top\|_{\mathrm{op}} = \|\Delta_t\Gamma_tB_t^\top + Q_tB_0\Gamma_t\Delta_t^\top\|_{\mathrm{op}} \le 2\theta_t H_{\mathrm{plan}}^2 t.$$

By (A3)(i), $\Gamma_t \succeq \gamma_h^{\mathrm{play}} t\, I_{k_h}$ for $t \ge T_0$, and $B_0B_0^\top \succeq (1-\varepsilon_h^2)I_r \succeq \frac{1}{2}I_r$, so $\lambda_{\min}(V_t^\star) \ge \lambda + \gamma_h^{\mathrm{play}} t/(2\sigma_{\mathrm{obs}}^2)$. Under (6), $2\theta_t H_{\mathrm{plan}}^2 t/\sigma_{\mathrm{obs}}^2 \le \gamma_h^{\mathrm{play}} t/(4\sigma_{\mathrm{obs}}^2) \le \frac{1}{2}\lambda_{\min}(V_t^\star)$, hence $V_t \succeq V_t^\star - \frac{1}{2}\lambda_{\min}(V_t^\star)I_r \succeq \frac{1}{2}V_t^\star$. The width bound follows from $\|B_t\tilde{H}(z)\| \le H_{\mathrm{plan}}$ and $\lambda_{\min}(V_t) \ge \gamma_h^{\mathrm{play}} t/(4\sigma_{\mathrm{obs}}^2)$.

(b) Error decomposition. Using (A1) and (A5), rewrite the observation in the gauge of B_t:

$$\tilde{U}_{i,s} = \tilde{\phi}_i^{(t)\top} B_t\tilde{H}_s + m_{i,s}^{(t)} + \eta_{i,s} + \nu'_{i,s}, \qquad m_{i,s}^{(t)} := \tilde{\phi}_i^{(t)\top}(Q_tB_0 - B_t)\tilde{H}_s,$$

where $\nu'_{i,s} := \nu_{i,s} + \phi_i^{\star\top} W^\star(I - PP^\top)H(x_s)$ satisfies $|\nu'_{i,s}| \le \varepsilon_{\mathrm{eff}}$ (the (A5) tail is at most $SH_{\mathrm{plan}}\varepsilon_h$) and $|m_{i,s}^{(t)}| \le S\,\theta_t\,H_{\mathrm{plan}}$ by (5). Writing $Z_t := B_t[\tilde{H}_1, \dots, \tilde{H}_t] \in \mathbb{R}^{r\times t}$ (so $V_t = \lambda I + \sigma_{\mathrm{obs}}^{-2} Z_t Z_t^\top$) and $\eta_i, \nu'_i, m_i^{(t)} \in \mathbb{R}^t$ for the corresponding vectors,

$$\hat{\phi}_i^{(t)} - \tilde{\phi}_i^{(t)} = V_t^{-1}\Big[\sigma_{\mathrm{obs}}^{-2} Z_t\big(\eta_i + \nu'_i + m_i^{(t)}\big) - \lambda\tilde{\phi}_i^{(t)}\Big],$$

so

$$\|\hat{\phi}_i^{(t)} - \tilde{\phi}_i^{(t)}\|_{V_t} \le \sigma_{\mathrm{obs}}^{-2}\|Z_t\eta_i\|_{V_t^{-1}} + \sigma_{\mathrm{obs}}^{-2}\|Z_t\nu'_i\|_{V_t^{-1}} + \sigma_{\mathrm{obs}}^{-2}\|Z_t m_i^{(t)}\|_{V_t^{-1}} + \sqrt{\lambda}\,S.$$

Bias terms. By Lemma 1, $\sigma_{\mathrm{obs}}^{-2}\|Z_t m_i^{(t)}\|_{V_t^{-1}} \le \|m_i^{(t)}\|_2/\sigma_{\mathrm{obs}} \le SH_{\mathrm{plan}}\theta_t\sqrt{t}/\sigma_{\mathrm{obs}} = \beta_t^W$ and $\sigma_{\mathrm{obs}}^{-2}\|Z_t\nu'_i\|_{V_t^{-1}} \le \|\nu'_i\|_2/\sigma_{\mathrm{obs}} \le \varepsilon_{\mathrm{eff}}\sqrt{t}/\sigma_{\mathrm{obs}} = \beta_t^{\varepsilon}$; both bounds are deterministic and hold for any B_t.

Noise term. Let $\xi_{i,t} := \sum_{s\le t}\tilde{H}_s\eta_{i,s} \in \mathbb{R}^{k_h}$, so $Z_t\eta_i = B_t\xi_{i,t} = Q_tB_0\xi_{i,t} + \Delta_t\xi_{i,t}$. For the first piece, part (a) gives $\|\cdot\|_{V_t^{-1}} \le \sqrt{2}\|\cdot\|_{(V_t^\star)^{-1}}$, and $V_t^\star = Q_tV_t^0Q_t^\top$ with $V_t^0 := \lambda I_r + \sigma_{\mathrm{obs}}^{-2}B_0\Gamma_tB_0^\top$, so

$$\sigma_{\mathrm{obs}}^{-2}\|Q_tB_0\xi_{i,t}\|_{V_t^{-1}} \le \sqrt{2}\,\sigma_{\mathrm{obs}}^{-2}\|B_0\xi_{i,t}\|_{(V_t^0)^{-1}} = \sqrt{2}\,\Big\|\sum_{s\le t}\tfrac{B_0\tilde{H}_s}{\sigma_{\mathrm{obs}}}\cdot\tfrac{\eta_{i,s}}{\sigma_{\mathrm{obs}}}\Big\|_{(V_t^0)^{-1}}.$$

The features $B_0\tilde{H}_s = W^\star PP^\top H(x_s)$ are $\mathcal{F}_{s-1}$-measurable (B_0 is fixed and x_s is chosen before $\eta_{i,s}$ is observed), $\|B_0\tilde{H}_s\| \le H_{\mathrm{plan}}$, and $\eta_{i,s}/\sigma_{\mathrm{obs}}$ is 1-sub-Gaussian, so Abbasi-Yadkori et al. (2011, Thm. 2) gives, with probability $\ge 1-\delta/(4N)$ simultaneously for all t, the bound $\sqrt{2\log(4N/\delta) + r\log(1 + tH_{\mathrm{plan}}^2/(\lambda\sigma_{\mathrm{obs}}^2 r))}$; together with $\sqrt{\lambda}S$ this is β_t^{noise}. For the second piece, $\|\Delta_t\xi_{i,t}\|_{V_t^{-1}} \le \|\Delta_t\xi_{i,t}\|_2/\sqrt{\lambda_{\min}(V_t)} \le 2\sigma_{\mathrm{obs}}\theta_t\|\xi_{i,t}\|_2/\sqrt{\gamma_h^{\mathrm{play}}t}$ by (a). Each coordinate of $\xi_{i,t}$ is a martingale with σ_{obs}-sub-Gaussian increments of size at most H_{plan}, so a union bound over the k_h coordinates, N users, and ET rounds gives $\|\xi_{i,t}\|_2 \le \sigma_{\mathrm{obs}}H_{\mathrm{plan}}\sqrt{2k_ht\log(8k_hNET/\delta)}$ for all (i,t) with probability $\ge 1-\delta/4$. Hence $\sigma_{\mathrm{obs}}^{-2}\|\Delta_t\xi_{i,t}\|_{V_t^{-1}} \le 2\sqrt{2}\,\theta_tH_{\mathrm{plan}}\sqrt{k_h\log(8k_hNET/\delta)/\gamma_h^{\mathrm{play}}} = \beta_t^{\times}$. The two probabilistic events together have probability $\ge 1-\delta/2$. □

Regret decomposition (separate-bound LCB). For each user i and round t, with $\hat{\Psi}_t(z) := B_t\tilde{H}(z) = \hat{W}_t^{\text{full}}H(z)$:

$$\underline{U}_i^{\max}(t) := \max_{z\in\mathcal{X}}\big[\hat{\phi}_i^{(t)\top}\hat{\Psi}_t(z) - \beta_t\sigma_i^{(t)}(z)\big], \quad \overline{U}_i(t)(x) := \hat{\phi}_i^{(t)\top}\hat{\Psi}_t(x) + \beta_t\sigma_i^{(t)}(x),$$

$$\mathrm{LCB}_t(D_i)(x) := \underline{U}_i^{\max}(t) - \overline{U}_i(t)(x), \quad \text{where} \quad \sigma_i^{(t)}(z) := \|\hat{\Psi}_t(z)\|_{V_t^{-1}}, \quad \beta_t := \bar{\beta}_t^{\text{full}}.$$

On the good event of Lemma 3(b) (which applies to $\bar{\beta}_t^{\text{full}} \ge \beta_t^{\text{full}}$), for every $z \in \mathcal{X}$,

$$\begin{aligned}\big|\hat{\phi}_i^{(t)\top}\hat{\Psi}_t(z) - U_i(z)\big| &\le \big|(\hat{\phi}_i^{(t)} - \tilde{\phi}_i^{(t)})^\top B_t\tilde{H}(z)\big| + \big|\tilde{\phi}_i^{(t)\top}(B_t - Q_tB_0)\tilde{H}(z)\big| \\ &\quad + \big|\phi_i^{\star\top}W^\star(I - PP^\top)H(z)\big| + |\delta_i(z)| \;\le\; \beta_t\sigma_i^{(t)}(z) + SH_{\text{plan}}\theta_t + \varepsilon_{\text{eff}},\end{aligned}$$

using Cauchy–Schwarz, (5), and (A5). Hence each bound is valid up to $SH_{\text{plan}}\theta_t + \varepsilon_{\text{eff}}$ ($\underline{U}_i^{\max}(t) \le U_i^{\max} + SH_{\text{plan}}\theta_t + \varepsilon_{\text{eff}}$, $\overline{U}_i(t)(x) \ge U_i(x) - SH_{\text{plan}}\theta_t - \varepsilon_{\text{eff}}$) and has width at most $2\beta_t\sigma_i^{(t)}(\cdot) + SH_{\text{plan}}\theta_t + \varepsilon_{\text{eff}}$ (with $\sup_z \sigma_i^{(t)}(z)$ for $U^{\max}$). Under (T1) the coordinator plays $x_t^{(e)} \in \arg\min_{x\in\mathcal{X}}\max_i \mathrm{LCB}_t(D_i)(x)$; with $x^\star \in \arg\min_x \Phi_e(x)$, the standard optimism chain $\Phi_e(x_t) \le \max_i \mathrm{LCB}_t(D_i)(x_t) + \mathrm{width}(x_t) \le \max_i \mathrm{LCB}_t(D_i)(x^\star) + \mathrm{width}(x_t) \le \Phi_e^\star +$ validity $\mathrm{gap}(x^\star) + \mathrm{width}(x_t)$ gives, for every $t \ge \max(T_0, t_1)$,

$$\Phi_e(x_t^{(e)}) - \Phi_e^\star \;\le\; 2\beta_t\Big[\max_i \sigma_i^{(t)}(x_t^{(e)}) + \max_i \sup_{z\in\mathcal{X}} \sigma_i^{(t)}(z)\Big] \;+\; 4SH_{\text{plan}}\theta_t \;+\; 4\varepsilon_{\text{eff}}. \tag{7}$$

(The factor 4 on θ_t counts the W-feature error in both brackets ($U^{\max}$ and $U(x)$) at each of $x^\star$ and x_t.) For $t < \max(T_0, t_1)$ we use the trivial per-round bound $\Phi_e(x_t) - \Phi_e^\star \le 2\sup_{i,x}|U_i(x)| \le 2(SH_{\text{plan}} + \varepsilon)$.

ϕ-side bound. By Lemma 3(a), both σ-terms in (7) are at most $2\sigma_{\text{obs}}H_{\text{plan}}/\sqrt{\gamma_h^{\text{play}}t}$, so $\sum_{t\le ET} 2[\cdots] \le 16\sigma_{\text{obs}}H_{\text{plan}}\sqrt{ET/\gamma_h^{\text{play}}}$. For $t \in [t_k, t_{k+1})$ with $t_k \ge t_1$, (A4)(ii) gives $\theta_t \le \bar{\theta}_t \le \sqrt{2}(c_6+1)(\rho_\infty(Nt_k) + \rho_\varepsilon) + \varepsilon_h^2$ with $t/t_k \le 2$, so $\bar{\theta}_t\sqrt{t} \le c\,\sigma_{\text{obs}}\sqrt{rk_h\log(d_h/\delta)/(N\gamma_\phi\gamma_h^{\text{play}})} + c\,(\varepsilon_{\text{eff}}/\sqrt{\gamma_\phi\gamma_h^{\text{play}}} + \varepsilon_h^2)\sqrt{t}$ (a constant plus a $\sqrt{t}$ floor). Multiplying the four radii of $\bar{\beta}_t^{\text{full}}$ against the σ-side aggregation:

- β^{noise}: $\tilde{O}\big(\sigma_{\text{obs}}H_{\text{plan}}\sqrt{rET\log(N/\delta)/\gamma_h^{\text{play}}}\big)$ — Term 1 of (3); the ridge-warmup piece $\sqrt{\lambda}S$ contributes $O(\sqrt{\lambda}SH_{\text{plan}}\sigma_{\text{obs}}\sqrt{ET/\gamma_h^{\text{play}}})$, dominated when $r\log(N/\delta) \ge \lambda S^2$.
- $\bar{\beta}_t^W = SH_{\text{plan}}\bar{\theta}_t\sqrt{t}/\sigma_{\text{obs}}$: the constant part gives $\tilde{O}\big(SH_{\text{plan}}^2\sigma_{\text{obs}}\sqrt{rk_hET/(N\gamma_\phi(\gamma_h^{\text{play}})^2)}\big)$ — Term 2; the $\sqrt{t}$ floor, together with $\beta_t^\varepsilon = \varepsilon_{\text{eff}}\sqrt{t}/\sigma_{\text{obs}}$, cancels the $t^{-1/2}$ width and gives per round $O\big(\varepsilon_{\text{eff}}H_{\text{plan}}/\sqrt{\gamma_h^{\text{play}}} + SH_{\text{plan}}^2\varepsilon_{\text{eff}}/(\gamma_h^{\text{play}}\sqrt{\gamma_\phi})\big)$ (the $SH_{\text{plan}}\varepsilon_h^2$ piece is $\le \varepsilon_{\text{eff}}$), hence Term 3 after summing over ET rounds — linear in ET, consistent with the $\Omega(\varepsilon\sqrt{d}\,T)$ misspecified-linear-bandit lower bound of Lattimore et al. (2020).
- $\bar{\beta}_t^\times = 2\sqrt{2}\,\bar{\theta}_tH_{\text{plan}}\sqrt{k_h\log(8k_hNET/\delta)/\gamma_h^{\text{play}}}$: the $t^{-1/2}$-decaying part of $\bar{\theta}_t$ gives $\sum_t t^{-1}$, i.e. $\tilde{O}\big(\sigma_{\text{obs}}^2H_{\text{plan}}^2k_h\sqrt{r}\log(ET)/((\gamma_h^{\text{play}})^{3/2}\sqrt{N\gamma_\phi})\big)$; the floor part gives $\tilde{O}\big(\varepsilon_{\text{eff}}\sigma_{\text{obs}}H_{\text{plan}}^2\sqrt{k_hET}/(\gamma_h^{\text{play}}\sqrt{\gamma_\phi\gamma_h^{\text{play}}})\big)$, sublinear in ET. Both are collected in R_{lo}.

W-side bound. By (A4)(ii) (the refit uses the current $\{\hat{\phi}_i^{(t_k)}\}$ in place of $\{\phi_i^\star\}$; the estimated participant Gram stays $\succeq (\gamma_\phi/2)I_r$ under the Gram condition of (A4)(ii), since $\max_i\|\hat{\phi}_i^{(t_k)} - Q_{t_k}\phi_i^\star\| \le \bar{e}_kS$ perturbs it by at most $O(\bar{e}_kS^2) \le \gamma_\phi/4$), $\theta_t \le \bar{\theta}_t \le \sqrt{2}(c_6+1)(\rho_\infty(Nt_k) + \rho_\varepsilon) + \varepsilon_h^2$ on $[t_k, t_{k+1})$ for $t_k \ge t_1$, so

$$4SH_{\text{plan}}\sum_{t\le ET}\theta_t \;\le\; c\,SH_{\text{plan}}\sigma_{\text{obs}}\sqrt{\tfrac{rk_hET\log(d_h/\delta)}{N\gamma_\phi\gamma_h^{\text{play}}}} + c\,SH_{\text{plan}}\Big(\tfrac{\varepsilon_{\text{eff}}}{\sqrt{\gamma_\phi\gamma_h^{\text{play}}}} + \varepsilon_h^2\Big)ET,$$

(the sum over refits gives $\sum_k K_{\text{refit}}/\sqrt{Nt_k} \asymp \sqrt{ET/N}$). Since $\gamma_h^{\text{play}} \le \lambda_{\max}(P^\top\Sigma_hP) \le H_{\text{plan}}^2$, i.e. $H_{\text{plan}}/\sqrt{\gamma_h^{\text{play}}} \ge 1$, both pieces are dominated by Term 2 and Term 3 respectively.

Burn-in. Rounds $t < \max(T_0, t_1)$ contribute at most $2(SH_{\text{plan}} + \varepsilon)\max(T_0, t_1) = O(SH_{\text{plan}}(T_0 + t_1))$, a constant independent of ET, collected in R_{lo}; from the definition of t_1 in (A4)(ii), $t_1 = O\big(K_{\text{refit}}\log\big(\rho_0/(\rho_\infty(NET) + \rho_\varepsilon)\big) + \sigma_{\text{obs}}^2rk_hH_{\text{plan}}^4\log(d_h/\delta)/(N\gamma_\phi(\gamma_h^{\text{play}})^3)\big)$, where the first term is the $O(\log(1/\rho_0))$ refits (at worst $O(\log ET)$) needed to contract the warm-start error below the statistical floor.

Combining. Adding the ϕ-side, W-side, and burn-in bounds and absorbing polylogarithmic factors into $\tilde{O}$ yields Equation (3) with

$$R_{\mathrm{lo}} = \tilde{O}\left(\frac{\sigma_{\mathrm{obs}}^2 H_{\mathrm{plan}}^2 k_h \sqrt{r}}{(\gamma_h^{\mathrm{play}})^{3/2}\sqrt{N\gamma_\phi}}\right) + \tilde{O}\left(\frac{\varepsilon_{\mathrm{eff}}\,\sigma_{\mathrm{obs}} H_{\mathrm{plan}}^2 \sqrt{k_h ET}}{\gamma_h^{\mathrm{play}}\sqrt{\gamma_\phi \gamma_h^{\mathrm{play}}}}\right) + O\big(SH_{\mathrm{plan}}(T_0 + t_1)\big).$$

The failure probability is at most δ: four events of probability $\leq \delta/4$ each: (A4)(i), (A4)(ii) over all refits, and the two events of Lemma 3(b). $\square$