

Offline Clustering of Preference Learning with Active-data Augmentation

JINGYUAN LIU, Nanjing University, China

FATEMEH GHAFARI, University of Massachusetts Amherst, USA

XUCHUANG WANG, University of Massachusetts Amherst, USA

XUTONG LIU, University of Washington, USA

MOHAMMAD HAJIESMAILI, University of Massachusetts Amherst, USA

CARLEE JOE-WONG, Carnegie Mellon University, USA

Preference learning from pairwise feedback is a widely adopted framework in practical applications such as reinforcement learning with human feedback (RLHF) and personalized recommendations. In many practical settings, however, user interactions are limited or costly (e.g. in safety-critical domains), making offline preference learning a necessary alternative. Moreover, real-world preference learning often involves heterogeneous users with different preferences. For example, annotators from different backgrounds may rank the same responses differently. This setting presents two central challenges: (1) identifying similarity across users to enable effective data aggregation, especially under scenarios where offline data is imbalanced across dimensions, and (2) handling the imbalanced offline data where some preference dimensions are underrepresented which leads unreliable estimation. To address these challenges, we study the Offline Clustering of Preference Learning problem, where the learner has access to fixed datasets from multiple users with potentially different preferences and aims to maximize utility for a test user. To tackle the first challenge, we first propose Off-C²PL for the pure offline setting, where the learner relies solely on offline data. Our theoretical analysis provides a suboptimality bound that explicitly captures the tradeoff between sample noise (arising from limited data) and bias (introduced by aggregating heterogeneous users). To address the second challenge of insufficient data coverage, we extend our framework to the setting with active-data augmentation where the learner is allowed to select a limited number of additional active-data for the test user based on the cluster structure learned by Off-C²PL. In this setting, our second algorithm, A²-Off-C²PL, actively selects samples that target the least-informative dimensions of the test user's preference. We prove that these actively collected samples contribute more effectively than offline ones, improving the suboptimality. Finally, we validate our theoretical results through simulations on synthetic and real-world datasets, demonstrating improvements of 61.47% in the pure offline setting and 57.51% in the setting with active-data augmentation over the closest baselines.

Additional Key Words and Phrases: Offline preference learning, clustering of bandits, dueling bandits

ACM Reference Format:

Jingyuan Liu, Fatemeh Ghaffari, Xuchuang Wang, Xutong Liu, Mohammad Hajiesmaili, and Carlee Joe-Wong. 2025. Offline Clustering of Preference Learning with Active-data Augmentation. 1, 1 (October 2025), 40 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Learning human preferences is a fundamental building block of modern AI systems. Whether aligning large language models (LLMs) with human values [3, 51], recommending movies or

Authors' addresses: Jingyuan Liu, Nanjing University, China, jingyuanliu@smail.nju.edu.cn; Fatemeh Ghaffari, University of Massachusetts Amherst, USA, fghaffari@umass.edu; Xuchuang Wang, University of Massachusetts Amherst, USA, xuchuangwang@cs.umass.edu; Xutong Liu, University of Washington, USA, xutongli@uw.edu; Mohammad Hajiesmaili, University of Massachusetts Amherst, USA, hajiesmaili@cs.umass.edu; Carlee Joe-Wong, Carnegie Mellon University, USA, cjoewong@andrew.cmu.edu.

products [2, 80], or personalizing digital assistants [49, 65], systems must understand not only what actions are available but which ones people actually prefer. Unlike traditional supervised learning tasks [27, 67] with clear ground-truth labels, preference learning must infer the subjective and often heterogeneous nature of human choices. In the examples above, a model that fails to capture preferences may generate fluent yet misaligned LLM outputs, or recommend items that frustrate rather than engage users.

A practical way to elicit such preferences is through pairwise feedback: rather than assigning absolute scores, users (or annotators) simply indicate which of two options they prefer. Pairwise comparisons are natural in practice: for instance, evaluators can more easily judge which of two LLM responses is better instead of assigning absolute scores to possible responses, and users often reveal preferences implicitly by choosing one product over another. This learning framework has been extensively modeled and studied under the dueling bandits problem, which uses sequences of pairwise comparisons to infer underlying preference structures [5, 20, 57, 58, 82].

Despite significant progress, most prior work assumes a single, shared preference vector, overlooking the fact that preferences vary across users in real-world applications. In LLM alignment, for instance, annotators from different backgrounds or cultures may rank responses differently. In recommendation, users routinely disagree on the same items. If we aggregate all feedback indiscriminately, the result is a one-size-fits-none policy. If we treat each user separately, limited per-user data leads to poor learning. The natural solution is to cluster users with similar preferences: pooling their data to increase sample sufficiency while preserving personalization.

However, clustering similar users becomes particularly challenging in the offline preference learning, where the learner has access only to fixed, pre-collected datasets of pairwise comparisons rather than interactive feedback that enables more accurate preference estimation. This setting is increasingly relevant in practice: in LLM alignment, reinforcement learning from human feedback (RLHF) [3, 10, 32] often relies on static datasets of human comparisons between possible responses, while in recommender systems [2, 21, 80], historical user logs provide pairwise evidence of preferences between different items. In both cases, the learner must leverage existing data to select actions that maximize user utility (i.e., satisfaction) from this fixed, given dataset [17, 37, 45, 89].

Motivated by this gap, we study the problem of *Offline Clustering of Preference Learning*, where U users are partitioned into J clusters. Users in the same cluster share a common preference vector, while those in different clusters do not. Each user has a fixed offline dataset of pairwise comparisons, where for a given context (the input condition or situation, e.g., a prompt in RLHF or user profile in recommendation), the user provides a binary preference between two candidate actions. We assume preferences follow the Bradley–Terry–Luce (BTL) model [7, 18]. The goal is to identify users that have similar preferences as the test user and aggregate their data to increase sample sufficiency to learn a personalized policy that selects actions with near-optimal expected reward.

This setting presents two central challenges: (1) *Identifying similarity across users without coverage assumptions*: A central challenge in leveraging offline data from users with potentially different preferences is to identify users that are similar to the test user. Prior works on clustering of bandits [22, 38, 39, 42, 44, 72, 73] typically rely on an *item regularity* assumption, which requires offline actions to provide balanced information across all preference dimensions, ensuring sufficient data coverage but limiting generality. However, this assumption becomes unrealistic in our setting with pairwise feedback [72], where the pairwise feedback may be interdependent and may distort data coverage. Thus, our problem demands identifying user similarity directly from imbalanced and potentially incomplete offline data, without relying on any coverage guarantees. This leads to the second challenge. (2) *Handling imbalanced offline data*: A natural way to mitigate imbalanced offline data is to collect new data samples of strategically chosen preferences, e.g., in a LLM setting, users may be presented with two carefully chosen responses and asked to indicate their preference

Table 1. Summary of main and additional theoretical results.

Comparisons of Algorithms for Pure Offline Model				
	Algorithm	Setting	Condition	Suboptimality
Previous [37, 89]	P-MLE [89] PDC [37]	Pure Offline Single User	—	$\tilde{O}\left(\sqrt{\frac{d}{\lambda_1}}\right)$
Main Result 1 (Theorem 1)			—	$\tilde{O}\left(\frac{\sqrt{d}(1+\hat{\gamma}\sqrt{N_1})}{\sqrt{\lambda_2}}\right)$
Additional Result 1 (Equation (9))	Off-C ² PL (Algorithm 1)	Pure Offline Multiple Users	Lower Threshold $\hat{\gamma} \leq \gamma$ (Definition 1)	$\tilde{O}\left(\sqrt{\frac{d}{\lambda_2}}\right)$
Additional Result 2 (Corollary 1)			Item Regularity (Assumption 1)	$\tilde{O}\left(\sqrt{\frac{d}{\lambda_a}}\left(\sqrt{\frac{1}{N_3}} + \hat{\gamma}\sqrt{\eta_1}\right)\right)$
Comparisons of Algorithms for Active-data Augmented Model				
	Algorithm	Setting	Condition	Suboptimality
Previous [17]	APO [17]	Pure Active Single User	—	$\tilde{O}\left(\frac{d}{\sqrt{N}}\right)$
Main Result 2 (Theorem 2)			—	$\tilde{O}\left(\frac{\sqrt{d}(1+\hat{\gamma}\sqrt{N_1})}{\sqrt{\lambda_3+N/d}}\right)$
Additional Result 3 (Corollary 2)	A ² -Off-C ² PL (Algorithm 2)	Hybrid (Offline + Active) Multiple Users	Imbalanced Dataset (Definition 2)	$\tilde{O}\left(\frac{\sqrt{d}(1+\hat{\gamma}\sqrt{N_1})}{\sqrt{\lambda_2+N}}\right)$
Additional Result 4 (Corollary 3)			Item Regularity (Assumption 1) + Imbalanced Dataset (Definition 2)	$\tilde{O}\left(\sqrt{\frac{d}{\lambda_a}}\left(\sqrt{\frac{1}{N_3}} + \hat{\gamma}\sqrt{\eta_2}\right)\right)$

Here, d denotes the dimension of each user's preference vector. λ_1 , λ_2 , and λ_3 represent the minimum eigenvalue of the (regularized) information matrix constructed from (i) the test user's offline data only, (ii) the test user's offline data combined with aggregated data from clustered neighbors, and (iii) case (ii) further augmented with N actively selected samples for the test user, respectively. $\tilde{\lambda}_a$ is the smoothed item regularity parameter, which lower bounds the information matrix in terms of the number of samples used. N_1 denotes the number of heterogeneous offline samples included, N_2 the total number of offline samples used, and N_3 the total number of samples combining offline and active data. Finally, $\eta_1 = N_1/N_2$ and $\eta_2 = N_1/N_3$ represent the fraction of heterogeneous samples among all offline samples and among the combined offline-active datasets, respectively.

between them, before receiving the final LLM response. Active learning approaches [11, 17, 26, 40, 48, 61] mitigate imbalance by querying new comparisons, but they assume fully interactive querying rather than the hybrid offline-active regime considered here. Hence, a key challenge is to effectively integrate actively collected data with fixed offline datasets, ensuring that new samples complement rather than exacerbate the imbalance in coverage across preference dimensions.

To address these challenges, we focus on two central research questions: (1) *Can we effectively identify users with similar preferences, especially under fixed and imbalanced offline data without relying on coverage assumptions?* (2) *How can we actively collect additional data to mitigate the impact of poor coverage in imbalanced offline datasets that fail to represent all preference dimensions?*

Table 1 summarizes the main contributions of our paper (with the key notations introduced at the bottom of the table). We highlight four key **contributions** as follows:

(i) Model Formulations: We are the first to introduce the *Offline Clustering of Preference Learning* framework, where the learner needs to learn heterogeneous user preferences from offline pairwise feedback, without assuming any data coverage assumption. This setting naturally leads to the two core challenges discussed earlier: identifying user similarity and handling imbalanced offline data. To formalize the problem, we first present the pure offline model, followed by its extended model with active-data augmentation. In the pure offline model, the learner relies solely on the fixed offline datasets to infer each user's preferences, cluster users with similar preferences, and aggregate their data to improve estimation accuracy. This reflects realistic scenarios such as aligning large language models using RLHF datasets collected from annotators across different regions, or personalizing recommendations from logged data of diverse user populations. Based on this, the active-data augmented model allows the learner to actively acquire a fixed number of additional samples to refine the estimation for the test user, while still leveraging the offline data. This setting captures

practical cases like requesting a small number of extra comparisons from annotators in RLHF, or collecting additional pairwise feedback from users in recommender systems.

(ii) Algorithm and Results for Pure Offline Model: In order to address the challenge of identifying similar users, we develop the first algorithm, Off-C²PL (Algorithm 1) for the pure offline model. Off-C²PL constructs confidence interval on preference estimation for each user based on the **minimum eigenvalue of each user's information matrix**, which captures the least informative dimension, and applies Maximum Likelihood Estimation (MLE) under the BTL model to estimate preferences. This design ensures that the confidence interval directly reflects data sufficiency and estimation accuracy without requiring any coverage assumption. A **clustering threshold parameter** $\hat{\gamma}$ is then used to determine similarity across users: intuitively, $\hat{\gamma}$ balances inclusiveness of clusters against the risk of aggregating heterogeneous users whose preferences are different with the test user. Building on this structure, the algorithm aggregates data across identified clusters to improve estimation. Main Result 1 in Table 1 shows that Off-C²PL achieves a suboptimality of $\tilde{O}((\sqrt{d} + \hat{\gamma}\sqrt{dN_1})/\sqrt{\lambda_2})$, where d is the preference dimension, N_1 the number of heterogeneous samples utilized, and λ_2 the minimum eigenvalue of the aggregated offline information matrix across those identified similar users. This bound has a numerator representing **noise** (\sqrt{d}) and **bias** ($\hat{\gamma}\sqrt{dN_1}$), and a denominator $\sqrt{\lambda_2}$ that reflects the **information gain from aggregating samples of similar users** (as determined by $\hat{\gamma}$). A smaller $\hat{\gamma}$ enforces stricter similarity, reducing N_1 but also lowering λ_2 , while a larger $\hat{\gamma}$ has the opposite effect. This quantifies the tradeoff in setting $\hat{\gamma}$. With a proper choice of $\hat{\gamma}$, the bias term can be eliminated (Additional Result 1), yielding guarantees that improve upon single-user baselines relying only on test user data [37, 89]. Further, by analyzing the item regularity assumption [22, 44, 72, 73] as a special case, Additional Result 2 highlights more clearly the balance between reducing noise and bias, which extends prior offline clustering of bandits result in traditional linear reward [44] to our setting with pairwise feedback.

(iii) Algorithm and Results for Active-data Augmented Model: Building on the structure learned by Off-C²PL, we introduce A²-Off-C²PL under the active-data augmented model, which extends Off-C²PL to address the imbalance of offline datasets. A²-Off-C²PL actively selects contexts and action pairs that maximize information gain along **the least-covered dimensions of the test user's information matrix**, thereby strengthening the weakest directions of the data. This active design yields significantly improved theoretical performance compared with only using pure offline data, as established in the following results. Main Result 2 shows that A²-Off-C²PL achieves suboptimality $\tilde{O}((\sqrt{d} + \hat{\gamma}\sqrt{dN_1})/\sqrt{\lambda_3 + N/d})$, where λ_3 is the minimum eigenvalue of the information matrix combining aggregated pure offline data from Off-C²PL with the N actively selected samples. Compared to Main Result 1, this active augmentation improves the suboptimality gap in two ways: (1) by **directly adding N new active samples**, which contributes an additional N/d term in the denominator; and (2) by **increasing the minimum eigenvalue of the information matrix** from λ_2 to λ_3 through targeted sampling of underrepresented directions. As formalized in Lemma 5 and Additional Result 3, when the offline data is imbalanced and performance is bottlenecked by a few weak dimensions, each active sample can be **as valuable as up to d equivalent offline samples**, yielding an additional N term in the denominator compared to the pure offline case (Main Result 1). Finally, Additional Result 4 demonstrates the further benefits of active augmentation under the item regularity assumption, where the bias is more tightly controlled, yielding performance that strictly outperforms the pure offline case with item regularity assumption (Additional Result 2).

(iv) Empirical Validation: We run experiments on a synthetic benchmark and on the Reddit TL;DR dataset. In the offline setting, we vary the number of samples per user from 10% to 100% of the available data and report the suboptimality gap. In this setting, Off-C²PL consistently achieves the lowest gap, leveraging cross-user information within clusters, especially when samples are

scarce. The improvements are 61.47% over KMeans and 80.07% over Off-DBSCAN. In the setting with active-data augmentation, each method is warm started with 20% of the data, followed by 500 rounds of learning. A²-Off-C²PL outperforms an online-only algorithm APO [17] and Off-C²PL with only random-data augmentation baseline by 87.58% and 57.51%, respectively.

This paper is organized as follows: We review crucial related works in Section 2. In Section 3, we introduce the offline clustering of preference learning problem along with its two settings: the pure offline setting and the active-data augmented setting. We then present the algorithm design and theoretical analysis for the pure offline model in Section 4, followed by those for active-data augmented model in Section 5. Finally, we validate our theoretical findings through experiments on both synthetic and real-world datasets in Section 6, and conclude the paper in Section 7.

2 RELATED WORKS

Offline RL and Bandit Learning. Offline statistical learning [9, 87] primarily focuses on parameter estimation, while offline reinforcement learning (batch RL) extends the scope to sequential decision-making problems using fixed offline datasets [28, 31, 34, 55, 75, 77], and has found wide applications in diverse domains such as dialogue generation [25], autonomous driving [83], educational technologies [63] and personal recommendations [6, 36]. Within this landscape, offline bandits—viewed as a special case of offline RL—extend the multi-armed bandit framework to learning solely from pre-collected data [62]. Prior studies have considered settings where the offline distributions align with the online reward distributions [4, 8] or where distribution shift arises between them [14, 86]. Among them, studies on offline contextual linear bandits [35, 70] are most closely related to our setting. However, our work goes beyond the standard contextual linear bandits formulation by studying pairwise feedback modeled through a logistic function, and by explicitly leveraging the clustering structure among users’ preferences for more efficient learning.

Preference Learning from Pairwise Feedback. Theoretical studies of preference learning from pairwise feedback trace back to the dueling bandit problem [5, 57, 82] and its extension, the contextual dueling bandit problem [20]. These ideas extend naturally to preference-based reinforcement learning [13, 71, 79, 85]. Recent work has emphasized offline preference-based RL, often motivated by reinforcement learning with human feedback (RLHF). Approaches include pessimism-driven methods [43, 84, 89] and KL-regularized formulations [66, 76, 78]. For instance, Xiong et al. [78] study active context selection under strong coverage assumptions, deriving sample-dependent bounds. Beyond RLHF, researchers have explored general preference structures [23, 56, 81], pure active preference learning without offline datasets [17], safety-constrained alignment [69], and sample-efficient learning under limited data [29]. Our work departs from these above mentioned works by explicitly incorporating clustering into pairwise preference learning and combining it with active data augmentation. This introduces two new challenges: (1) reliably inferring clusters from noisy offline comparisons, and (2) selecting informative queries when both contexts and actions matter. Importantly, learning from pairwise feedback provides weaker supervision than full-reward feedback, making these challenges sharper. We address them with algorithms and bounds that reveal the interplay between clustering, data coverage, and active exploration in both pure offline and hybrid settings.

Heterogeneous Preference Learning. Heterogeneous preference learning has been widely studied under the clustering of bandits [22, 38, 39] and multi-task learning [19], where data from users with distinct preference vectors could be used to accelerate learning. Later works investigate privacy [46], model misspecification [73], and robustness to corrupted users [74]. More recent studies by Liu et al. [44] and Wang et al. [72] are closely related to our setting, respectively providing offline and online algorithms for clustering of bandits, whereas we study the preference learning from

pairwise feedback under the offline and active-data augmented settings. With growing interest in RLHF, recent efforts have addressed scenarios involving users with diverse preferences, which are often referred to as personalized RLHF [15, 24, 30, 41, 53, 54]. Theoretically, Liu et al. [45] study heterogeneous user rationality, Zhong et al. [88] focus on meta-learning and social welfare aggregation, and Park et al. [52] analyze representation-based aggregation under assumptions on uniqueness, diversity, and concentrability. Compared to these directions, our work is the first to establish a general clustering-based framework for heterogeneous preference learning without imposing assumptions on the underlying clustering structure or data coverage, and to extend beyond the conventional pure offline setting by incorporating an active-data augmentation mechanism that adaptively improves underrepresented dimensions.

3 SETTING

Notations. Throughout this paper, we use $[s] = \{1, 2, \dots, s\}$ to denote the set of integers from 1 to s . For any matrix $M \in \mathbb{R}^{d \times d}$, we write $\lambda_{\min}(M) = \lambda_1(M)$ to denote its smallest eigenvalue, and $\lambda_i(M)$ to denote its i -th smallest eigenvalue. For vector norms, we use $\|\cdot\|_2$ to denote the Euclidean (ℓ_2) norm, and $\|\cdot\|_M$ to denote the Mahalanobis norm defined with respect to matrix M .

3.1 Problem Formulation

We consider a set of U users, denoted by $\mathcal{U} = [U]$, where each user $u \in \mathcal{U}$ is associated with a preference vector $\theta_u \in \Theta$, with $\Theta := \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq 1\}$. To model preference heterogeneity, the users are partitioned into J clusters ($J \leq U$), where all users within the same cluster $j \in [J]$ share a common preference vector θ^j . Specifically, let $\mathcal{U}(j)$ denote the set of users in cluster j , so that $\mathcal{U} = \bigcup_{j=1}^J \mathcal{U}(j)$ and $\mathcal{U}(j) \cap \mathcal{U}(j') = \emptyset$ for any $j \neq j'$. By construction, users in the same cluster share the same preference vector*, i.e., $\theta_u = \theta_{u'}$ if and only if there exists a cluster j such that $u, u' \in \mathcal{U}(j)$. We further denote by j_u the cluster index to which user u belongs. Note that both the true clustering and the number of clusters are **unknown** to the learner. For a given user u , we refer to users in the same cluster as *homogeneous users* and those in different clusters as *heterogeneous users*.

In the offline preference learning setting, each user $u \in \mathcal{U}$ is provided with an offline dataset $\mathcal{D}_u = \{(\mathbf{x}_u^i, \mathbf{a}_u^i, \mathbf{a}'_u^i, y_u^i)\}_{i=1}^{N_u}$ where N_u denotes the number of samples for each user, and we further define $N_S = \sum_{u \in \mathcal{S}} N_u$ as the total number of samples from all users in a set \mathcal{S} . Within each dataset \mathcal{D}_u , $\mathbf{x}_u^i \in \mathcal{X}$ represents a context for selecting actions (e.g., prompts in RLHF or specific user features in recommendation systems) randomly drawn from the context set \mathcal{X} , and $\mathbf{a}_u^i, \mathbf{a}'_u^i \in \mathcal{A}$ represent a pair of candidate actions (e.g., responses in RLHF or items in recommendation systems) randomly drawn from the action set \mathcal{A} . The binary feedback y_u^i indicates user u 's preference: $y_u^i = 1$ implies that user u prefers action \mathbf{a}_u^i over \mathbf{a}'_u^i given context \mathbf{x}_u^i , whereas $y_u^i = 0$ implies the opposite. Preferences y_u^i are assumed to follow the Bradley–Terry–Luce (BTL) model [7, 18, 89]:

$$\begin{aligned} \mathbb{P}[y_u^i = 1 \mid u, \mathbf{x}_u^i, \mathbf{a}_u^i, \mathbf{a}'_u^i] &= \frac{1}{1 + \exp(-(r_u(\mathbf{x}_u^i, \mathbf{a}_u^i) - r_u(\mathbf{x}_u^i, \mathbf{a}'_u^i)))} \\ &= \sigma\left(\theta_u^\top (\phi(\mathbf{x}_u^i, \mathbf{a}_u^i) - \phi(\mathbf{x}_u^i, \mathbf{a}'_u^i))\right), \end{aligned}$$

where $r_u(\mathbf{x}, \mathbf{a}) = \theta_u^\top \phi(\mathbf{x}, \mathbf{a})$ is a linear reward function parameterized by an unknown vector θ_u and a known feature mapping $\phi: \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$ with $\|\phi(\mathbf{x}, \mathbf{a})\|_2 \leq 1$ for all $(\mathbf{x}, \mathbf{a}) \in \mathcal{X} \times \mathcal{A}$, and $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the sigmoid function. The interpretations of the context, action, and feature

*In practice, users within a cluster may have similar but not identical preferences (e.g., individuals from similar backgrounds often exhibit minor differences). Our results remain valid under such variations, as discussed in Remark 5 and verified in Section 6. For clarity and consistency with prior works [22, 38, 39, 44], we still assume identical preferences in each cluster.

map ϕ in practical applications are discussed in detail in Section 3.2. Additionally, we define the feature difference $z_u^i = \phi(\mathbf{x}_u^i, \mathbf{a}_u^i) - \phi(\mathbf{x}_u^i, \mathbf{a}'^i_u)$, noting that $(\boldsymbol{\theta}^\top \mathbf{z})$ is 2-subgaussian for any $\boldsymbol{\theta} \in \Theta$.

A policy $\pi : \mathcal{X} \rightarrow \mathcal{A}$ is a mapping from contexts to actions. Given an arbitrary test user $u_t \in \mathcal{U}$, we define the *suboptimality gap* of a policy π_{u_t} as:

$$\text{SubOpt}_{u_t}(\pi_{u_t}) \triangleq J_{u_t}(\pi_{u_t}^*) - J_{u_t}(\pi_{u_t}) = \mathbb{E}_{\mathbf{x} \sim \rho_p} [\boldsymbol{\theta}_{u_t}^\top \phi(\mathbf{x}, \pi_{u_t}^*(\mathbf{x})) - \boldsymbol{\theta}_{u_t}^\top \phi(\mathbf{x}, \pi_{u_t}(\mathbf{x}))], \quad (1)$$

where $J_u(\pi) = \mathbb{E}_{\mathbf{x} \sim \rho_p} [r_u(\mathbf{x}, \pi(\mathbf{x}))]$ denotes the expected reward for user u under policy π , $\pi_u^* = \arg \max_{\pi} J_u(\pi)$ is the optimal policy, and ρ_p denotes the distribution over contexts.

We consider two settings based on dataset availability:

- **Pure Offline Model:** In this setting, the policy π_{u_t} for the test user u_t is derived from fixed, pre-collected offline datasets $\mathcal{D} = \bigcup_{u \in \mathcal{U}} \mathcal{D}_u$. The objective is to minimize the suboptimality gap in Equation (1) using solely offline data.
- **Active-data Augmented Model:** In addition to the fixed offline dataset \mathcal{D} , the learner actively selects N additional data points specifically for the test user u_t . At each active selection round $n \in [N]$, the learner chooses a data tuple $(\hat{\mathbf{x}}_{u_t}^n, \hat{\mathbf{a}}_{u_t}^n, \hat{\mathbf{a}}'^n_{u_t}) \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}$, obtains preference feedback $\hat{y}_{u_t}^n$, and forms an active dataset $\hat{\mathcal{D}} = \{(\hat{\mathbf{x}}_{u_t}^n, \hat{\mathbf{a}}_{u_t}^n, \hat{\mathbf{a}}'^n_{u_t}, \hat{y}_{u_t}^n)\}_{n=1}^N$ after N rounds. The objective is to minimize Equation (1) by leveraging both offline and actively collected datasets $\mathcal{D} \cup \hat{\mathcal{D}}$.

Remark 1 (Distinctions from Classical Clustering of Bandits Works). In addition to the setting differences discussed in Section 2, we highlight the differences in assumptions between this paper and classical clustering of bandits works [22, 38, 39, 42, 72]. Previous studies typically rely on three assumptions: (i) *user randomness*, ensuring balanced data across users; (ii) *sufficient data* with a large heterogeneity gap for correct clustering; and (iii) *item regularity*, guaranteeing adequate coverage across all preference dimensions. While the only prior offline work [44] relaxes user randomness and data sufficiency, it still depends on item regularity. However, this assumption is overly restrictive in our setting and real-life scenarios, as pairwise feedback may be interdependent and distort coverage. In contrast, we remove all three assumptions to develop a more general and practical framework, treating the setting with item regularity assumption only as a special case.

3.2 Representative Applications

Our framework is closely related to the reinforcement learning from human feedback (RLHF) paradigm [17, 37, 89]. In this setting, \mathbf{x}_u^i represents a prompt shown to labeler u , $(\mathbf{a}_u^i, \mathbf{a}'^i_u)$ are two candidate responses, and y_u^i indicates the labeler's preference over two responses. The reward $r_u(\mathbf{x}, \mathbf{a})$ reflects the labeler's underlying evaluation, while $\phi(\mathbf{x}_u^i, \mathbf{a}_u^i)$ can be interpreted as the output of all but the final layer of a pre-trained language model and $\boldsymbol{\theta}_u$ as the personalized weights in its final layer [37, 52, 89]. In this view, the pure offline setting aims to aggregate offline pairwise preference data from multiple labelers to align the base model for the test labeler, whereas the active-data augmented setting focuses on the test labeler by carefully selecting prompt-response pairs based on the offline data. For instance, the learner may target prompts where the model's responses are more uncertain or diverse, and pair them with contrasting candidate responses, so that the resulting preference feedback provides additional information for refining the user's preference estimate.

Beyond RLHF, our framework also applies to recommendation systems [2, 36, 80], where u denotes a user, \mathbf{x}_u^i captures contextual information (e.g., time, recommendation category, or interface variant), $(\mathbf{a}_u^i, \mathbf{a}'^i_u)$ are two candidate items (such as movies or products), and y_u^i indicates which item was preferred. The pure offline case models cold-start recommendation, estimating the test user's preferences from historical interactions of similar users. The active-data augmented setting

Algorithm 1 Offline Connection-based Clustering of Preference Learning

-
- 1: **Input:** Test user $u_t \in \mathcal{U}$; offline dataset $\mathcal{D} = \bigcup_{u \in \mathcal{U}} \mathcal{D}_u$; parameters $\alpha \geq 1, \lambda > 0, \delta > 0, \kappa > 0, \hat{\gamma} \geq 0$; and reference vector \mathbf{w} .
 - 2: **Initialization:** Construct a null graph $\mathcal{G} = (\mathcal{V}, \emptyset)$ where $\mathcal{V} = \mathcal{U}$. For each user $u \in \mathcal{V}$, compute $\hat{\theta}_u$ and CI_u as in Equation (2).
 - 3: // Offline Cluster Learning
 - 4: **for** each pair of users $u_1, u_2 \in \mathcal{V}$ **do**
 - 5: Connect (u_1, u_2) if the condition in Equation (3) holds.
 - 6: **end for**
 - 7: Let $\mathcal{G}_{\hat{\gamma}} = (\mathcal{V}, \mathcal{E}_{\hat{\gamma}})$ denote the updated graph.
 - 8: // Data Aggregation
 - 9: **for** each user $u \in \mathcal{V}$ **do**
 - 10: Aggregate data and update statistics:

$$\mathcal{V}_{\hat{\gamma}}(u) = \{v \mid (u, v) \in \mathcal{E}_{\hat{\gamma}}\} \cup \{u\}, \quad \tilde{M}_u = \frac{\lambda}{\kappa} I + \sum_{v \in \mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \mathbf{z}_v^i (\mathbf{z}_v^i)^\top, \quad \tilde{N}_u = \sum_{v \in \mathcal{V}_{\hat{\gamma}}(u)} N_v,$$

$$\tilde{\theta}_u = \arg \min_{\theta} \left[- \sum_{v \in \mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} (y_v^i \log \sigma(\theta^\top \mathbf{z}_v^i) + (1 - y_v^i) \log \sigma(-\theta^\top \mathbf{z}_v^i)) + \frac{\lambda}{2} \|\theta\|_2^2 \right].$$
 - 11: **end for**
 - 12: // Policy Output
 - 13: Calculate the pessimistic value estimate $\tilde{J}_{u_t}(\pi)$ for any policy π as in Equation (4).
 - 14: **Output:** $\pi_{u_t} = \arg \max_{\pi} \tilde{J}_{u_t}(\pi)$.
-

extends this by interactively querying the user with designed contextual features and item pairs, collecting feedback to improve preference estimation.

4 ALGORITHM FOR PURE OFFLINE MODEL

To address the first research question in Section 1 on how to learn cluster structures under fixed and imbalanced offline data without coverage assumptions, we begin with the pure offline model. In this section, we introduce our algorithm, *Offline Connection-based Clustering of Preference Learning* (Off-C²PL) in Section 4.1, followed by the theoretical analysis in Section 4.2. We further examine a special case under the commonly adopted *item regularity* assumption (Assumption 1) from the clustering of bandits literature [22, 39, 42, 44, 73], connecting our framework to prior studies.

4.1 Algorithm Design: Off-C²PL

We detail the procedure of Off-C²PL in Algorithm 1. To address scenarios without any coverage assumption, Off-C²PL constructs confidence intervals for each user's estimated preference vector based on the minimum eigenvalue of the user's information (Gramian) matrix, enabling reliable confidence estimation even with uneven data coverage across dimensions. The algorithm initializes a null graph and connects edges only between users whose estimated preferences are confidently identified as similar, ensuring safe data aggregation. To handle binary pairwise feedback $(\mathbf{a}_u^i, \mathbf{a}'_u^i, y_u^i)$ under a logistic model, Off-C²PL adopts a maximum likelihood estimation (MLE) approach, estimating $\hat{\theta}_u$ by minimizing the regularized negative log-likelihood of observed comparisons.

Input and Initialization. The inputs (line 1) include test user u_t , offline dataset $\mathcal{D} = \bigcup_{u \in \mathcal{U}} \mathcal{D}_u$, parameters $(\alpha, \lambda, \delta, \kappa, \hat{\gamma})$ explained later, and a reference vector $\mathbf{w} \in \mathbb{R}^d$ used for theoretical simplification which does not affect the induced policy [37, 89]. The algorithm initializes a null graph \mathcal{G} , representing each user in \mathcal{U} as an isolated node (line 2), and then computes key statistics:

$$\begin{aligned} \hat{\theta}_u &= \underset{\theta}{\operatorname{argmin}} \left[- \sum_{i=1}^{N_u} (y_u^i \log \sigma(\theta^\top z_u^i) + (1 - y_u^i) \log \sigma(-\theta^\top z_u^i)) + \frac{\lambda}{2} \|\theta\|_2^2 \right], \\ M_u &= \frac{\lambda}{\kappa} I + \sum_{i=1}^{N_u} z_u^i (z_u^i)^\top, \text{ CI}_u = \frac{\sqrt{\lambda\kappa} + 2\sqrt{d \log(1 + \frac{4\kappa N_u}{\lambda d}) + 2 \log(\frac{2U}{\delta})}}{\kappa \sqrt{\lambda_{\min}(M_u)}}. \end{aligned} \quad (2)$$

Here, $\hat{\theta}_u$ estimates user preferences under pairwise feedback, M_u is a Gramian matrix regularized by λ/κ , and CI_u denotes the confidence interval constructed based on the minimum eigenvalue of M_u , rather than the number of available samples, making it more suitable for scenarios without coverage assumptions in our setting.

Offline Cluster Learning. Unlike traditional online clustering of bandits algorithms [22, 38, 39, 72] which typically begin with a complete user graph and iteratively delete edges based on online feedback, our algorithm starts with a null graph \mathcal{G} and incrementally connects users whose preferences are sufficiently similar. This connection-based strategy is better suited to offline settings, where limited data per user make edge deletion unreliable and prone to bias. To determine similarity, we use the key threshold parameter $\hat{\gamma}$, which controls whether two users should be clustered together. Specifically, as shown in line 5, the algorithm connects two users u_1 and u_2 if they satisfy:

$$\left\| \hat{\theta}_{u_1} - \hat{\theta}_{u_2} \right\|_2 < \hat{\gamma} - \alpha (\text{CI}_{u_1} + \text{CI}_{u_2}), \quad (3)$$

where the parameter α controls the conservativeness of clustering: a larger α inflates confidence intervals, making the algorithm less likely to mistakenly cluster users with noisy estimates. This condition guarantees that the estimated difference between the preference vectors of u_1 and u_2 remains within the acceptable range $\hat{\gamma}$ with high confidence (see Section 4.2 for details). In this way, the algorithm only connects users whose behaviors are similar enough under the offline data, progressively building a graph that accurately reflects the underlying cluster structure.

Data Aggregation. Let $\mathcal{G}_{\hat{\gamma}}$ denote the graph obtained after the cluster learning phase. Based on this graph, the algorithm aggregates data from users who are identified to have similar preferences (line 10). Specifically, we define $\mathcal{V}_{\hat{\gamma}}(u)$ as the set containing user u and its one-shot neighbors, representing all users estimated to share similar preferences with u . Using this set, the algorithm constructs the aggregated Gramian matrix \tilde{M}_u by combining samples from all users in $\mathcal{V}_{\hat{\gamma}}(u)$ and calculates the total number of samples \tilde{N}_u within this set. The preference estimate for user u is then refined by applying MLE to the aggregated data, yielding $\tilde{\theta}_u$.

Policy Output. In the final step, the algorithm computes a pessimistic estimate [28, 35, 55] of the value function for any policy π for the test user u_t which downweights underrepresented dimensions and emphasizes directions with sufficient data coverage, thereby mitigating the risk of overestimating performance in poorly explored dimensions:

$$\tilde{J}_{u_t}(\pi) = \left(\mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi(\mathbf{x}))] - \mathbf{w} \right)^\top \tilde{\theta}_{u_t} - \tilde{\beta}_{u_t} \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_{u_t}^{-1}}, \quad (4)$$

where the confidence term $\tilde{\beta}_u = \left(2\sqrt{d \log(1 + \frac{4\tilde{N}_u \kappa}{\lambda d}) + 2 \log(\frac{2U}{\delta})} + \sqrt{\lambda\kappa} \right) / \kappa$ accounts for estimation uncertainty (line 13). The algorithm outputs the final policy π_{u_t} that maximizes a pessimistic estimate $\tilde{J}_{u_t}(\pi)$, following the principle of pessimism in offline learning [28, 35]. This estimate is

designed to down-weight underrepresented dimensions and prioritize actions in regions of the feature space where the data provides more reliable information. Note that obtaining the exact π_{u_t} in Algorithm 1 requires an exhaustive search, which is feasible for small context and action spaces \mathcal{X} and \mathcal{A} . For large-scale settings, one can instead employ policy optimization methods such as PPO [17, 60] to efficiently approximate π_{u_t} .

4.2 Theoretical Results for Algorithm 1

We present the theoretical results for Algorithm 1 (Off-C²PL), with detailed proofs in Appendix B and key notations summarized in Table 2. Lemma 1 bounds the estimation error of each user's preference vector $\hat{\theta}_u$ based on individual data (line 2); Lemma 2 characterizes the homogeneous and heterogeneous neighbor sets ($\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$), quantifying data aggregation quality in the learned graph $\mathcal{G}_{\hat{\gamma}}$; and Lemma 3 extends this analysis to the aggregated estimator $\tilde{\theta}_u$ (line 10). Finally, Theorem 1 provides the main suboptimality bound. We begin by introducing the minimum heterogeneity gap between different clusters in Definition 1.

Table 2. Summary of neighbor set notations.

Notation	Definition	Interpretation
$\mathcal{V}_{\hat{\gamma}}(u)$	$\{u\} \cup \{v \mid (u, v) \in \mathcal{E}_{\hat{\gamma}}\}$	Set containing user u and all its neighbors in the graph $\mathcal{G}_{\hat{\gamma}}$.
$\mathcal{R}_{\hat{\gamma}}(u)$	$\{v \mid v \in \mathcal{V}_{\hat{\gamma}}(u), \theta_u = \theta_v\}$	Set of <i>homogeneous neighbors</i> of u , i.e., users in $\mathcal{V}_{\hat{\gamma}}(u)$ sharing the same preference vector. Their data can be safely aggregated with u 's without introducing bias.
$\mathcal{W}_{\hat{\gamma}}(u)$	$\{v \mid v \in \mathcal{V}_{\hat{\gamma}}(u), \theta_u \neq \theta_v\}$	Set of <i>heterogeneous neighbors</i> of u , i.e., users in $\mathcal{V}_{\hat{\gamma}}(u)$ with different preference vectors. Aggregating their data with u 's may introduce bias and should be carefully controlled.

Definition 1 (Minimum Heterogeneity Gap). *The preference vectors of users from different clusters are separated by at least a gap of γ . Specifically, for any two users u and v belonging to different clusters (i.e., $j_u \neq j_v$), it holds that $\|\theta_u - \theta_v\|_2 \geq \gamma$.*

Lemma 1 (Confidence Ellipsoid of $\hat{\theta}_u$). *For any user u , under the initialization in Equation (2) with $\kappa = 1/(2 + e^2 + e^{-2})$, it holds with probability at least $1 - \delta$ that*

$$\left\| \hat{\theta}_u - \theta_u \right\|_2 \leq \frac{\sqrt{\lambda \kappa} + 2\sqrt{2 \log\left(\frac{1}{\delta}\right) + d \log\left(1 + \frac{4N_u \kappa}{d\lambda}\right)}}{\kappa \sqrt{\lambda_{\min}(M_u)}}.$$

Lemma 1 provides a high-probability bound on the estimation error of $\hat{\theta}_u$ defined in Equation (2), which guarantees that the estimation error for each $\hat{\theta}_u$ is controlled by the minimum eigenvalue of the information matrix M_u for user u . Note that the estimate $\tilde{\theta}_u$ is obtained by aggregating all data from users in the neighborhood $\mathcal{V}_{\hat{\gamma}}(u)$, which includes both homogeneous and heterogeneous neighbors. Since Algorithm 1 relies on $\tilde{\theta}_u$ to determine the final policy, it is crucial to analyze the cardinality of both sets $\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$, since the former provides additional homogeneous samples that help reduce the estimation error, while the latter may introduce biased samples that can sometimes adversely affect the estimate. We formalize this in the following lemma:

Lemma 2 (Cardinality of $\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$). *Let parameter inputs in Algorithm 1 satisfy $\alpha \geq 1$, λ and δ be such that $\lambda \leq 2 \log\left(\frac{2U}{\delta}\right) + d \log\left(1 + \frac{4\kappa \min_v \{N_v\}}{d\lambda}\right)$, $\delta \leq \frac{d\lambda}{4\kappa \min_v \{N_v\} + d\lambda}$, and $\kappa = 1/(2 + e^2 + e^{-2})$. Define $\varepsilon = \hat{\gamma} - \gamma$ as the gap between the selected clustering threshold and the true minimum heterogeneity gap. Then there exist some $\alpha_r \in \left(\frac{\kappa}{3(\alpha+1)\sqrt{2 \max\{2, d\} \log(2U/\delta)}}, \frac{\kappa}{2(\alpha-1)\sqrt{2 \log(2U/\delta)}}\right)$ and*

$\alpha_w \in \left(0, \frac{\kappa}{2(\alpha-1)\sqrt{2\log(2U/\delta)}}\right)$ such that for any user u , with probability at least $1 - \delta$, the cardinalities of the homogeneous and heterogeneous neighbor sets can be characterized as:

$$\mathcal{R}_{\hat{\gamma}}(u) = \left\{v \mid \theta_u = \theta_v \text{ and } \frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha_r \hat{\gamma}\right\} \cup \{u\}, \quad (5)$$

$$\mathcal{W}_{\hat{\gamma}}(u) = \left\{v \mid \gamma \leq \|\theta_u - \theta_v\|_2 < \hat{\gamma} \text{ and } \frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha_w \varepsilon\right\}. \quad (6)$$

In Lemma 2, the notation $\lambda_{\min}(M_u)$, denoting the minimum eigenvalue of the information matrix M_u , quantifies the sufficiency of data in the dataset \mathcal{D}_u . Since the preference vector θ_u lies in \mathbb{R}^d , the dataset must provide adequate coverage along each dimension to ensure a sufficiently large $\lambda_{\min}(M_u)$, i.e., an informative Gramian matrix.

By definition, $\mathcal{R}_{\hat{\gamma}}(u)$ consists of user u and its homogeneous neighbors, indicating those samples that are beneficial for accurately estimating the true preference vector $\tilde{\theta}_u$. The first condition in Equation (5) ensures the homogeneity of users within $\mathcal{R}_{\hat{\gamma}}(u)$, while the second condition shows that only when both u and v have sufficiently informative data can v be identified as a reliable neighbor. Moreover, the right-hand side of Equation (5) depends linearly on $\hat{\gamma}$, implying that increasing the clustering threshold $\hat{\gamma}$ allows more homogeneous neighbors to be included. On the other hand, $\mathcal{W}_{\hat{\gamma}}(u)$ captures the heterogeneous neighbors of u , which may introduce bias. The first condition in Equation (6) shows that only users with a preference difference smaller than $\hat{\gamma}$ may be mistakenly clustered together, while the second condition imposes a stricter data sufficiency requirement for these heterogeneous neighbors. Notably, since $\varepsilon = \hat{\gamma} - \gamma$, the required information level for heterogeneous connections is more stringent than that for homogeneous ones.

With Lemma 2, we can now bound the estimation error of $\tilde{\theta}_u$ in terms of the total number of aggregated samples, denoted by $N_{\mathcal{V}_{\hat{\gamma}}(u)}$, and the number of samples coming from heterogeneous neighbors, denoted by $N_{\mathcal{W}_{\hat{\gamma}}(u)}$. This is formalized in the following lemma.

Lemma 3 (Confidence Ellipsoid of $\tilde{\theta}_u$). *For any user u , under the data aggregation step of Algorithm 1 and the same conditions as in Lemma 2, it holds with probability at least $1 - \delta$ that*

$$\left\|\tilde{\theta}_u - \theta_u\right\|_{\tilde{M}_u} \leq \frac{\sqrt{\lambda\kappa} + 2\sqrt{2\log\left(\frac{2U}{\delta}\right) + d\log\left(1 + \frac{4\kappa N_{\mathcal{V}_{\hat{\gamma}}(u)}}{d\lambda}\right)}}{\kappa} + \frac{\hat{\gamma}\sqrt{dN_{\mathcal{W}_{\hat{\gamma}}(u)}}}{2}.$$

Lemma 3 shows that the estimation error of $\tilde{\theta}_u$ with respect to the information matrix built from its local neighborhood in $\mathcal{G}_{\hat{\gamma}}$ can be decomposed into two sources: the noise term (the first term), which captures the randomness due to finite samples, and the bias term (the second term), which reflects the heterogeneity arising from including neighbors in $\mathcal{W}_{\hat{\gamma}}(u)$. Building on this result, our first theorem characterizes the suboptimality gap of Algorithm 1 in the offline setting.

Theorem 1. *Under the same conditions as in Lemma 2, the suboptimality gap of Algorithm 1 for any test user u_t can be bounded with probability at least $1 - \delta$ as:*

$$\text{SubOpt}_{u_t}(\pi_{u_t}) \leq \tilde{O}\left(\sqrt{d}\left(1 + \hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}\right)\left\|\mathbb{E}_{\mathbf{x} \sim \rho_p}[\phi(\mathbf{x}, \pi_{u_t}(\mathbf{x}))] - \mathbf{w}\right\|_{\tilde{M}_{u_t}^{-1}}\right) \quad (7)$$

$$\leq \tilde{O}\left(\frac{\sqrt{d}\left(1 + \hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}\right)}{\sqrt{\lambda_{\min}(\tilde{M}_{u_t})}}\right), \quad (8)$$

where \tilde{O} hides absolute constants and logarithmic factors. The matrix $\tilde{M}_{u_t} = \frac{\lambda}{\kappa} I + \sum_{v \in \mathcal{V}_{\hat{\gamma}}(u_t)} \sum_{i=1}^{N_v} \mathbf{z}_v^i (\mathbf{z}_v^i)^\top$ denotes the information matrix constructed from the aggregated data of all users in $\mathcal{V}_{\hat{\gamma}}(u_t)$. Furthermore, when the threshold satisfies $\hat{\gamma} \leq \gamma$, the heterogeneous set becomes empty according to Lemma 2, i.e., $\mathcal{W}_{\hat{\gamma}}(u_t) = \emptyset$, and the suboptimality bound simplifies to

$$\text{SubOpt}_{u_t}(\pi_{u_t}) \leq \tilde{O} \left(\sqrt{d} \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_{u_t}(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_{u_t}^{-1}} \right) \leq \tilde{O} \left(\sqrt{\frac{d}{\lambda_{\min}(\tilde{M}_{u_t})}} \right). \quad (9)$$

The suboptimality gap in Equation (7) of Theorem 1 consists of the product of two interpretable terms. The first term, $\sqrt{d} \left(1 + \hat{\gamma} \sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}} \right)$, can be further decomposed into two parts. Up to logarithmic factors, the first part can be seen as a fundamental term that arises from the inherent sample noise and reflects the baseline statistical uncertainty. The second part, which grows linearly with $\hat{\gamma}$ and sublinearly with the number of samples from heterogeneous neighbors $N_{\mathcal{W}_{\hat{\gamma}}(u_t)}$, captures the bias introduced by potential heterogeneity from the neighbors. As represented in Lemma 2, choosing a larger $\hat{\gamma}$ can increase $N_{\mathcal{W}_{\hat{\gamma}}(u_t)}$, thereby amplifying this bias term.

The second term in Equation (8), $\left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_{u_t}(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_{u_t}^{-1}}$, is known as the *concentratability coefficient*, a concept widely used in offline learning and policy evaluation [28, 37, 89]. This term quantifies the mismatch between the context-action distribution induced by the optimal policy and the distribution supported by the offline data from u_t and its neighbors in the resulting graph $\mathcal{G}_{\hat{\gamma}}$. A smaller concentratability coefficient implies that the offline data provides better coverage of the distribution under the optimal policy. Furthermore, choosing the reference vector \mathbf{w} as a representative feature (e.g., the most frequent feature vector ϕ observed in the data) [37, 89] aligns the concentratability term with the data-supported subspace, leading to a tighter suboptimality.

The dependence on \tilde{M}_{u_t} in Equation (8) through its minimum eigenvalue indicates that the overall sample efficiency is constrained by how well the data covers each dimension of the parameter space. Specifically, $\lambda_{\min}(\tilde{M}_{u_t})$ appearing in the denominator implies that, in the worst case, the effective number of samples per dimension is determined by the least informative direction. As indicated in Lemma 2, increasing $\hat{\gamma}$ can expand the neighborhood $\mathcal{V}_{\hat{\gamma}}(u_t)$, enlarging \tilde{M}_{u_t} and potentially improving coverage, though at the cost of introducing more heterogeneity bias.

Remark 2 (Selection of $\hat{\gamma}$). As shown in Lemma 2, the cardinalities of both $\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$ depend critically on the choice of $\hat{\gamma}$. Increasing $\hat{\gamma}$ generally enlarges both sets: a larger $\mathcal{R}_{\hat{\gamma}}(u)$ provides more homogeneous samples that can improve the accuracy of estimating θ_u , whereas a larger $\mathcal{W}_{\hat{\gamma}}(u)$ may introduce greater bias due to the inclusion of heterogeneous neighbors (as analyzed in Lemma 3 and Theorem 1). Therefore, careful selection of $\hat{\gamma}$ is crucial. Notably, Equation (9) shows that choosing $\hat{\gamma} \leq \gamma$ simplifies the suboptimality bound to a bias-free form. This provides a practical strategy to avoid large bias when a lower bound of γ is available, but at the cost of reducing $\mathcal{R}_{\hat{\gamma}}(u_t)$ and thus increasing the noise due to fewer aggregated samples. Due to space limitations, we defer detailed guidelines on selecting this parameter in practice to Appendix A.

Remark 3 (Comparison with Single User Case). When we choose $\hat{\gamma} = 0$, Algorithm 1 reduces to the special case where no clustering is learned and only the data from the test user, \mathcal{D}_{u_t} , is used for estimation. In this scenario, the bound in Theorem 1 specializes to $\tilde{O} \left(\sqrt{d} \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_{u_t}(\mathbf{x}))] - \mathbf{w} \right\|_{M_{u_t}^{-1}} \right)$, which matches the suboptimality bound derived for the single-user case in previous works [37, 89].

Remark 4 (Discussions on Parameter κ). The input parameter κ in Algorithm 1 serves as a non-linearity coefficient, lower bounding the minimum slope of the sigmoid function, i.e.,

$$\min_{(\mathbf{x}, \mathbf{a}, \mathbf{a}') \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}, \theta \in \Theta} \nabla \sigma(\phi(\mathbf{x}, \mathbf{a})^\top \theta - \phi(\mathbf{x}, \mathbf{a}')^\top \theta) \geq \kappa > 0. \quad (10)$$

In our setting, κ can be safely fixed to the constant $1/(2 + e^2 + e^{-2})$, which guarantees the validity of our theoretical results (e.g., Theorem 1). This is because we assume $\|\theta_u\|_2 \leq 1$ and $\|\phi(\mathbf{x}, \mathbf{a})\|_2 \leq 1$, following prior works on contextual logistic bandits [12, 33, 50] and clustering of bandits literature [22, 44, 72, 73]. In more general scenarios where the ℓ_2 -norm of either θ_u or $\phi(\mathbf{x}, \mathbf{a})$ is not bounded by a constant, the margin can become arbitrarily large, and $1/\kappa$ may grow exponentially. In such cases, as shown in Lemma 3 and Appendix B.4 (proof of Theorem 1), our suboptimality bound scales linearly with $1/\kappa$. By contrast, prior work in the single-user setting exploits mirror-descent techniques to improve this dependence to $1/\sqrt{\kappa}$ [37], which is argued to be tight [17, 37]. Extending this improved $\sqrt{\kappa}$ dependence to our heterogeneous multi-user setting with clustering remains an interesting open problem.

4.3 Further Results and Comparisons under Item Regularity Assumption

In the traditional clustering of bandits literature [16, 22, 39, 44, 72, 73], a common assumption is that the offline datasets provide sufficient coverage across all dimensions of the preference vector. This condition ensures that the information matrix is well-conditioned, which is crucial for accurate estimation. We first introduce this standard requirement, known as the *item regularity assumption*, and then discuss how our algorithm and theoretical results change under this setting.

Assumption 1 (Item Regularity). Let ρ be a distribution over $\{(\mathbf{x}, \mathbf{a}, \mathbf{a}') \in \mathcal{X} \times \mathcal{A} \times \mathcal{A} : \|\phi(\mathbf{x}, \mathbf{a})\|_2 \leq 1, \|\phi(\mathbf{x}, \mathbf{a}')\|_2 \leq 1\}$ where covariance matrix $\mathbb{E}_{(\mathbf{x}, \mathbf{a}, \mathbf{a}') \sim \rho} [(\phi(\mathbf{x}, \mathbf{a}) - \phi(\mathbf{x}, \mathbf{a}'))(\phi(\mathbf{x}, \mathbf{a}) - \phi(\mathbf{x}, \mathbf{a}'))^\top]$ is full rank with minimum eigenvalue $\lambda_a > 0$. For any fixed unit vector $\theta \in \mathbb{R}^d$, the random variable $(\theta^\top (\phi(\mathbf{x}, \mathbf{a}) - \phi(\mathbf{x}, \mathbf{a}')))^2$, with $(\mathbf{x}, \mathbf{a}, \mathbf{a}') \sim \rho$, has sub-Gaussian tails with variance upper bounded by σ^2 . Each context-action pair $(\mathbf{x}_u^i, \mathbf{a}_u^i, \mathbf{a}'_u^i)$ in \mathcal{D}_u is selected from a finite candidate set \mathcal{S}_u^i with size $|\mathcal{S}_u^i| \leq S$ for any $i \in [N_u]$, where the actions in \mathcal{S}_u^i are independently drawn from ρ . Moreover, we assume the smoothed regularity parameter $\tilde{\lambda}_a = \int_0^{\lambda_a} \left(1 - e^{-\frac{(\lambda_a - x)^2}{2\sigma^2}}\right)^S dx$ is known to the algorithm.

Assumption 1 ensures that the data distribution is sufficiently rich to provide informative samples in all directions of the preference vector θ_u . This assumption is especially relevant when offline data are collected from finite action spaces with bounded size, such as datasets generated by logging policies in online bandits [20, 72]. Under this condition, preference estimates become accurate once enough data are observed, since the minimum eigenvalue of the information matrix grows directly with the number of samples. Consequently, our confidence bounds decrease with the amount of offline data rather than depending solely on the minimum eigenvalue itself. Lemma 4 summarizes the modified clustering conditions and resulting characterizations.

Lemma 4 (Extension of Lemma 2). Under Assumption 1, replace the confidence interval by $\text{CI}_u = \left(\sqrt{\lambda\kappa} + 2\sqrt{d \log\left(1 + \frac{4\kappa N_u}{\lambda d}\right)} + 2\log\left(\frac{2U}{\delta}\right)\right) / \left(\kappa\sqrt{\tilde{\lambda}_a} N_u / 2\right)$, and adjust the condition in Equation (3) to:

$$\|\hat{\theta}_{u_1} - \hat{\theta}_{u_2}\|_2 < \hat{\gamma} - \alpha(\text{CI}_{u_1} + \text{CI}_{u_2}) \quad \text{and} \quad \min\{N_{u_1}, N_{u_2}\} \geq N_{\min},$$

where $N_{\min} = \frac{16}{\lambda_a^2} \log\left(\frac{8Ud}{\lambda_a^2 \delta}\right)$. All other conditions remain as in Lemma 2. Then there exist some $\alpha'_r \in \left(\frac{\kappa\sqrt{\tilde{\lambda}_a}}{3(\alpha+1)\sqrt{\max\{2,d\}\log(2U/\delta)}}, \frac{\kappa\sqrt{\tilde{\lambda}_a}}{2(\alpha-1)\sqrt{2\log(2U/\delta)}}\right)$ and $\alpha'_w \in \left(0, \frac{\kappa\sqrt{\tilde{\lambda}_a}}{2(\alpha-1)\sqrt{\log(2U/\delta)}}\right)$ such that the

cardinalities of $\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$ are given by:

$$\mathcal{R}_{\hat{\gamma}}(u) = \begin{cases} \{v \mid \theta_u = \theta_v, \frac{1}{\sqrt{N_u}} + \frac{1}{\sqrt{N_v}} < \alpha'_r \hat{\gamma}, N_v \geq N_{\min}\} \cup \{u\}, & N_u \geq N_{\min} \\ \{u\}, & \text{otherwise} \end{cases}, \quad (11)$$

$$\mathcal{W}_{\hat{\gamma}}(u) = \begin{cases} \{v \mid \gamma \leq \|\theta_u - \theta_v\|_2 < \hat{\gamma}, \frac{1}{\sqrt{N_u}} + \frac{1}{\sqrt{N_v}} < \alpha'_w \epsilon\}, & N_u \geq N_{\min} \\ \emptyset, & \text{otherwise} \end{cases}. \quad (12)$$

The expressions above show that, under Assumption 1, the ability to correctly identify homogeneous and heterogeneous neighbors depends explicitly on the sample size rather than the conditioning of the Gramian matrix. This aligns with the results in standard offline clustering of bandits frameworks [44]. Below we present Corollary 1, which characterizes the suboptimality of our algorithm under Assumption 1.

Corollary 1. *Under the same conditions as in Lemma 4, the suboptimality of the algorithm is bounded with probability at least $1 - \delta$ as:*

$$\text{SubOpt}_{u_t}(\pi_{u_t}) \leq \tilde{O} \left(\sqrt{\frac{d}{\tilde{\lambda}_a}} \left(\sqrt{\frac{1}{N_{\mathcal{V}_{\hat{\gamma}}(u_t)}}} + \hat{\gamma} \sqrt{\eta_{\mathcal{W}_{\hat{\gamma}}(u_t)}} \right) \right),$$

where $\eta_{\mathcal{W}_{\hat{\gamma}}(u_t)} = \frac{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}{N_{\mathcal{V}_{\hat{\gamma}}(u_t)}}$ denotes the fraction of samples from heterogeneous neighbors among all samples aggregated for u_t in the graph $\mathcal{G}_{\hat{\gamma}}$.

Corollary 1 takes a form similar to the suboptimality bounds in classical offline clustering of bandits [44]. Specifically, the term $\sqrt{1/N_{\mathcal{V}_{\hat{\gamma}}(u_t)}}$ captures the *noise*, arising from the inherent variance in estimating the preference vector. This term decreases as the number of aggregated samples $N_{\mathcal{V}_{\hat{\gamma}}(u_t)}$ increases, implying that a larger $\hat{\gamma}$, which connects more users, reduces the noise. In contrast, the term $\hat{\gamma} \sqrt{\eta_{\mathcal{W}_{\hat{\gamma}}(u_t)}}$ captures the *bias*, introduced by aggregating data from neighbors whose preferences differ from u_t . This bias grows linearly with $\hat{\gamma}$ and depends on the fraction of heterogeneous samples included. Thus, while increasing $\hat{\gamma}$ reduces noise, it also risks introducing greater bias. This tradeoff underscores the importance of carefully tuning $\hat{\gamma}$ to balance sample efficiency with robustness against heterogeneity, as discussed in Remark 2. Finally, the scaling factor $\sqrt{d/\tilde{\lambda}_a}$ arises from Assumption 1, reflecting that each offline sample contributes only partial information across dimensions. As a result, the overall suboptimality must be scaled by $\sqrt{d/\tilde{\lambda}_a}$ to capture performance across all preference dimensions.

Remark 5 (Robustness of Algorithm 1). As noted in prior works on clustering of bandits [16, 73], it can be restrictive to assume that users within the same cluster share exactly identical preferences, as small gaps may exist even among users with similar backgrounds. To address this, those works developed additional algorithms to handle intra-cluster bias, often based on edge-deletion strategies [22, 39]. In contrast, we argue that our proposed Algorithm 1 is inherently robust to such cases. Specifically, when small preference gaps exist within a cluster, the setting can be interpreted as if each user forms its own cluster (i.e., $U = J$). In this case, $\mathcal{R}_{\hat{\gamma}}(u) = \{u\}$ in Lemma 2, while other users with similar (though not identical) preferences may be included in $\mathcal{W}_{\hat{\gamma}}(u)$ when $\hat{\gamma}$ is chosen larger than this gap, provided their data sufficiency satisfies the second condition in Equation (6) or Equation (12). According to Theorem 1 and Corollary 1, such users still contribute to the aggregated information matrix M_{u_t} and to the neighbor set $\mathcal{V}_{\hat{\gamma}}(u_t)$ which helps decrease noise with some additional bias, reflected in larger $N_{\mathcal{W}_{\hat{\gamma}}(u_t)}$ and thus $\eta_{\mathcal{W}_{\hat{\gamma}}(u_t)}$ (noting that

$\eta_{\mathcal{W}_{\hat{y}}(u_t)} \leq 1$ always holds). Therefore, in practice, when small intra-cluster gaps exist, it is often preferable to select a relatively small \hat{y} to better control the bias.

5 ALGORITHM FOR ACTIVE-DATA AUGMENTED MODEL

In Section 4, we analyzed the algorithm designed for clustering-based preference learning under the pure offline setting. However, as shown in Theorem 1, a key limitation of the pure offline case is its reliance on the distribution of the available datasets. More specifically, if the data collected from a user's neighbors fail to adequately cover the distribution induced by the optimal policy, the resulting concentratability coefficient may become large, which can significantly degrade performance. This phenomenon corresponds to the second research question introduced in Section 1: *how to mitigate the impact of insufficient coverage in offline datasets*. In many real-world applications, it is often feasible to collect a small amount of additional online or interactive data to complement existing offline datasets. Motivated by this, we extend our offline algorithm in Section 4 to the active-data augmented model defined in Section 3, which aims to address the distributional limitation challenge of the pure offline model by combining offline clustering with active-data augmentation.

Algorithm 2 Active-data Augmented - Offline Connection-based Clustering of Preference Learning

- 1: **Input:** Test user $u_t \in \mathcal{U}$, offline dataset $\mathcal{D} = \bigcup_{u \in \mathcal{U}} \mathcal{D}_u$, and online rounds N ; Graph $\mathcal{G}_{\hat{y}}$, neighbor set $\mathcal{V}_{\hat{y}}(u_t)$, aggregated Gramian matrix \tilde{M}_{u_t} , and initial preference estimate $\tilde{\theta}_{u_t}$ from Algorithm 1.
 - 2: **Initialization:** Set $\tilde{M}_{u_t}^0 \leftarrow \tilde{M}_{u_t}$ and $\tilde{\theta}_{u_t}^0 \leftarrow \tilde{\theta}_{u_t}$.
 - 3: // Active-data Augmentation
 - 4: **for** $n = 1, \dots, N$ **do**
 - 5: Select $(\hat{x}_{u_t}^n, \hat{a}_{u_t}^n, \hat{a}'_{u_t}^n)$ according to Equation (13).
 - 6: Observe feedback $\hat{y}_{u_t}^n$.
 - 7: Compute $\hat{z}_{u_t}^n = \phi(\hat{x}_{u_t}^n, \hat{a}_{u_t}^n) - \phi(\hat{x}_{u_t}^n, \hat{a}'_{u_t}^n)$.
 - 8: Update $\tilde{M}_{u_t}^n = \tilde{M}_{u_t}^{n-1} + \hat{z}_{u_t}^n (\hat{z}_{u_t}^n)^\top$ and $\tilde{\theta}_{u_t}^n$ as in Equation (14).
 - 9: **end for**
 - 10: // Policy Output
 - 11: Construct $\bar{\theta}_{u_t}$ as Equation (15).
 - 12: **Output:** $\pi_{u_t}(x) = \arg\max_{a \in \mathcal{A}} \phi(x, a)^\top \bar{\theta}_{u_t}$.
-

5.1 Algorithm Design: A²-Off-C²PL

We now introduce our algorithm for the active-data augmented model, which extends the cluster structure learned in Off-C²PL (Algorithm 1). Recall from Section 3 that in active-data augmented model, the learner can interact with the environment for a limited number of rounds to collect additional feedback. Specifically, it is allowed to select N rounds of active data for the target user u_t to mitigate the poor coverage of the offline datasets. We refer our algorithm in this setting as *Active-data Augmented - Offline Connection-based Clustering of Preference Learning* (A²-Off-C²PL). The core idea of A²-Off-C²PL is to actively select N rounds of data for the test user to complement the offline data by improving the coverage of the feature space (e.g. in conversational recommendation systems the website adopts N rounds of further dialogues to identify the users' preferences). Since the clustering structure has been learned offline, the active-data augmentation phase should be based on the aggregated Gramian matrix \tilde{M}_{u_t} , which summarizes the information from the test user's neighborhoods. As shown by the suboptimality bound in Theorem 1, the estimation error

is largely determined by the minimum eigenvalue of \tilde{M}_{u_t} . Therefore, the goal of this phase is to actively collect new data to increase this eigenvalue, ensuring that each dimension is sufficiently covered. The detailed procedure is summarized in Algorithm 2.

Input and Initialization. The inputs and initialization directly use the results from Algorithm 1. Specifically, in addition to test user u_t and offline dataset \mathcal{D} , the algorithm also takes the learned cluster graph $\mathcal{G}_{\hat{Y}}$, the neighbor set $\mathcal{V}_{\hat{Y}}(u_t)$, and the initial Gramian matrix \tilde{M}_{u_t} and preference estimate $\tilde{\theta}_{u_t}$ (Line 1). These are used to initialize the active-data augmentation phase (Line 2).

Active-data Augmentation. The key component of Algorithm 2 is the active-data augmentation procedure. In each round n , the algorithm selects the context-action pair on the most underrepresented dimensions to broaden the information matrix:

$$(\hat{\mathbf{x}}_{u_t}^n, \hat{\mathbf{a}}_{u_t}^n, \hat{\mathbf{a}}_{u_t}^n) = \underset{(\mathbf{x}, \mathbf{a}, \mathbf{a}') \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}}{\operatorname{argmax}} \left\{ \|\phi(\mathbf{x}, \mathbf{a}) - \phi(\mathbf{x}, \mathbf{a}')\|_{(\tilde{M}_{u_t}^{n-1})^{-1}} \right\}. \quad (13)$$

After selection, the feedback $\hat{y}_{u_t}^n$ is observed, and the difference feature $\hat{\mathbf{z}}_{u_t}^n$ is computed. The Gramian matrix is then updated as $\tilde{M}_{u_t}^n = \tilde{M}_{u_t}^{n-1} + \hat{\mathbf{z}}_{u_t}^n (\hat{\mathbf{z}}_{u_t}^n)^\top$, and the preference estimate is refined by solving the regularized maximum likelihood problem (regularized by the same λ as that in Algorithm 1) that combines both the offline aggregated data and all active-data up to round n :

$$\begin{aligned} \tilde{\theta}_{u_t}^n = \underset{\theta}{\operatorname{argmin}} & \left(- \sum_{v \in \mathcal{V}_{\hat{Y}}(u_t)} \sum_{i=1}^{N_v} \left[y_v^i \log \sigma(\theta^\top \mathbf{z}_v^i) + (1 - y_v^i) \log \sigma(-\theta^\top \mathbf{z}_v^i) \right] \right. \\ & \left. - \sum_{s=1}^n \left[\hat{y}_{u_t}^s \log \sigma(\theta^\top \hat{\mathbf{z}}_{u_t}^s) + (1 - \hat{y}_{u_t}^s) \log \sigma(-\theta^\top \hat{\mathbf{z}}_{u_t}^s) \right] + \frac{\lambda}{2} \|\theta\|_2^2 \right). \end{aligned} \quad (14)$$

Policy Output. Finally, the algorithm constructs the final preference estimate $\bar{\theta}_{u_t}$ by taking a weighted average of all historical estimates $\tilde{\theta}_{u_t}^n$ for $n = 1, \dots, N$:

$$\bar{\theta}_{u_t} = \frac{1}{d \lambda_{\min}(\tilde{M}_{u_t}^N) + N} \left(d \lambda_{\min}(\tilde{M}_{u_t}^N) \tilde{\theta}_{u_t}^N + \sum_{n=1}^N \tilde{\theta}_{u_t}^n \right). \quad (15)$$

This weighting places more emphasis on the final estimate, extending prior approach in Das et al. [17] which only uses a simple average for the pure active setting. The learned policy then selects the action that maximizes the expected reward as: $\pi_{u_t}(\mathbf{x}) = \operatorname{argmax}_{\mathbf{a} \in \mathcal{A}} \phi(\mathbf{x}, \mathbf{a})^\top \bar{\theta}_{u_t}$.

5.2 Theoretical Results for Algorithm 2

We now present the theoretical guarantee for Algorithm 2, A²-Off-C²PL, in Theorem 2.

Theorem 2. *Under the same assumptions as in Lemma 2 and Theorem 1, the suboptimality gap of Algorithm 2 for the test user u_t can be bounded with probability at least $1 - \delta$ as:*

$$\text{SubOpt}_{u_t}(\pi_{u_t}) \leq \tilde{O} \left(\frac{\sqrt{d} \left(1 + \hat{Y} \sqrt{N \mathcal{W}_{\hat{Y}}(u_t)} \right)}{\sqrt{\lambda_{\min}(\tilde{M}_{u_t}^N) + N/d}} \right),$$

where $\tilde{M}_{u_t}^N = \frac{1}{\kappa} I + \sum_{v \in \mathcal{V}_{\hat{Y}}(u_t)} \sum_{i=1}^{N_v} \mathbf{z}_v^i (\mathbf{z}_v^i)^\top + \sum_{i=1}^N \hat{\mathbf{z}}_{u_t}^i (\hat{\mathbf{z}}_{u_t}^i)^\top$ denotes the final Gramian matrix that combines both the offline aggregated data and the actively selected data in Algorithm 2.

In Theorem 2, the numerator mirrors the structure of Theorem 1: it is composed of two parts, where the first one representing the inherent sample noise, and the other capturing the bias

introduced by heterogeneous neighbors. The term inside the square root of the denominator, $\lambda_{\min}(\tilde{M}_{u_t}^N) + N/d$, quantifies the effective number of "useful" samples that contribute to accurately estimating the preference vector for each dimension, just as in Theorem 1. Specifically, $\lambda_{\min}(\tilde{M}_{u_t}^N)$ reflects the normal contribution of the aggregated information matrix, $\tilde{M}_{u_t}^N$, from both pure offline samples and active selected samples in each dimension, while N/d corresponds to the additional contribution of the N active samples, distributed across d dimensions.

Remark 6 (Comparison with Prior Results). When $N = 0$, the setting reduces to the pure offline scenario, and the suboptimality bound in Theorem 2 naturally recovers the bound from Theorem 1. Additionally, as discussed in Remark 3, setting $\hat{\gamma} = 0$ to only use samples from the test user itself allows us to specialize our result to the single-user case. Building on this, our framework can be further specialized to scenarios involving only active data without any offline data when $\mathcal{D} = \emptyset$, as explored in prior work [17]. In this case, $\tilde{M}_{u_t}^N$ consists solely of active samples, and the suboptimality bound in Theorem 2 outperforms the result in Das et al. [17] (which achieves $\tilde{O}(d/\sqrt{N})$) by incorporating $\lambda_{\min}(\tilde{M}_{u_t}^N)$ into the denominator, yielding a more refined bound.

As shown in Theorem 2, the final Gramian matrix under active-data augmentation, denoted by $\tilde{M}_{u_t}^N$, differs from \tilde{M}_{u_t} in that it not only aggregates the offline samples but also includes the actively selected samples. According to the selection rule in Equation (13), the algorithm deliberately targets the dimensions with the sparsest information, which is fundamentally different from passively using the given offline dataset. In scenarios where the offline data is imbalanced (i.e. with some dimensions severely underrepresented while others are sufficiently covered), this active selection allows the algorithm to focus additional samples on the least informative directions, effectively "filling in" the gaps and improving the estimation.

Therefore, a key quantity of interest is the improvement in the information matrix through our actively selected data, captured by the gap $\lambda_{\min}(\tilde{M}_{u_t}^N) - \lambda_{\min}(\tilde{M}_{u_t})$. We first give Definition 2 that characterizes such cases where active selection brings significant improvement.

Definition 2 ((d^*, N) -Sample Imbalanced Gramian Matrix). *A Gramian matrix M is called (d^*, N) -sample imbalanced if d^* is the smallest value in $\{1, \dots, d\}$ such that $\lambda_{d^*+1}(M) - \lambda_{\min}(M) \geq \lceil N/d^* \rceil$. By convention, any matrix is at least (d, N) -sample imbalanced, since there are only d dimensions and we treat $\lambda_{d+1}(M)$ as $+\infty$.*

Intuitively, this definition implies that there is a large discrepancy in sample sufficiency between the least well-informed dimension and the $(d^* + 1)$ -th dimension. For a (d^*, N) -sample imbalanced matrix, actively selecting samples according to Equation (13) can substantially boost the minimum eigenvalue by concentrating new samples where they are most needed. This is formalized in the following lemma.

Lemma 5 (Quantification of the Minimum Eigenvalue Improvement). *Assume that the feature difference vector $\mathbf{z} = \phi(\mathbf{x}, \mathbf{a}) - \phi(\mathbf{x}, \mathbf{a}')$ can span the entire Euclidean unit ball $\{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\|_2 \leq 1\}$ for all $(\mathbf{x}, \mathbf{a}, \mathbf{a}') \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}$. Further suppose that \tilde{M}_{u_t} is (d^*, N) -sample imbalanced as defined in Definition 2. Then, under the active selection rule in Equation (13) for a total of N rounds, it holds that*

$$\lambda_{\min}(\tilde{M}_{u_t}^N) - \lambda_{\min}(\tilde{M}_{u_t}) \geq \lfloor N/d^* \rfloor.$$

Combining Lemma 5 with Theorem 2, we can explicitly show how the active sampling improves the bound relative to the pure offline setting.

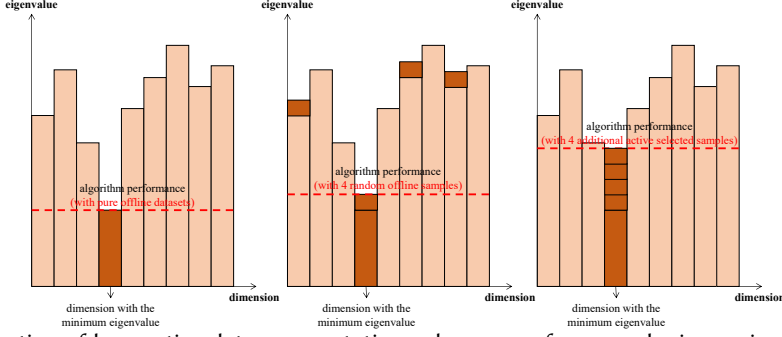


Fig. 1. Illustration of how active-data augmentation enhances performance by increasing the minimum eigenvalue of the information matrix. **Left:** Pure offline data suffers from underrepresented dimensions, limiting performance. **Middle:** Adding random offline samples offers limited improvement. **Right:** Actively selected samples focus on underrepresented dimensions, substantially increasing the minimum eigenvalue and improving performance.

Corollary 2. Suppose that the assumptions in Lemma 5 hold. Then the suboptimality gap in Theorem 2 can be rewritten as:

$$\text{SubOpt}_{u_t}(\pi_{u_t}) \leq \tilde{O} \left(\frac{\sqrt{d} \left(1 + \hat{\gamma} \sqrt{N \mathcal{W}_{\hat{\gamma}}(u_t)} \right)}{\sqrt{\lambda_{\min}(\tilde{M}_{u_t}) + N/d^*}} \right).$$

Moreover, the bound can be simplified to: $\tilde{O} \left(\frac{\sqrt{d} \left(1 + \hat{\gamma} \sqrt{N \mathcal{W}_{\hat{\gamma}}(u_t)} \right)}{\sqrt{\lambda_{\min}(\tilde{M}_{u_t}) + N}} \right)$ when \tilde{M}_{u_t} is $(1, N)$ -sample imbalanced.

As shown in Lemma 5 and Corollary 2, when the offline Gramian matrix \tilde{M}_{u_t} is highly imbalanced (i.e., well covered in some dimensions but sparse in others) our active-data selection rule explicitly targets the underrepresented dimensions. In this case, each actively selected sample can contribute more than a single effective observation. Specifically, comparing Theorem 1 with Corollary 2, the denominator improves by $\tilde{O}(N/d^*)$ for some $d^* \leq d$, rather than the $O(N/d)$ scaling in the general case. Intuitively, the active samples only need to be distributed across d^* dimensions instead of all d dimensions. Consequently, for a (d^*, N) -sample imbalanced matrix \tilde{M}_{u_t} , one actively selected sample is equivalent to d/d^* fully informative samples and yields a suboptimality gain. Figure 1 depicts this phenomenon. This result highlights how active-data augmentation can effectively mitigate imbalance in offline coverage by reinforcing the sparse directions of the preference.

Finally, we present a special-case result under the item regularity assumption (Assumption 1) and the condition that \tilde{M}_{u_t} is (d^*, N) -sample imbalanced, which illustrates the benefit of active-data augmentation even in a traditional bandit context:

Corollary 3. Suppose Assumption 1 holds and that \tilde{M}_{u_t} is (d^*, N) -sample imbalanced. Following the proof of Corollary 1, it holds that

$$\text{SubOpt}_{u_t}(\pi_{u_t}) \leq \tilde{O} \left(\sqrt{\frac{d}{\tilde{\lambda}_a}} \left(\frac{1}{\sqrt{N \mathcal{V}_{\hat{\gamma}}(u_t)} + N/(d^* \tilde{\lambda}_a)} + \frac{\hat{\gamma} \sqrt{N \mathcal{W}_{\hat{\gamma}}(u_t)}}{\sqrt{N \mathcal{V}_{\hat{\gamma}}(u_t)} + N/(d^* \tilde{\lambda}_a)} \right) \right).$$

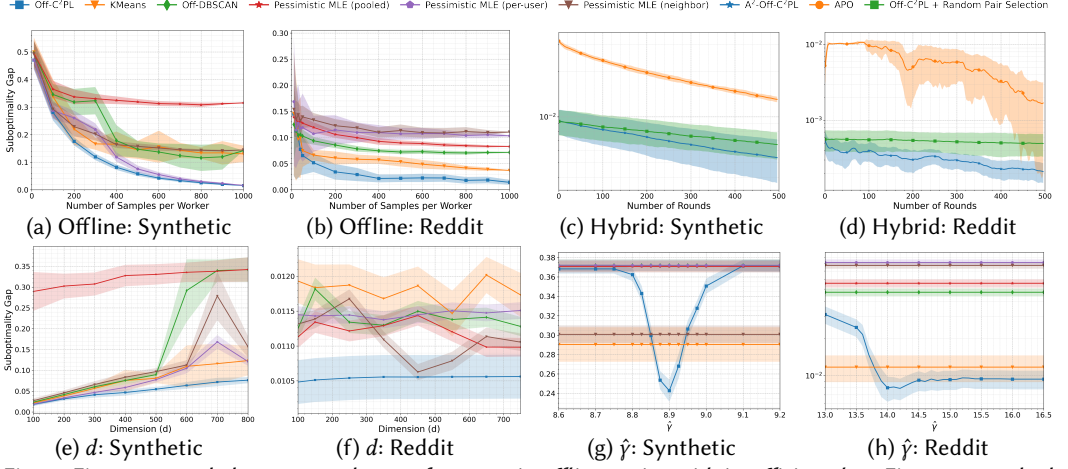


Fig. 2. Figures 2a and 2b correspond to *performance in offline setting with insufficient data*, Figures 2c and 2d correspond to *performance in hybrid setting*, Figures 2g and 2h correspond to *the impact of dimension d* , and Figures 2e and 2f correspond to *the impact of clustering-threshold $\hat{\gamma}$* .

Corollary 3 can be interpreted in terms of *noise* (the first term) and *bias* (the second term). Importantly, under Assumption 1, each actively selected sample is equivalent to at least $1/(d^* \tilde{\lambda}_a)$ offline samples (which is strictly greater than one, since $\tilde{\lambda}_a \leq 1/d \leq 1/d^*$ holds by Wang et al. [73]). This advantage arises because active samples offer better coverage through the active selection rule than the coverage offered by Assumption 1 for offline samples. Consequently, this result strengthens Corollary 1, yielding a strictly better suboptimality bound by reducing both noise and bias.

6 EXPERIMENTS

In this section, we evaluate the performance of Off-C²PL and A²-Off-C²PL using synthetic and real-world data. All experiments are averaged over 20 independent rounds.

Baselines. We compare Off-C²PL with both enhanced versions of traditional clustering algorithms and prior methods for contextual logistic bandits. Specifically, we adapt classical clustering algorithms such as KMeans [47] (with $\sqrt{\#}$ of users as cluster number) and DBSCAN [59] to our setting by incorporating the same policy output phase as in Algorithm 1 with their clustering procedures. We also include variants of Pessimistic MLE [89] for contextual logistic bandits: *Pessimistic MLE (per-user)* uses only the test user's data, *Pessimistic MLE (pooled)* aggregates data from all users, and *Pessimistic MLE (neighbor)* leverages data from the test user's neighbors identified by a KNN algorithm using cosine similarity on θ . For evaluating A²-Off-C²PL, we compare against the pure offline algorithm Off-C²PL trained on randomly generated offline samples and the pure active learning algorithm Active Preference Optimization (APO) from Das et al. [17] that operates without any offline data.

Synthetic Dataset. We construct a synthetic pairwise-preference dataset with $U = 40$ users partitioned into $J = 8$ clusters uniformly at random. Each cluster j has a ground-truth vector $\theta^j \in \mathbb{R}^d$ with $d = 768$, matching the dimensionality of the real-world embeddings used in our experiments. For a user u in cluster c , we set $\theta_u = \theta^c + \epsilon_u$, where $\epsilon_u \sim \mathcal{N}(0, s^2 I_d)$. This adds mild within-cluster heterogeneity so users are similar but not identical, better reflecting real data. We then generate 1000 pairwise comparisons per user under a Bradley-Terry-Luce model: for

a pair-difference feature $\mathbf{z} \sim \mathcal{N}(0, I_d)$, the preferred item is sampled with probability $\sigma(\beta \boldsymbol{\theta}_u \cdot \mathbf{z})$, where $\sigma(x) = (1 + e^{-x})^{-1}$ and β controls noise (larger β implies cleaner preferences).

Real-World Dataset. We use the Reddit TL;DR summarization [68] alongside human preferences collected by Stiennon et al. [64]. Each sample in our dataset consists of a forum post from Reddit, paired with two distinct summaries generated by the GPT-2 language model. Human annotators then indicate their preference for one of the summaries. This dataset contains preference annotations from 76 users, with individual contributions ranging from as few as 2 to more than 18,000 prompts. For evaluation, we focus on 42 annotators who each provide more than 1,000 annotations, and from each of these, we uniformly sample 1,000 preferences for testing. In order to calculate the suboptimality gap, it is necessary to have access to an optimal policy. However, the true optimal policy is unknown when working with real-world data. Therefore, we must rely on the available dataset to approximate the most optimal policy. Thus, we leverage maximum likelihood estimation (MLE) regression through a gradient descent on the full dataset, to ensure that the derived optimal policy is optimal relative to the given dataset.

Experiment 1: Performance under pure offline model. We examine Off-C²PL against a suite of baselines on both the synthetic and the Reddit dataset, varying the per-user sample budget from 100 to 1000 pairs, considering 40 users. On the synthetic data (Figure 2a), Off-C²PL has the smallest suboptimality gap across the entire range. Relative to the baselines in this run, it improves performance by 88.1% over KMeans, 89.1% over Off-DBSCAN, and 95.1%, 89.2%, and 3.39% over Pessimistic MLE (pooled), (neighbor), and (per-user). Pessimistic MLE (per-user) becomes competitive only after using more than 80% of the samples and remains clearly worse in the low-sample regime. On the Reddit dataset (Figure 2b), no baseline matches Off-C²PL. With only ≈ 400 pairs per user it achieves a near-zero suboptimality gap and delivers relative improvements of 61.5% over KMeans, 80.1% over Off-DBSCAN, 82.8% over Pessimistic MLE (pooled), 87.1% over the neighbor, and 86.2% over the per-user variant.

Experiment 2: Performance under active-data augmented model. We compare A²-Off-C²PL against APO and an algorithm which uses Off-C²PL as offline initialization but replaces our active-data augmentation strategy with *random* pair selection. We allocate 20% of the data to the offline phase and then run 500 rounds of active-data selection. On the Reddit dataset, A²-Off-C²PL yields relative improvements of 87.6% over the online-only baseline and 57.5% over the random-selection baseline. On the synthetic dataset, the corresponding improvements are 58.7% and 18.0%. As shown in Figures 2c and 2d, the pure active method begins with a large suboptimality gap due to the missing offline head start. Although the active phase reduces this gap over rounds, it remains substantially worse. The random-selection baseline starts at the same gap as A²-Off-C²PL but fails to discover sufficiently informative pairs and therefore makes little progress. In contrast, A²-Off-C²PL consistently drives the gap downward across active rounds, achieving the best performance throughout.

Experiment 3: The impact of dimension d . We vary dimension d from 100 to 800 on synthetic data and from 100 to 768 on Reddit. For Reddit, we obtain lower-dimensional features by applying PCA to the original 768-dimensional embeddings, so 768 is the maximum. On the synthetic dataset (Figure 2e), the gap increases with d at a fixed sample size, as expected from higher estimation complexity. Notably, Off-C²PL degrades the slowest as it uses data across users within clusters and regularizes effectively in high dimensions. On Reddit, however (Figure 2f), there is no noticeable trend in performance across d , which is consistent with PCA preserving the dominant variance directions. Truncating to lower d primarily removes low-variance components that contribute little to preference prediction.

Experiment 4: The impact of clustering-threshold $\hat{\gamma}$. Sweeping the clustering-threshold $\hat{\gamma}$ reveals a bias–variance trade-off: overly small values merge unrelated users, while overly large values prevent

pooling users in true clusters (Figures 2g and 2h). With a well-calibrated $\hat{\gamma}$, Off-C²PL recovers the correct cluster structure and substantially reduces the suboptimality gap, demonstrating that accurate control of cluster connectivity is crucial when data is scarce.

7 CONCLUSION

In this paper, we introduce and systematically study the Offline Clustering of Preference Learning problem, where user preferences naturally vary. We propose Algorithm 1 (Off-C²PL), which leverages maximum likelihood estimation to cluster users with similar preferences without relying on any coverage assumption, enabling accurate aggregation of heterogeneous offline data. Our theoretical analysis characterizes the tradeoff between variance reduction from data aggregation and bias introduced by heterogeneity. We further extend this framework with active-data augmentation in Algorithm 2 (A²-Off-C²PL), which selectively samples underrepresented dimensions, achieving notable theoretical and empirical gains over purely offline methods.

A promising direction for future work is to refine our suboptimality bounds in cases where the ℓ_2 norm of θ_u is not constant. While prior single-user analyses improve the dependence on the nonlinearity parameter from $1/\kappa$ to $1/\sqrt{\kappa}$, extending this improvement to heterogeneous multi-user clustering remains open. Developing techniques to achieve a $1/\sqrt{\kappa}$ dependency within our framework would mark a significant theoretical advancement.

REFERENCES

- [1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* 24 (2011).
- [2] Nicolás Aramayo, Mario Schiappacasse, and Marcel Goic. 2023. A multiarmed bandit approach for house ads recommendations. *Marketing Science* 42, 2 (2023), 271–292.
- [3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
- [4] Siddhartha Banerjee, Sean R Sinclair, Milind Tambe, Lily Xu, and Christina Lee Yu. 2022. Artificial replay: a meta-algorithm for harnessing historical data in Bandits. *arXiv preprint arXiv:2210.00025* (2022).
- [5] Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. 2021. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research* 22, 7 (2021), 1–108.
- [6] Léon Bottou, Jonas Peters, Joaquin Quiñero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14, 1 (2013), 3207–3260.
- [7] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.
- [8] Jinzhi Bu, David Simchi-Levi, and Yunzong Xu. 2020. Online pricing with offline data: Phase transition and inverse square law. In *International Conference on Machine Learning*. PMLR, 1202–1210.
- [9] T Tony Cai and Zijian Guo. 2017. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. (2017).
- [10] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).
- [11] Ian Char, Youngseog Chung, Willie Neiswanger, Kirthevasan Kandasamy, Andrew O Nelson, Mark Boyer, Egemen Kolemen, and Jeff Schneider. 2019. Offline contextual bayesian optimization. *Advances in Neural Information Processing Systems* 32 (2019).
- [12] Xi Chen, Yining Wang, and Yuan Zhou. 2020. Dynamic assortment optimization with changing contextual information. *Journal of machine learning research* 21, 216 (2020), 1–44.
- [13] Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. 2022. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*. PMLR, 3773–3793.
- [14] Wang Chi Cheung and Lixing Lyu. 2024. Leveraging (Biased) Information: Multi-armed Bandits with Offline Data. *arXiv preprint arXiv:2405.02594* (2024).
- [15] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. 2024. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271* (2024).
- [16] Xiangxiang Dai, Zhiyong Wang, Jize Xie, Xutong Liu, and John CS Lui. 2024. Conversational recommendation with online learning and clustering on misspecified users. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2024), 7825–7838.
- [17] Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. 2024. Active preference optimization for sample efficient RLHF. *arXiv preprint arXiv:2402.10500* (2024).
- [18] Gerard Debreu. 1960. Individual choice behavior: A theoretical analysis.
- [19] Yaqi Duan and Kaizheng Wang. 2023. Adaptive and robust multi-task learning. *The Annals of Statistics* 51, 5 (2023), 2015–2039.
- [20] Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. 2015. Contextual dueling bandits. In *Conference on Learning Theory*. PMLR, 563–587.
- [21] Fredrik Ek and Robert Stigsson. 2015. Recommender Systems; Contextual Multi-Armed Bandit Algorithms for the purpose of targeted advertisement within e-commerce. (2015).
- [22] Claudio Gentile, Shuai Li, and Giovanni Zappella. 2014. Online clustering of bandits. In *International conference on machine learning*. PMLR, 757–765.
- [23] Lin Gui, Cristina Gărbacea, and Victor Veitch. 2024. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *arXiv preprint arXiv:2406.00832* (2024).
- [24] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564* (2023).

- [25] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456* (2019).
- [26] Kaixuan Ji, Jiafan He, and Quanan Gu. 2024. Reinforcement learning from human feedback with active queries. *arXiv preprint arXiv:2402.09401* (2024).
- [27] Tammy Jiang, Jaimie L Gradus, and Anthony J Rosellini. 2020. Supervised machine learning: a brief primer. *Behavior therapy* 51, 5 (2020), 675–687.
- [28] Ying Jin, Zhuoran Yang, and Zhaoran Wang. 2021. Is pessimism provably efficient for offline rl?. In *International Conference on Machine Learning*. PMLR, 5084–5096.
- [29] Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2025. Spread preference annotation: Direct preference judgment for efficient llm alignment. In *The Thirteenth International Conference on Learning Representations*.
- [30] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453* (2023).
- [31] Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*. Springer, 45–73.
- [32] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. (2023).
- [33] Joongkyu Lee and Min-hwan Oh. 2024. Nearly minimax optimal regret for multinomial logistic bandit. *Advances in Neural Information Processing Systems* 37 (2024), 109003–109065.
- [34] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).
- [35] Gene Li, Cong Ma, and Nati Srebro. 2022. Pessimism for Offline Linear Contextual Bandits using ℓ_p Confidence Sets. *Advances in Neural Information Processing Systems* 35 (2022), 20974–20987.
- [36] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.
- [37] Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. 2025. Provably Efficient RLHF Pipeline: A Unified View from Contextual Bandits. *arXiv preprint arXiv:2502.07193* (2025).
- [38] Shuai Li, Wei Chen, and Kwong-Sak Leung. 2019. Improved algorithm on online clustering of bandits. *arXiv preprint arXiv:1902.09162* (2019).
- [39] Shuai Li and Shengyu Zhang. 2018. Online clustering of contextual cascading bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [40] Xiang Li, Viraj Mehta, Johannes Kirschner, Ian Char, Willie Neiswanger, Jeff Schneider, Andreas Krause, and Ilija Bogunovic. 2022. Near-optimal policy identification in active reinforcement learning. *arXiv preprint arXiv:2212.09510* (2022).
- [41] Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133* (2024).
- [42] Zhuohua Li, Maoli Liu, Xiangxiang Dai, and John Lui. 2025. Demystifying online clustering of bandits: Enhanced exploration under stochastic and smoothed adversarial contexts. *arXiv preprint arXiv:2501.00891* (2025).
- [43] Zihao Li, Zhuoran Yang, and Mengdi Wang. 2023. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438* (2023).
- [44] Jingyuan Liu, Zeyu Zhang, Xuchuang Wang, Xutong Liu, John Lui, Mohammad Hajiesmaili, and Carlee Joe-Wong. 2025. Offline Clustering of Linear Bandits: Unlocking the Power of Clusters in Data-Limited Environments. *arXiv preprint arXiv:2505.19043* (2025).
- [45] Pangpang Liu, Chengchun Shi, and Will Wei Sun. 2024. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504* (2024).
- [46] Xutong Liu, Haoru Zhao, Tong Yu, Shuai Li, and John CS Lui. 2022. Federated online clustering of bandits. In *Uncertainty in Artificial Intelligence*. PMLR, 1221–1231.
- [47] James B McQueen. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.* 281–297.
- [48] Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. 2023. Sample efficient reinforcement learning from human feedback via active exploration. *arXiv preprint arXiv:2312.00267* (2023).
- [49] Cataldo Musto, Fedelucio Narducci, Marco Polignano, Marco De Gemmis, Pasquale Lops, and Giovanni Semeraro. 2021. Myrrorbot: A digital assistant based on holistic user models for personalized access to online services. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–34.

- [50] Min-hwan Oh and Garud Iyengar. 2019. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems* 32 (2019).
- [51] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
- [52] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. 2024. Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254* (2024).
- [53] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075* (2024).
- [54] Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems* 37 (2024), 37100–37137.
- [55] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. 2021. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems* 34 (2021), 11702–11716.
- [56] Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715* (2024).
- [57] Aadirupa Saha. 2021. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems* 34 (2021), 30050–30062.
- [58] Aadirupa Saha and Akshay Krishnamurthy. 2022. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*. PMLR, 968–994.
- [59] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* 42, 3 (2017), 1–21.
- [60] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [61] Burr Settles. 2009. Active learning literature survey. (2009).
- [62] Pannagadatta Shivaswamy and Thorsten Joachims. 2012. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*. PMLR, 1046–1054.
- [63] Adish Singla, Anna N Rafferty, Goran Radanovic, and Neil T Heffernan. 2021. Reinforcement learning for education: Opportunities and challenges. *arXiv preprint arXiv:2107.08828* (2021).
- [64] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems* 33 (2020), 3008–3021.
- [65] Maurice E Stucke and Ariel Ezrachi. 2017. How digital assistants can harm our economy, privacy, and democracy. *Berkeley Technology Law Journal* 32, 3 (2017), 1239–1300.
- [66] Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Alexey Naumov, Pierre Perrault, Michal Valko, and Pierre Menard. 2023. Regularized rl. *arXiv preprint arXiv:2310.17303* (2023).
- [67] Rajat Verma, Vishal Nagar, and Satyasundara Mahapatra. 2021. Introduction to supervised learning. *Data Analytics in Bioinformatics: A Machine Learning Perspective* (2021), 1–34.
- [68] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*. 59–63.
- [69] Akifumi Wachi, Thien Tran, Rei Sato, Takumi Tanabe, and Youhei Akimoto. 2024. Stepwise alignment for constrained language model policy optimization. *Advances in Neural Information Processing Systems* 37 (2024), 104471–104520.
- [70] Lequn Wang, Akshay Krishnamurthy, and Alex Slivkins. 2024. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 766–774.
- [71] Yuanhao Wang, Qinghua Liu, and Chi Jin. 2023. Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems* 36 (2023), 76006–76032.
- [72] Zhiyong Wang, Jiahang Sun, Mingze Kong, Jize Xie, Qinghua Hu, John Lui, and Zhongxiang Dai. 2025. Online Clustering of Dueling Bandits. *arXiv preprint arXiv:2502.02079* (2025).
- [73] Zhiyong Wang, Jize Xie, Xutong Liu, Shuai Li, and John Lui. 2023. Online clustering of bandits with misspecified user models. *Advances in Neural Information Processing Systems* 36 (2023), 3785–3818.
- [74] Zhiyong Wang, Jize Xie, Tong Yu, Shuai Li, and John Lui. 2023. Online corrupted user detection and regret minimization. *Advances in Neural Information Processing Systems* 36 (2023), 33262–33287.
- [75] Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. 2021. On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*. PMLR, 11362–11371.

- [76] Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. 2024. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455* (2024).
- [77] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. 2021. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems* 34 (2021), 6683–6694.
- [78] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2023. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456* (2023).
- [79] Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. 2020. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems* 33 (2020), 18784–18794.
- [80] Cairong Yan, Haixia Han, Yanting Zhang, Dandan Zhu, and Yongquan Wan. 2022. Dynamic clustering based contextual combinatorial multi-armed bandit for online recommendation. *Knowledge-Based Systems* 257 (2022), 109927.
- [81] Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. 2024. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems* 37 (2024), 81773–81807.
- [82] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. 2012. The k-armed dueling bandits problem. *J. Comput. System Sci.* 78, 5 (2012), 1538–1556.
- [83] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access* 8 (2020), 58443–58469.
- [84] Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. 2023. Provable offline reinforcement learning with human feedback. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*.
- [85] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. 2023. How to Query Human Feedback Efficiently in RL? (2023).
- [86] Chicheng Zhang, Alekh Agarwal, Hal Daumé III, John Langford, and Sahand N Negahban. 2019. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301* (2019).
- [87] Cun-Hui Zhang and Stephanie S Zhang. 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76, 1 (2014), 217–242.
- [88] Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. 2024. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006* (2024).
- [89] Banghua Zhu, Michael Jordan, and Jiantao Jiao. 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*. PMLR, 43037–43067.

Table 3. Summary of key notations.

Notation	Description
U, \mathcal{U}	Number of users and the user set $\{1, \dots, U\}$.
J	Number of clusters which is unknown to the learner.
θ_u	True d -dimensional preference vector of user u with $\ \theta_u\ _2 \leq 1$.
θ^j	Preference shared by all users in cluster j .
$\mathcal{U}(j)$	Users in cluster j .
$\phi(\mathbf{x}, \mathbf{a})$	Feature map $\phi : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$ with $\ \phi(\mathbf{x}, \mathbf{a})\ _2 \leq 1$.
\mathcal{D}_u	Offline data of user u : $\{(\mathbf{x}_u^i, \mathbf{a}_u^i, \mathbf{a}'_u^i, y_u^i)\}_{i=1}^{N_u}$.
\mathbf{z}_u^i	Feature difference $\phi(\mathbf{x}_u^i, \mathbf{a}_u^i) - \phi(\mathbf{x}_u^i, \mathbf{a}'_u^i)$.
$\sigma(\cdot)$	Sigmoid function in the BTL preference model.
κ	Non-linearity coefficient (Equation (10)); lower bound on $\nabla \sigma(\cdot)$ across comparisons.
M_u	Regularized Gramian from \mathcal{D}_u : $\frac{\lambda}{\kappa} I + \sum_{i \in \mathcal{D}_u} \mathbf{z}_u^i (\mathbf{z}_u^i)^\top$.
$\lambda_{\min}(M)$	Minimum eigenvalue of matrix M .
CI_u	Confidence radius for the MLE $\hat{\theta}_u$.
$\hat{\gamma}$	Clustering threshold controlling when two users are connected.
$\mathcal{V}_{\hat{\gamma}}(u)$	Set of user u and its neighbors connected under threshold $\hat{\gamma}$.
\tilde{M}_u, \tilde{N}_u	Aggregated Gramian and sample count over $\mathcal{V}_{\hat{\gamma}}(u)$.
π_u^*	Optimal policy for user u .
$\text{SubOpt}_u(\pi)$	Suboptimality gap of policy π for user u (Equation (1)).

APPENDIX

A DETAILED DISCUSSION OF REMARK 2

This appendix elaborates practical policies for choosing the clustering threshold $\hat{\gamma}$. Our treatment closely follows the guidance in Liu et al. [44]; we include their spirit here for completeness and refer readers there for additional discussion.

A.1 Case 1: Known γ

When the minimum heterogeneity gap γ (defined in Definition 1) is known, a natural choice is $\hat{\gamma} = \gamma$, which exactly separates users across clusters.

Remark 7 (Discussions on γ Known Cases). Setting $\hat{\gamma} = \gamma$ eliminates bias from heterogeneous neighbors because the graph connects only users with the same preference vectors, implying $\mathcal{W}_{\hat{\gamma}}(u_t) = \emptyset$. The bound thus reflects only sampling noise from the homogeneous neighborhood $\mathcal{V}_{\hat{\gamma}}(u_t)$. Lemma 2 and Equation (9) together show that setting $\hat{\gamma} = \gamma$ allows Algorithm 1 to maximize $\mathcal{R}_{\hat{\gamma}}(u_t)$ while still ensuring zero bias, making this choice practical. Notably, choosing $\hat{\gamma} < \gamma$ would also make $\mathcal{W}_{\hat{\gamma}}(u_t) = \emptyset$, but at the cost of potentially shrinking $\mathcal{R}_{\hat{\gamma}}(u_t)$ and losing valuable homogeneous samples which leads to smaller $\mathcal{V}_{\hat{\gamma}}(u_t)$ and thus increases the noise.

A.2 Case 2: Unknown γ

When γ is unknown, the threshold $\hat{\gamma}$ must be estimated from the offline data. We define

$$\Gamma(u, v) = \|\hat{\theta}_u - \hat{\theta}_v\|_2 - \alpha(\text{CI}_u + \text{CI}_v), \quad M(u) = \{v \in \mathcal{U} \setminus \{u\} : \Gamma(u, v) > 0\}, \quad (16)$$

where CI_u is given in Equation (2). For $\alpha \geq 1$, $\Gamma(u, v) \leq \|\theta_u - \theta_v\|_2$ is a lower bound on the true preference gap, and $M(u)$ collects users deemed heterogeneous relative to u . We consider two complementary policies.

Definition 3 (Underestimation policy). *The underestimation policy is defined as:*

$$\hat{\gamma} = \mathbb{I}\{M(u_t) \neq \emptyset\} \cdot \min_{v \in M(u_t)} \Gamma(u_t, v). \quad (17)$$

Theorem 3 (Effect of the underestimation policy). *With $\hat{\gamma}$ chosen by Equation (17) and $\alpha'_w = \frac{\kappa}{3(\alpha+1)\sqrt{2 \max\{2, d\} \log(2U/\delta)}}$, any user v in the heterogeneous neighbor set $\mathcal{W}_{\hat{\gamma}}(u_t)$ of Lemma 2 also satisfies*

$$\frac{1}{\sqrt{\lambda_{\min}(M_{u_t})}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} \geq \alpha'_w \|\theta_{u_t} - \theta_v\|_2.$$

Remark 8 (When an underestimation policy is preferable). This conservative choice keeps $\mathcal{W}_{\hat{\gamma}}(u_t)$ small—only users with limited information enter—thereby controlling bias. The tradeoff is fewer homogeneous neighbors ($\mathcal{R}_{\hat{\gamma}}(u_t)$ and $\mathcal{V}_{\hat{\gamma}}(u_t)$ may shrink), which can increase noise. It is therefore preferable when bias is the primary concern—for example, in RLHF with annotators from diverse regions where mis-clustering can inject systematic preference bias or in fairness-sensitive applications (e.g., healthcare or education) where even small cross-group bias is more harmful than the extra noise from using fewer neighbors.

Definition 4 (Overestimation policy). *The overestimation policy is defined as:*

$$\hat{\gamma} = \mathbb{I}\{M(u_t) \neq \emptyset\} \cdot \min_{v \in M(u_t)} \tilde{\Gamma}(u_t, v), \quad (18)$$

where $\tilde{\Gamma}(u_t, v) = \|\hat{\theta}_{u_t} - \hat{\theta}_v\|_2 + \alpha(CI_{u_t} + CI_v)$ is an upper bound on the gap between users u_t and v .

Theorem 4 (Effect of the overestimation policy). *Under the policy in Definition 4, if $M(u_t) \neq \emptyset$ then $\hat{\gamma} \geq \gamma$.*

Remark 9 (When an overestimation policy is preferable). Ensuring $\hat{\gamma} \geq \gamma$ expands both the homogeneous neighbor set $\mathcal{R}_{\hat{\gamma}}(u_t)$ and the heterogeneous neighbor set $\mathcal{W}_{\hat{\gamma}}(u_t)$. This typically reduces noise but may also increase bias through more heterogeneous neighbors. This policy is therefore well-suited to noise-dominated regimes, such as recommendation cohorts with sparse but relatively homogeneous histories; or high-dimension scenarios where the number of dimensions d is large.

Both policies introduced here have their advantages and disadvantages. Underestimation reduces bias at the expense of higher noise; while overestimation does the opposite. In practice, the preferred policy depends on whether bias or noise is the main bottleneck. For additional discussion and complementary proofs of Lemmas 3 and 4, see Liu et al. [44].

B DETAILED PROOFS

B.1 Proof of Lemma 1

Proof. First, for any $\theta_s \in \mathbb{R}^d$, define

$$G_u(\theta_s) := \sum_{i=1}^{N_u} (\sigma(\theta_s^\top z_u^i) - \sigma(\theta_u^\top z_u^i)) z_u^i + \lambda \theta_s.$$

By the mean value theorem, for any two parameter vectors θ_{s_1} and θ_{s_2} , we have

$$G_u(\theta_{s_1}) - G_u(\theta_{s_2}) = \left(\sum_{i=1}^{N_u} \nabla \sigma(\theta_s^\top z_u^i) z_u^i (z_u^i)^\top + \lambda I \right) (\theta_{s_1} - \theta_{s_2}) = W_u(\theta_{s_1} - \theta_{s_2}),$$

where we define

$$W_u := \sum_{i=1}^{N_u} \nabla \sigma(\theta_s^\top z_u^i) z_u^i (z_u^i)^\top + \lambda I \quad \text{and} \quad \theta_s = \xi \theta_{s_1} + (1 - \xi) \theta_{s_2}, \quad \xi \in [0, 1].$$

In particular, for each user $u \in \mathcal{U}$, the mean value theorem implies that there exists $\xi_u \in [0, 1]$ such that the intermediate point is given by $\theta_u = \xi_u \theta_u + (1 - \xi_u) \hat{\theta}_u$.

Furthermore, we define

$$W_u := \sum_{i=1}^{N_u} \nabla \sigma(\theta_u^\top z_u^i) z_u^i (z_u^i)^\top + \lambda I.$$

Recall that

$$M_u = \sum_{i=1}^{N_u} z_u^i (z_u^i)^\top + \frac{\lambda}{\kappa} I.$$

By Equation (10), we have $W_u \succeq \kappa M_u$ and $M_u^{-1} \succeq \kappa W_u^{-1}$ since $\nabla \sigma(\theta_u^\top z_u^i) \geq \kappa$. Here, for two symmetric matrices A_1 and A_2 , the notation $A_1 \succeq A_2$ means that $A_1 - A_2$ is positive semi-definite.

Using these properties, we can show that

$$\begin{aligned} \|G_u(\hat{\theta}_u) - \lambda \theta_u\|_{M_u^{-1}}^2 &= \|G_u(\hat{\theta}_u) - G_u(\theta_u)\|_{M_u^{-1}}^2 = \|W_u(\theta_u - \hat{\theta}_u)\|_{M_u^{-1}}^2 \\ &= (\theta_u - \hat{\theta}_u)^\top W_u M_u^{-1} W_u (\theta_u - \hat{\theta}_u) \\ &\stackrel{(a)}{\geq} \kappa (\theta_u - \hat{\theta}_u)^\top W_u (\theta_u - \hat{\theta}_u) \\ &\stackrel{(b)}{\geq} \kappa^2 (\theta_u - \hat{\theta}_u)^\top M_u (\theta_u - \hat{\theta}_u) = \kappa^2 \|\theta_u - \hat{\theta}_u\|_{M_u}^2, \end{aligned} \quad (19)$$

where (a) follows from $M_u^{-1} \succeq \kappa W_u^{-1}$ and (b) from $W_u \succeq \kappa M_u$.

Moreover, observe that

$$\|\lambda \theta_u\|_{M_u^{-1}} = \lambda \sqrt{\theta_u^\top M_u^{-1} \theta_u} \leq \sqrt{\lambda \kappa} \|\theta_u\|_2 \leq \sqrt{\lambda \kappa}, \quad (20)$$

where the first inequality uses $M_u \succeq \frac{\lambda}{\kappa} I$ and the second follows from $\|\theta_u\|_2 \leq 1$.

Combining these results, we have

$$\begin{aligned} \|\theta_u - \hat{\theta}_u\|_{M_u} &\stackrel{(a)}{\leq} \frac{1}{\kappa} \|G_u(\hat{\theta}_u) - \lambda \theta_u\|_{M_u^{-1}} \\ &\stackrel{(b)}{\leq} \frac{1}{\kappa} \|G_u(\hat{\theta}_u)\|_{M_u^{-1}} + \frac{1}{\kappa} \|\lambda \theta_u\|_{M_u^{-1}} \\ &\stackrel{(c)}{\leq} \frac{1}{\kappa} \|G_u(\hat{\theta}_u)\|_{M_u^{-1}} + \sqrt{\frac{\lambda}{\kappa}}, \end{aligned} \quad (21)$$

where (a) follows from (19), (b) uses the triangle inequality, and (c) applies (20).

We then bound the term $\|G_u(\hat{\theta}_u)\|_{M_u^{-1}}$ as follows:

$$\|G_u(\hat{\theta}_u)\|_{M_u^{-1}} = \left\| \sum_{i=1}^{N_u} \left(\sigma(\hat{\theta}_u^\top z_u^i) - \sigma(\theta_u^\top z_u^i) \right) z_u^i + \lambda \hat{\theta}_u \right\|_{M_u^{-1}}$$

$$\begin{aligned}
&= \left\| \sum_{i=1}^{N_u} \left(\sigma(\hat{\theta}_u^\top \mathbf{z}_u^i) - (y_u^i - \varepsilon_u^i) \right) \mathbf{z}_u^i + \sum_{i=1}^{N_u} \varepsilon_u^i \mathbf{z}_u^i + \lambda \hat{\theta}_u \right\|_{M_u^{-1}} \\
&\stackrel{(a)}{\leq} \left\| \sum_{i=1}^{N_u} \varepsilon_u^i \mathbf{z}_u^i \right\|_{M_u^{-1}}, \tag{22}
\end{aligned}$$

where inequality (a) follows from the fact that $\hat{\theta}_u$ is chosen to minimize the regularized log-likelihood:

$$\hat{\theta}_u = \arg \min_{\theta} \left[- \sum_{i=1}^{N_u} (y_u^i \log \sigma(\theta^\top \mathbf{z}_u^i) + (1 - y_u^i) \log \sigma(-\theta^\top \mathbf{z}_u^i)) + \frac{\lambda}{2} \|\theta\|_2^2 \right], \tag{23}$$

and thus its gradient satisfies

$$\sum_{i=1}^{N_u} \left(\sigma(\hat{\theta}_u^\top \mathbf{z}_u^i) - y_u^i \right) \mathbf{z}_u^i + \lambda \hat{\theta}_u = 0.$$

Therefore, it follows from (22) that

$$\frac{1}{\kappa} \|G_u(\hat{\theta}_u)\|_{M_u^{-1}} \leq \frac{1}{\kappa} \left\| \sum_{i=1}^{N_u} \varepsilon_u^i \mathbf{z}_u^i \right\|_{M_u^{-1}}.$$

Next, let $V = \frac{\lambda}{\kappa} I$. Since ε_u^i is 2-subgaussian, we apply Theorem 1 in Abbasi-Yadkori et al. [1] to obtain

$$\left\| \sum_{i=1}^{N_u} \varepsilon_u^i \mathbf{z}_u^i \right\|_{M_u^{-1}}^2 \leq 8 \log \left(\frac{\det(M_u)^{1/2}}{\delta \det(V)^{1/2}} \right) \tag{24}$$

with probability at least $1 - \delta$. Since $\|\mathbf{z}_u^i\|_2 \leq 2$, we have

$$\det(M_u) \leq \left(\frac{\lambda}{\kappa} + \frac{4N_u}{d} \right)^d, \quad \det(V) = \left(\frac{\lambda}{\kappa} \right)^d, \quad \text{and thus} \quad \sqrt{\frac{\det(M_u)}{\det(V)}} \leq \left(1 + \frac{4N_u\kappa}{d\lambda} \right)^{d/2}.$$

Therefore,

$$\left\| \sum_{i=1}^{N_u} \varepsilon_u^i \mathbf{z}_u^i \right\|_{M_u^{-1}}^2 \leq 8 \log \left(\frac{1}{\delta} \right) + 4d \log \left(1 + \frac{4N_u\kappa}{d\lambda} \right) \quad \text{with probability at least } 1 - \delta.$$

Putting everything together, we conclude that

$$\|\theta_u - \hat{\theta}_u\|_{M_u} \leq \frac{\sqrt{\lambda\kappa} + 2\sqrt{2 \log(1/\delta) + d \log(1 + 4N_u\kappa/(d\lambda))}}{\kappa} \quad \text{with probability at least } 1 - \delta,$$

which follows from combining (19), (21), (22), and (24). \blacksquare

B.2 Proof of Lemma 2

Proof. In order to prove Lemma 2, it suffices to show the following statement: under the same conditions as in Lemma 2, both sets can be characterized as

$$\begin{aligned}
\mathcal{R}_{\hat{\gamma}}(u) &= \left\{ v \mid \theta_u = \theta_v \text{ and } \frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha_r \hat{\gamma} \right\} \cup \{u\}, \\
\mathcal{W}_{\hat{\gamma}}(u) &= \left\{ v \mid \gamma \leq \|\theta_u - \theta_v\|_2 < \hat{\gamma} \text{ and } \frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha_w \varepsilon \right\}
\end{aligned}$$

for some $\alpha_r \in \left(\frac{\kappa}{3(\alpha+1)\sqrt{2\max\{2,d\}\log(2U/\delta)}}, \frac{\kappa}{2(\alpha-1)\sqrt{2\log(2U/\delta)}} \right)$ and $\alpha_w \in \left(0, \frac{\kappa}{2(\alpha-1)\sqrt{2\log(2U/\delta)}} \right)$ with probability at least $1 - \delta$.

First, by applying Lemma 1 and a union bound, we have that the event

$$\mathcal{E} \doteq \bigcap_{u \in \mathcal{U}} \left\{ \|\hat{\theta}_u - \theta_u\|_2 \leq \text{CI}_u \right\}$$

holds with probability at least $1 - \delta/2$.

Recall that the connection condition in Algorithm 1 is given by

$$\left\| \hat{\theta}_{u_1} - \hat{\theta}_{u_2} \right\|_2 < \hat{\gamma} - \alpha (\text{CI}_{u_1} + \text{CI}_{u_2}),$$

which implies

$$\begin{aligned} \hat{\gamma} &> \left\| \hat{\theta}_{u_1} - \hat{\theta}_{u_2} \right\|_2 + \alpha (\text{CI}_{u_1} + \text{CI}_{u_2}) \\ &\geq \left\| \hat{\theta}_{u_1} - \hat{\theta}_{u_2} \right\|_2 + \text{CI}_{u_1} + \text{CI}_{u_2} \\ &\stackrel{(a)}{\geq} \left\| \hat{\theta}_{u_1} - \hat{\theta}_{u_2} \right\|_2 + \left\| \hat{\theta}_{u_1} - \theta_{u_1} \right\|_2 + \left\| \hat{\theta}_{u_2} - \theta_{u_2} \right\|_2 \\ &\stackrel{(b)}{\geq} \left\| \theta_{u_1} - \theta_{u_2} \right\|_2, \end{aligned}$$

where (a) follows from the event \mathcal{E} and (b) follows by the triangle inequality. Therefore, any pair of connected users must have preference vectors whose difference is no greater than $\hat{\gamma}$.

Next, we calculate the cardinality of $\mathcal{R}_{\hat{\gamma}}(u)$. Note that for any user $v \in \mathcal{R}_{\hat{\gamma}}(u)$, it holds that $\theta_u = \theta_v$. To prove the claim for $\mathcal{R}_{\hat{\gamma}}(u)$ in Lemma 2, it suffices to show the following two conditions under event \mathcal{E} :

- (i) If $\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \frac{\kappa\hat{\gamma}}{3(\alpha+1)\sqrt{2\max\{2,d\}\log(2U/\delta)}}$ then v must be included in $\mathcal{R}_{\hat{\gamma}}(u)$.
- (ii) If $\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} \geq \frac{\kappa\hat{\gamma}}{2(\alpha-1)\sqrt{2\log(2U/\delta)}}$ then v must not be included in $\mathcal{R}_{\hat{\gamma}}(u)$.

For (i). Given

$$\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \frac{\kappa\hat{\gamma}}{3(\alpha+1)\sqrt{2\max\{2,d\}\log(2U/\delta)}},$$

we have

$$(\alpha+1)(\text{CI}_u + \text{CI}_v) \tag{25}$$

$$\begin{aligned} &\leq \frac{3(\alpha+1)\sqrt{2\log(2U/\delta)} + d\log(1+4N_u\kappa/(d\lambda))}{\kappa\sqrt{\lambda_{\min}(M_u)}} + \frac{3(\alpha+1)\sqrt{2\log(2U/\delta)} + d\log(1+4N_v\kappa/(d\lambda))}{\kappa\sqrt{\lambda_{\min}(M_v)}} \\ &\leq \frac{3(\alpha+1)\sqrt{2\max\{2,d\}\log(2U/\delta)}}{\kappa} \left(\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} \right) < \hat{\gamma}, \end{aligned} \tag{26}$$

where the second last inequality holds if λ and δ satisfy $\lambda\kappa \leq 2\log(2U/\delta) + d\log(1+4N_s\kappa/(d\lambda))$ and $\delta \leq d\lambda/(4N_s\kappa + d\lambda)$ for all $s \in \mathcal{U}$.

Therefore, under event \mathcal{E} , we obtain

$$\left\| \hat{\theta}_u - \hat{\theta}_v \right\|_2 \leq \left\| \theta_u - \theta_v \right\|_2 + \text{CI}_u + \text{CI}_v \stackrel{(a)}{=} \text{CI}_u + \text{CI}_v \stackrel{(b)}{\leq} \hat{\gamma} - \alpha(\text{CI}_u + \text{CI}_v),$$

where (a) uses $\theta_u = \theta_v$, and (b) follows from (25). Hence the connection condition in Equation (3) holds, which implies that v will be connected to u with probability at least $1 - \delta$.

For (ii). If

$$\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} \geq \frac{\kappa \hat{\gamma}}{2(\alpha-1)\sqrt{2\log(2U/\delta)}},$$

then we have

$$\begin{aligned} (\alpha-1)(\text{CI}_u + \text{CI}_v) &\geq \frac{2(\alpha-1)}{\kappa} \sqrt{\frac{2\log(2U/\delta)}{\lambda_{\min}(M_u)}} + \frac{2(\alpha-1)}{\kappa} \sqrt{\frac{2\log(2U/\delta)}{\lambda_{\min}(M_v)}} \\ &\geq \hat{\gamma}. \end{aligned} \quad (27)$$

Therefore, it follows that

$$\hat{\gamma} - \alpha(\text{CI}_u + \text{CI}_v) \leq -(\text{CI}_u + \text{CI}_v) = \|\theta_u - \theta_v\|_2 - (\text{CI}_u + \text{CI}_v) \leq \|\hat{\theta}_u - \hat{\theta}_v\|_2.$$

Hence, the connection condition in Equation (3) does not hold under event \mathcal{E} . This verifies that any v satisfying this bound cannot be included in $\mathcal{R}_{\hat{\gamma}}(u)$, implying

$$\alpha_r \in \left(\frac{\kappa}{3(\alpha+1)\sqrt{2\max\{2,d\}\log(2U/\delta)}}, \frac{\kappa}{2(\alpha-1)\sqrt{2\log(2U/\delta)}} \right).$$

For the cardinality of $\mathcal{W}_{\hat{\gamma}}(u)$, note that since both $\lambda_{\min}(M_u)$ and $\lambda_{\min}(M_v)$ are positive, we trivially have $\alpha_w > 0$. It remains to show that any heterogeneous user v with

$$\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} \geq \frac{\kappa \hat{\gamma}}{2(\alpha-1)\sqrt{2\log(2U/\delta)}}$$

cannot be included in $\mathcal{W}_{\hat{\gamma}}(u)$ under event \mathcal{E} . By the same argument as in (27), we have $(\alpha-1)(\text{CI}_u + \text{CI}_v) \geq \varepsilon$. This yields

$$\varepsilon - \alpha(\text{CI}_u + \text{CI}_v) \leq -(\text{CI}_u + \text{CI}_v) \leq \|\theta_u - \theta_v\|_2 - (\text{CI}_u + \text{CI}_v) - \gamma \leq \|\hat{\theta}_u - \hat{\theta}_v\|_2 - \gamma,$$

which implies

$$\hat{\gamma} - \alpha(\text{CI}_u + \text{CI}_v) \leq \|\hat{\theta}_u - \hat{\theta}_v\|_2.$$

Thus, the connection condition in Equation (3) does not hold for such v , confirming that it cannot be included in $\mathcal{W}_{\hat{\gamma}}(u)$. ■

B.3 Proof of Lemma 3

Proof. First, we define

$$\tilde{G}_u(\theta_s) = \sum_{v \in \mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} (\sigma(\theta_s^\top z_v^i) - \sigma(\theta_u^\top z_v^i)) z_v^i + \lambda \theta_s, \quad \forall \theta_s \in \mathbb{R}^d.$$

By the mean value theorem, for any θ_{s_1} and θ_{s_2} , we have

$$\tilde{G}_u(\theta_{s_1}) - \tilde{G}_u(\theta_{s_2}) = \left(\sum_{v \in \mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \nabla \sigma(\theta_s^\top z_v^i) z_v^i z_v^{i\top} + \lambda I \right) (\theta_{s_1} - \theta_{s_2}),$$

for some intermediate point $\theta_s = \xi \theta_{s_1} + (1-\xi) \theta_{s_2}$ with $\xi \in [0, 1]$. In particular, for each $u \in \mathcal{U}$, we let $\xi_u \in [0, 1]$ and define the corresponding intermediate point $\theta_u = \xi_u \theta_u + (1-\xi_u) \hat{\theta}_u$.

We further define

$$\tilde{W}_u = \sum_{v \in \mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \nabla \sigma(\theta_u^\top z_v^i) z_v^i z_v^{i\top} + \lambda I \quad \text{and} \quad \tilde{M}_u = \sum_{v \in \mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} z_v^i z_v^{i\top} + \frac{\lambda}{\kappa} I.$$

By construction, it holds that $\tilde{W}_u \succeq \kappa \tilde{M}_u$ and thus $\tilde{M}_u^{-1} \succeq \kappa \tilde{W}_u^{-1}$ for all $u \in \mathcal{U}$.

Then, we have

$$\begin{aligned}
 \|\tilde{G}_u(\tilde{\theta}_u) - \lambda \theta_u\|_{\tilde{M}_u^{-1}}^2 &= \|\tilde{G}_u(\tilde{\theta}_u) - \tilde{G}_u(\theta_u)\|_{\tilde{M}_u^{-1}}^2 = \|\tilde{W}_u(\theta_u - \tilde{\theta}_u)\|_{\tilde{M}_u^{-1}}^2 \\
 &= (\theta_u - \tilde{\theta}_u)^\top \tilde{W}_u \tilde{M}_u^{-1} \tilde{W}_u (\theta_u - \tilde{\theta}_u) \\
 &\stackrel{(a)}{\geq} \kappa (\theta_u - \tilde{\theta}_u)^\top \tilde{W}_u (\theta_u - \tilde{\theta}_u) \\
 &\stackrel{(b)}{\geq} \kappa^2 (\theta_u - \tilde{\theta}_u)^\top \tilde{M}_u (\theta_u - \tilde{\theta}_u) = \kappa^2 \|\theta_u - \tilde{\theta}_u\|_{\tilde{M}_u}^2,
 \end{aligned} \tag{28}$$

where (a) follows from $\tilde{M}_u^{-1} \succeq \kappa \tilde{W}_u^{-1}$ and (b) follows from $\tilde{W}_u \succeq \kappa \tilde{M}_u$.

Moreover, since $\tilde{M}_u \succeq \frac{\lambda}{\kappa} I$, we have

$$\|\lambda \theta_u\|_{\tilde{M}_u^{-1}} = \lambda \sqrt{\theta_u^\top \tilde{M}_u^{-1} \theta_u} \leq \lambda \sqrt{\theta_u^\top \left(\frac{\kappa}{\lambda} I\right) \theta_u} = \sqrt{\lambda \kappa} \|\theta_u\|_2 \leq \sqrt{\lambda \kappa}. \tag{29}$$

Hence, we obtain

$$\begin{aligned}
 \|\theta_u - \tilde{\theta}_u\|_{\tilde{M}_u} &\stackrel{(a)}{\leq} \frac{1}{\kappa} \|\tilde{G}_u(\tilde{\theta}_u) - \lambda \theta_u\|_{\tilde{M}_u^{-1}} \\
 &\stackrel{(b)}{\leq} \frac{1}{\kappa} \|\tilde{G}_u(\tilde{\theta}_u)\|_{\tilde{M}_u^{-1}} + \frac{1}{\kappa} \|\lambda \theta_u\|_{\tilde{M}_u^{-1}} \\
 &\stackrel{(c)}{\leq} \frac{1}{\kappa} \|\tilde{G}_u(\tilde{\theta}_u)\|_{\tilde{M}_u^{-1}} + \sqrt{\frac{\lambda}{\kappa}},
 \end{aligned} \tag{30}$$

where (a) follows from Equation (31), (b) applies the triangle inequality, and (c) uses the bound in Equation (29).

Furthermore, we can bound $\tilde{G}_u(\tilde{\theta}_u)$ as follows:

$$\begin{aligned}
 &\frac{1}{\kappa^2} \|\tilde{G}_u(\tilde{\theta}_u)\|_{\tilde{M}_u^{-1}}^2 \\
 &\stackrel{(a)}{=} \frac{1}{\kappa^2} \left\| \sum_{v \in \mathcal{V}_Y(u)} \sum_{i=1}^{N_v} \left(\sigma(\tilde{\theta}_u^\top z_v^i) - \sigma(\theta_u^\top z_v^i) \right) z_v^i + \lambda \tilde{\theta}_u \right\|_{\tilde{M}_u^{-1}}^2 \\
 &= \frac{1}{\kappa^2} \left\| \sum_v \sum_i \left(\sigma(\tilde{\theta}_u^\top z_v^i) - y_v^i + y_v^i - \sigma(\theta_u^\top z_v^i) \right) z_v^i + \lambda \tilde{\theta}_u \right\|_{\tilde{M}_u^{-1}}^2 \\
 &= \frac{1}{\kappa^2} \left\| \sum_v \sum_i \left(\sigma(\tilde{\theta}_u^\top z_v^i) - y_v^i \right) z_v^i + \lambda \tilde{\theta}_u + \sum_v \sum_i (y_v^i - \sigma(\theta_u^\top z_v^i)) z_v^i \right\|_{\tilde{M}_u^{-1}}^2 \\
 &\stackrel{(b)}{=} \frac{1}{\kappa^2} \left\| \sum_v \sum_i (y_v^i - \sigma(\theta_v^\top z_v^i) + \sigma(\theta_v^\top z_v^i) - \sigma(\theta_u^\top z_v^i)) z_v^i \right\|_{\tilde{M}_u^{-1}}^2 \\
 &= \frac{1}{\kappa^2} \left\| \underbrace{\sum_v \sum_i \varepsilon_v^i z_v^i}_{\text{noise}} + \underbrace{\sum_v \sum_i (\sigma(\theta_v^\top z_v^i) - \sigma(\theta_u^\top z_v^i)) z_v^i}_{\text{bias}} \right\|_{\tilde{M}_u^{-1}}^2 \\
 &\stackrel{(c)}{\leq} \left(\frac{1}{\kappa} \left\| \sum_v \sum_i \varepsilon_v^i z_v^i \right\|_{\tilde{M}_u^{-1}} + \frac{1}{\kappa} \left\| \sum_v \sum_i (\sigma(\theta_v^\top z_v^i) - \sigma(\theta_u^\top z_v^i)) z_v^i \right\|_{\tilde{M}_u^{-1}} \right)^2
 \end{aligned}$$

$$\stackrel{(d)}{=} \left(\frac{1}{\kappa} \left\| \sum_v \sum_i \varepsilon_v^i z_v^i \right\|_{\tilde{M}_u^{-1}} + \frac{1}{\kappa} \left\| \sum_{v \in \mathcal{W}_{\hat{Y}}(u)} \sum_i (\sigma(\theta_v^\top z_v^i) - \sigma(\theta_u^\top z_v^i)) z_v^i \right\|_{\tilde{M}_u^{-1}} \right)^2. \quad (31)$$

Here, (a) follows from the definition of $\tilde{G}_u(\tilde{\theta}_u)$; (b) holds since $\tilde{\theta}_u$ minimizes the negative log-likelihood regularized by λ , implying

$$\sum_v \sum_i \left(\sigma(\tilde{\theta}_u^\top z_v^i) - y_v^i \right) z_v^i + \lambda \tilde{\theta}_u = 0;$$

(c) uses the triangle inequality; and (d) uses the fact that for any homogeneous neighbor $v \in \mathcal{R}_{\hat{Y}}(u)$, we have $\theta_u = \theta_v$, so only the heterogeneous neighbors contribute to the bias term.

Next, we bound the term

$$\left\| \sum_{v \in \mathcal{W}_{\hat{Y}}(u)} \sum_{i=1}^{N_v} (\sigma(\theta_v^\top z_v^i) - \sigma(\theta_u^\top z_v^i)) z_v^i \right\|_{\tilde{M}_u^{-1}}.$$

By the triangle inequality, we have

$$\begin{aligned} & \left\| \sum_{v \in \mathcal{W}_{\hat{Y}}(u)} \sum_{i=1}^{N_v} (\sigma(\theta_v^\top z_v^i) - \sigma(\theta_u^\top z_v^i)) z_v^i \right\|_{\tilde{M}_u^{-1}} \\ & \leq \sum_{v \in \mathcal{W}_{\hat{Y}}(u)} \sum_{i=1}^{N_v} \left| \sigma(\theta_v^\top z_v^i) - \sigma(\theta_u^\top z_v^i) \right| \|z_v^i\|_{\tilde{M}_u^{-1}} \\ & \stackrel{(a)}{\leq} \sum_v \sum_i \frac{1}{4} \|\theta_v^\top z_v^i - \theta_u^\top z_v^i\| \|z_v^i\|_{\tilde{M}_u^{-1}} \leq \frac{\hat{Y}}{4} \sum_v \sum_i \|z_v^i\|_2 \|z_v^i\|_{\tilde{M}_u^{-1}} \\ & \stackrel{(b)}{\leq} \frac{\hat{Y}}{2} \sum_{v \in \mathcal{W}_{\hat{Y}}(u)} \sum_{i=1}^{N_v} \|z_v^i\|_{\tilde{M}_u^{-1}}, \end{aligned} \quad (32)$$

where (a) follows from the Lipschitz continuity of the sigmoid function with constant $L_\sigma = \frac{1}{4}$, and (b) uses $\|z_v^i\|_2 \leq 2$.

Furthermore, observe that

$$\sum_{v \in \mathcal{W}_{\hat{Y}}(u)} \sum_{i=1}^{N_v} \|z_v^i\|_{\tilde{M}_u^{-1}}^2 = \text{tr} \left(\tilde{M}_u^{-1} \left(\tilde{M}_u - \frac{\lambda}{\kappa} I \right) \right) \leq d.$$

By applying Cauchy–Schwarz inequality, we get

$$\sum_v \sum_i \|z_v^i\|_{\tilde{M}_u^{-1}} \leq \sqrt{\left(\sum_v N_v \right) \left(\sum_v \sum_i \|z_v^i\|_{\tilde{M}_u^{-1}}^2 \right)} \leq \sqrt{d \cdot N_{\mathcal{W}_{\hat{Y}}(u)}}. \quad (33)$$

Combining the above, the bias term due to heterogeneous neighbors is bounded accordingly.

Therefore, by applying Equation (32) and (33), we obtain

$$\left\| \sum_{v \in \mathcal{W}_{\hat{Y}}(u)} \sum_{i=1}^{N_v} (\sigma(\theta_v^\top z_v^i) - \sigma(\theta_u^\top z_v^i)) z_v^i \right\|_{\tilde{M}_u^{-1}} \leq \frac{\hat{Y}}{2} \sqrt{d N_{\mathcal{W}_{\hat{Y}}(u)}}, \quad (34)$$

where $N_{\mathcal{W}_{\hat{Y}}(u)} = \sum_{v \in \mathcal{W}_{\hat{Y}}(u)} N_v$.

Furthermore, for the noise term in Equation (31), by applying Theorem 1 in Abbasi-Yadkori et al. [1] with $V = \frac{\lambda}{\kappa}I$, we have

$$\left\| \sum_{v \in \mathcal{V}_{\hat{Y}}(u)} \sum_{i=1}^{N_v} \varepsilon_v^i z_v^i \right\|_{\tilde{M}_u^{-1}} \leq 2\sqrt{2 \log \left(\frac{\det(\tilde{M}_u)^{1/2}}{\delta \det(V)^{1/2}} \right)} \leq 2\sqrt{2 \log \left(\frac{1}{\delta} \right) + d \log \left(1 + \frac{4 N_{\mathcal{V}_{\hat{Y}}(u)} \kappa}{d\lambda} \right)} \quad (35)$$

with probability at least $1 - \delta$, where $N_{\mathcal{V}_{\hat{Y}}(u)} = \sum_{v \in \mathcal{V}_{\hat{Y}}(u)} N_v$.

Combining Equation (30), Equation (31), Equation (34), and (35), we finally have

$$\left\| \theta_u - \tilde{\theta}_u \right\|_{\tilde{M}_u} \leq \frac{\sqrt{\lambda \kappa} + 2\sqrt{2 \log \left(\frac{2U}{\delta} \right) + d \log \left(1 + \frac{4 N_{\mathcal{V}_{\hat{Y}}(u)} \kappa}{d\lambda} \right)}}{\kappa} + \frac{\hat{Y}}{2} \sqrt{d N_{\mathcal{W}_{\hat{Y}}(u)}},$$

which holds for all $u \in \mathcal{U}$ with probability at least $1 - \delta$. This completes the proof of Lemma 3. ■

B.4 Proof of Theorem 1

Proof. By Lemmas 1 and 3, we have

$$\left\| \theta_u - \tilde{\theta}_u \right\|_{\tilde{M}_u} \leq \tilde{\beta}_u + \frac{\hat{Y}}{2} \sqrt{d N_{\mathcal{W}_{\hat{Y}}(u)}} \quad (36)$$

for all $u \in \mathcal{U}$ with probability at least $1 - \delta$.

For simplicity, let $u = u_t$ denote the test user. Define $J'_u(\pi) = J_u(\pi) - \langle \theta_u, \mathbf{w} \rangle$. Then, the suboptimality gap can be written as:

$$\begin{aligned} \text{SubOpt}_u(\pi_u) &= J_u(\pi_u^*) - J_u(\pi_u) = J'_u(\pi_u^*) - J'_u(\pi_u) \\ &= \left(J'_u(\pi_u^*) - \tilde{J}_u(\pi_u^*) \right) + \left(\tilde{J}_u(\pi_u^*) - \tilde{J}_u(\pi_u) \right) + \left(\tilde{J}_u(\pi_u) - J'_u(\pi_u) \right). \end{aligned}$$

For the second term, since $\pi_u = \arg \max_{\pi} \tilde{J}_u(\pi)$, we have $\tilde{J}_u(\pi_u^*) - \tilde{J}_u(\pi_u) \leq 0$.

For the third term:

$$\begin{aligned} \tilde{J}_u(\pi_u) - J'_u(\pi_u) &= \left(\mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u(\mathbf{x}))] - \mathbf{w} \right)^\top (\tilde{\theta}_u - \theta_u) - \tilde{\beta}_u \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}} \\ &\leq \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}} \left(\left\| \tilde{\theta}_u - \theta_u \right\|_{\tilde{M}_u} - \tilde{\beta}_u \right) \\ &\stackrel{(a)}{\leq} \frac{\hat{Y}}{2} \sqrt{d N_{\mathcal{W}_{\hat{Y}}(u)}} \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}}, \end{aligned}$$

where (a) uses (36).

Similarly, for the first term:

$$\begin{aligned} J'_u(\pi_u^*) - \tilde{J}_u(\pi_u^*) &= (\theta_u - \tilde{\theta}_u)^\top \left(\mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u^*(\mathbf{x}))] - \mathbf{w} \right) + \tilde{\beta}_u \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u^*(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}} \\ &\leq \left(\left\| \theta_u - \tilde{\theta}_u \right\|_{\tilde{M}_u} + \tilde{\beta}_u \right) \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u^*(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}} \\ &\leq \left(2\tilde{\beta}_u + \frac{\hat{Y}}{2} \sqrt{d N_{\mathcal{W}_{\hat{Y}}(u)}} \right) \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u^*(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}}. \end{aligned}$$

Putting everything together, we obtain:

$$\text{SubOpt}_u(\pi_u) \leq \left(2\tilde{\beta}_u + \frac{\hat{Y}}{2} \sqrt{d N_{\mathcal{W}_{\hat{Y}}(u)}} \right) \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u^*(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}}$$

$$\begin{aligned}
&\leq \left(\frac{2\sqrt{2\log\left(\frac{2U}{\delta}\right) + d\log\left(1 + \frac{4\tilde{N}_u\kappa}{d\lambda}\right)}}{\kappa} + \hat{\gamma}\sqrt{dN_{\mathcal{W}_{\hat{\gamma}}(u)}} \right) \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u^*(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}} \\
&= \tilde{O} \left(\sqrt{d} \left(1 + \hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u)}} \right) \left\| \mathbb{E}_{\mathbf{x} \sim \rho_p} [\phi(\mathbf{x}, \pi_u^*(\mathbf{x}))] - \mathbf{w} \right\|_{\tilde{M}_u^{-1}} \right),
\end{aligned}$$

which concludes the proof of Theorem 1. \blacksquare

B.5 Proof of Lemma 4

Proof. In this proof, we define

$$\text{CI}_u = \frac{\sqrt{\lambda\kappa} + 2\sqrt{d\log\left(1 + \frac{4\kappa N_u}{\lambda d}\right) + 2\log\left(\frac{2U}{\delta}\right)}}{\kappa\sqrt{\tilde{\lambda}_a N_u/2}}.$$

By Lemma 1, Lemma J.1 in Wang et al. [73] and Lemma 7 in [39], it holds that $\lambda_{\min}(M_u) \geq \tilde{\lambda}_a N_u/2$ for all users connected to user u with probability at least $1 - \delta/2$. Therefore, we have

$$\left\| \hat{\theta}_u - \theta_u \right\|_2 \leq \frac{\sqrt{\lambda\kappa} + 2\sqrt{2\log\left(\frac{2U}{\delta}\right) + d\log\left(1 + \frac{4N_u\kappa}{d\lambda}\right)}}{\kappa\sqrt{\lambda_{\min}(M_u)}} \leq \text{CI}_u$$

with probability at least $1 - \delta$.

Finally, by following the same argument used in the proof of Lemma 2, but replacing $\lambda_{\min}(M_u)$ with the explicit bound on N_u under Assumption 1, we obtain the desired result in Lemma 4. \blacksquare

B.6 Proof of Corollary 1

Proof. We denote $\eta_{\mathcal{W}_{\hat{\gamma}}(u)} \doteq N_{\mathcal{W}_{\hat{\gamma}}(u)}/N_{\mathcal{V}_{\hat{\gamma}}(u)}$ for clarity, then it follows that

$$\begin{aligned}
\text{SubOpt}_u(\pi_u) &\leq \tilde{O} \left(\frac{\sqrt{d} \left(1 + \hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}} \right)}{\sqrt{\lambda_{\min}(\tilde{M}_{u_t})}} \right) \\
&\leq \tilde{O} \left(\sqrt{\frac{d}{\tilde{\lambda}_a}} \left(\sqrt{\frac{1}{N_{\mathcal{V}_{\hat{\gamma}}(u)}}} + \frac{\hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u)}}}{\sqrt{N_{\mathcal{V}_{\hat{\gamma}}(u)}}} \right) \right) \\
&\leq \tilde{O} \left(\sqrt{\frac{d}{\tilde{\lambda}_a}} \left(\sqrt{\frac{1}{N_{\mathcal{V}_{\hat{\gamma}}(u)}}} + \hat{\gamma}\sqrt{\eta_{\mathcal{W}_{\hat{\gamma}}(u)}} \right) \right).
\end{aligned}$$

Here the first inequality follows directly from Theorem 1, while the second inequality applies Lemma J.1 in Wang et al. [73] and Lemma 7 in Li and Zhang [39]. This completes the proof of Corollary 1. \blacksquare

B.7 Proof of Theorem 2

Proof. To simplify the notation, we write $u = u_t$. We define

$$\text{SubOpt}_u(\pi_u, \mathbf{x}) \doteq \theta_u^\top (\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))),$$

$$\bar{\beta}_u^n := \frac{2\sqrt{d \log \left(1 + \frac{4\kappa(\tilde{N}_u+n)}{\lambda d}\right)} + 2\log(2N/\delta) + \sqrt{\lambda\kappa}}{\kappa}.$$

First, note that by Lemma 3 and Lemma 6, since the cardinality of the heterogeneous neighbor set $\mathcal{W}_{\hat{Y}}(u)$ remains unchanged during the online phase, we have

$$\left\| \theta_u - \tilde{\theta}_u^n \right\|_{\tilde{M}_u^n} \leq \bar{\beta}_u^n + \frac{\hat{Y}}{2} \sqrt{d N_{\mathcal{W}_{\hat{Y}}(u)}} \quad \text{for each } n \in [N], \quad (37)$$

with probability at least $1 - \frac{\delta}{2N}$. By applying a union bound over all $n \in [N]$, this bound holds uniformly for all rounds with probability at least $1 - \delta$.

We now bound $\text{SubOpt}_u(\pi_u, \mathbf{x})$. It holds that

$$\begin{aligned} & \text{SubOpt}_u(\pi_u, \mathbf{x}) \\ &= \theta_u^\top (\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))) \\ &\stackrel{(a)}{\leq} \theta_u^\top (\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))) + \bar{\theta}_u^\top (\phi(\mathbf{x}, \pi_u(\mathbf{x})) - \phi(\mathbf{x}, \pi_u^*(\mathbf{x}))) \\ &= (\theta_u - \bar{\theta}_u)^\top (\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))) \\ &= \left(\theta_u - \frac{1}{d \lambda_{\min}(\tilde{M}_u^N) + N} \left(d \lambda_{\min}(\tilde{M}_u^N) \tilde{\theta}_u^N + \sum_{n=1}^N \tilde{\theta}_u^n \right) \right)^\top (\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))) \\ &= \frac{1}{d \lambda_{\min}(\tilde{M}_u^N) + N} \left(d \lambda_{\min}(\tilde{M}_u^N) (\theta_u - \tilde{\theta}_u^N)^\top + \sum_{n=1}^N (\theta_u - \tilde{\theta}_u^n)^\top \right) (\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))), \end{aligned} \quad (38)$$

where (a) holds due to the fact that π_u maximizes the pessimistic value (line 9 in Algorithm 2).

Next, for the first term in (39), we have:

$$\begin{aligned} & (\theta_u - \tilde{\theta}_u^N)^\top (\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))) \\ &\stackrel{(a)}{\leq} \|\theta_u - \tilde{\theta}_u^N\|_2 \|\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))\|_2 \\ &\stackrel{(b)}{\leq} 2 \frac{\|\theta_u - \tilde{\theta}_u^N\|_{\tilde{M}_u^N}}{\sqrt{\lambda_{\min}(\tilde{M}_u^N)}} \\ &\stackrel{(c)}{\leq} \frac{2\bar{\beta}_u^N + \hat{Y} \sqrt{d N_{\mathcal{W}_{\hat{Y}}(u)}}}{\sqrt{\lambda_{\min}(\tilde{M}_u^N)}}. \end{aligned} \quad (40)$$

Here, (a) follows from the Cauchy-Schwarz inequality; (b) uses the fact that feature vectors are bounded by 1 in norm and the definition of the minimum eigenvalue; (c) follows from (37).

For the summation term in (39), we have:

$$\sum_{n=1}^N (\theta_u - \tilde{\theta}_u^n)^\top (\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x})))$$

$$\begin{aligned}
&\leq \sum_{n=1}^N \|\theta_u - \tilde{\theta}_u^n\|_{\tilde{M}_u^n} \|\phi(\mathbf{x}, \pi_u^*(\mathbf{x})) - \phi(\mathbf{x}, \pi_u(\mathbf{x}))\|_{(\tilde{M}_u^n)^{-1}} \\
&\stackrel{(a)}{\leq} \sum_{n=1}^N \|\theta_u - \tilde{\theta}_u^n\|_{\tilde{M}_u^n} \|\phi(\dot{\mathbf{x}}_u^n, \dot{\mathbf{a}}_u^n) - \phi(\dot{\mathbf{x}}_u^n, \dot{\mathbf{a}}_u'^n)\|_{(\tilde{M}_u^n)^{-1}} \\
&\stackrel{(b)}{\leq} \sum_{n=1}^N \left(2\bar{\beta}_u^n + \hat{\gamma} \sqrt{d N \mathcal{W}_{\hat{\gamma}}(u)}\right) \|\phi(\dot{\mathbf{x}}_u^n, \dot{\mathbf{a}}_u^n) - \phi(\dot{\mathbf{x}}_u^n, \dot{\mathbf{a}}_u'^n)\|_{(\tilde{M}_u^n)^{-1}} \\
&\stackrel{(c)}{\leq} \left(2\bar{\beta}_u^N + \hat{\gamma} \sqrt{d N \mathcal{W}_{\hat{\gamma}}(u)}\right) \sum_{n=1}^N \|\phi(\dot{\mathbf{x}}_u^n, \dot{\mathbf{a}}_u^n) - \phi(\dot{\mathbf{x}}_u^n, \dot{\mathbf{a}}_u'^n)\|_{(\tilde{M}_u^n)^{-1}} \\
&\stackrel{(d)}{\leq} \left(2\bar{\beta}_u^N + \hat{\gamma} \sqrt{d N \mathcal{W}_{\hat{\gamma}}(u)}\right) \sqrt{N} \sqrt{\sum_{n=1}^N \|\phi(\dot{\mathbf{x}}_u^n, \dot{\mathbf{a}}_u^n) - \phi(\dot{\mathbf{x}}_u^n, \dot{\mathbf{a}}_u'^n)\|_{(\tilde{M}_u^n)^{-1}}^2} \\
&\stackrel{(e)}{\leq} \left(2\bar{\beta}_u^N + \hat{\gamma} \sqrt{d N \mathcal{W}_{\hat{\gamma}}(u)}\right) \sqrt{2dN \log \left(1 + \frac{4\kappa N}{\lambda d}\right)}. \tag{41}
\end{aligned}$$

Here, (a) holds by the active data augmentation rule in line 4 of Algorithm 2; (b) uses the ellipsoid bound (37); (c) holds because $\bar{\beta}_u^n$ is non-decreasing in n ; (d) applies the Cauchy–Schwarz inequality; and (e) follows from the elliptical potential lemma (Lemma 7).

Combining Equation (39), Equation (40), and Equation (41) yields:

$$\begin{aligned}
\text{SubOpt}_u(\pi_u, \mathbf{x}) &\leq \left(\frac{1}{d \lambda_{\min}(\tilde{M}_u^N) + N} \right) \left(2\bar{\beta}_u^N + \hat{\gamma} \sqrt{d N \mathcal{W}_{\hat{\gamma}}(u)} \right) \left(d \sqrt{\lambda_{\min}(\tilde{M}_u^N)} + \sqrt{2dN \log \left(1 + \frac{4\kappa N}{\lambda d}\right)} \right) \\
&\leq \left(\frac{1}{d \lambda_{\min}(\tilde{M}_u^N) + N} \right) \left(2\bar{\beta}_u^N \sqrt{d} + \hat{\gamma} d \sqrt{N \mathcal{W}_{\hat{\gamma}}(u)} \right) \sqrt{2 \left(d \lambda_{\min}(\tilde{M}_u^N) + 2N \log \left(1 + \frac{4\kappa N}{\lambda d}\right) \right)} \\
&= \tilde{O} \left(\frac{d \left(1 + \hat{\gamma} \sqrt{N \mathcal{W}_{\hat{\gamma}}(u)} \right)}{\sqrt{d \lambda_{\min}(\tilde{M}_u^N) + N}} \right).
\end{aligned}$$

Since $\text{SubOpt}_u(\pi_u) = \mathbb{E}_{\mathbf{x} \sim \rho_p} [\text{SubOpt}_u(\pi_u, \mathbf{x})]$, it follows that

$$\text{SubOpt}_u(\pi_u) \leq \tilde{O} \left(\frac{d \left(1 + \hat{\gamma} \sqrt{N \mathcal{W}_{\hat{\gamma}}(u)} \right)}{\sqrt{d \lambda_{\min}(\tilde{M}_u^N) + N}} \right),$$

which completes the proof of Theorem 2. ■

B.8 Proof of Lemma 5

Proof. According to Lemma 10, under the active data augmentation rule in Equation (13), it can be shown that in each block of d^* rounds, the minimum eigenvalue of the Gramian matrix

increases by at least 1, that is, for any $i \in \{1, \dots, \lfloor \frac{N}{d^*} \rfloor\}$,

$$\lambda_{\min}(\tilde{M}_{u_t}^{d^* i}) - \lambda_{\min}(\tilde{M}_{u_t}^{d^* (i-1)}) \geq 1.$$

Therefore, we have:

$$\begin{aligned} \lambda_{\min}(\tilde{M}_{u_t}^N) - \lambda_{\min}(\tilde{M}_{u_t}) &\geq \lambda_{\min}(\tilde{M}_{u_t}^{d^* \lfloor \frac{N}{d^*} \rfloor}) - \lambda_{\min}(\tilde{M}_{u_t}) \\ &\geq \sum_{i=0}^{\lfloor \frac{N}{d^*} \rfloor - 1} \left(\lambda_{\min}(\tilde{M}_{u_t}^{d^* (i+1)}) - \lambda_{\min}(\tilde{M}_{u_t}^{d^* i}) \right) \geq \left\lfloor \frac{N}{d^*} \right\rfloor, \end{aligned}$$

where we define $\lambda_{\min}(\tilde{M}_{u_t}^0) = \lambda_{\min}(\tilde{M}_{u_t})$ to be the minimum eigenvalue of the Gramian matrix constructed from the aggregated offline data. This completes the proof of Lemma 5. ■

C TECHNICAL LEMMAS

Lemma 6 (Confidence Ellipsoid). *Let $\{F_t\}_{t=0}^\infty$ be a filtration. Let $\{\varepsilon_t\}_{t=1}^\infty$ be a real-valued stochastic process such that ε_t is F_t -measurable and ε_t is conditionally R -subgaussian for some $R > 0$. Moreover, let $\{X_t\}_{t=1}^\infty$ be an \mathbb{R}^d -valued stochastic process such that X_t is F_{t-1} -measurable. Assume that $V = \lambda I$ for $\lambda > 0$ is a $d \times d$ positive definite matrix. For any $t \geq 0$, define*

$$\bar{V}_t = V + \sum_{s=1}^n X_s X_s^\top, \quad S_t = \sum_{s=1}^n \varepsilon_s X_s.$$

Let $Y_t = \langle X_t, \theta^ \rangle + \varepsilon_t$ and assume that $\|\theta^*\|_2 \leq S$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$, θ^* lies in the set*

$$C_t = \left\{ \theta \in \mathbb{R}^d : \left\| \hat{\theta}_t - \theta \right\|_{\bar{V}_t} \leq R \sqrt{2 \log \left(\frac{\det(\bar{V}_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right\}$$

where $\hat{\theta}_t = (X_{1:t}^\top X_{1:t} + \lambda I)^{-1} X_{1:t}^\top Y_{1:t}$ is the least squares estimate of θ^ , for $X_{1:t}$ being the matrix whose rows are $X_1^\top, \dots, X_t^\top$ and $Y_{1:t} = (Y_1, \dots, Y_t)^\top$. Furthermore, if for all $t \geq 1$, $\|X_t\|_2 \leq L$ then with probability at least $1 - \delta$, for all $t \geq 0$, θ^* lies in the set*

$$C'_t = \left\{ \theta \in \mathbb{R}^d : \left\| \hat{\theta}_t - \theta \right\|_{\bar{V}_t} \leq R \sqrt{d \log \left(\frac{1 + tL^2/\lambda}{\delta} \right)} + \lambda^{1/2} S \right\}.$$

Proof. Lemma 6 comes from Theorem 2 in Abbasi-Yadkori et al. [1]. ■

Lemma 7 (Elliptic Potential Lemma). *Let $\{z_s\}_{s=1}^n$ be a sequence of vectors in \mathbb{R}^d such that $\|z_s\| \leq L$ for any $s \in [t]$. Let $V_t = \sum_{s=1}^{t-1} z_s z_s^\top + \lambda I$. Then,*

$$\sum_{s=1}^n \|z_s\|_{V_s^{-1}}^2 \leq 2d \log \left(1 + \frac{tL^2}{\lambda d} \right).$$

Proof. Lemma 7 comes from Lemma C.2 in Das et al. [17]. ■

Lemma 8 (Lower Bound on the Minimum Eigenvalue). *Let $\mathbf{a}_s, n \geq 1$ be generated sequentially from a random distribution such that $\|\mathbf{a}\|_2 \leq 1$ and $\mathbb{E}[\mathbf{a}\mathbf{a}^\top]$ is full rank with minimal eigenvalue $\lambda_a > 0$. Let $M_n = \sum_{s=1}^n \mathbf{a}_s \mathbf{a}_s^\top$. Then event*

$$\lambda_{\min}(M_n) \geq \left(n\lambda_a - \frac{1}{3}\sqrt{18nA(\delta) + A(\delta)^2} - \frac{1}{3}A(\delta) \right)$$

holds with probability at least $1 - \delta$ for $n \geq 0$ where $A(n, \delta) = \log \left(\frac{(n+1)(n+3)d}{\delta} \right)$. Furthermore,

$$\lambda_{\min}(M_n) \geq \frac{1}{2}\lambda_a n, \quad \forall n \geq \frac{16}{\lambda_a^2} \log \left(\frac{8d}{\lambda_a^2 \delta} \right)$$

holds with probability at least $1 - \delta$.

Proof. Lemma 8 comes from Lemma 7 in Li and Zhang [39] and Lemma B.2 in Wang et al. [72]. ■

Lemma 9 (One-step Update on the Euclidean Unit Ball). *Let $M \in \mathbb{R}^{d \times d}$ be symmetric positive semidefinite with eigenvalues $\lambda_1(M) \leq \lambda_2(M) \leq \dots \leq \lambda_d(M)$, and corresponding orthonormal eigenvectors q_1, \dots, q_d . Let*

$$z^* := \arg \max_{\|z\|_2 \leq 1} z^\top M^{-1} z, \quad (42)$$

and define the rank-one update $M^+ = M + z^(z^*)^\top$. Then the increase in the smallest eigenvalue satisfies*

$$\lambda_{\min}(M^+) - \lambda_{\min}(M) = \min\{1, \lambda_2(M) - \lambda_1(M)\}.$$

Moreover, the original eigenvector q_1 remains an eigenvector of M^+ , now with eigenvalue

$$M^+ q_1 = (\lambda_1(M) + 1) q_1.$$

Proof. Write the spectral decomposition

$$M = Q \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d) Q^\top,$$

with $Q = [q_1, \dots, q_d]$ where $Q^{-1} = Q^\top$ due to its semi-definite property. For any z with $\|z\| \leq 1$, let $y = Q^\top z$, so $\|y\| \leq 1$ and

$$z^\top M^{-1} z = y^\top \text{diag}(1/\lambda_1, \dots, 1/\lambda_d) y = \sum_{i=1}^d \frac{y_i^2}{\lambda_i}.$$

Since $1/\lambda_1 \geq 1/\lambda_2 \geq \dots$, this quadratic form is maximized by concentrating all mass on the first coordinate:

$$y^* = \pm e_1, \implies z^* = Q y^* = \pm q_1,$$

and without loss of generality $z^* = q_1$. Moreover, because we chose an orthonormal eigenbasis, $\|q_1\| = 1$, so $\|z^*\| = 1$.

Now consider $M^+ = M + q_1 q_1^\top$. Observe:

$$M^+ q_1 = \lambda_1 q_1 + q_1 = (\lambda_1 + 1) q_1, \quad M^+ q_i = \lambda_i q_i \quad (i \geq 2),$$

since $q_i^\top q_1 = 0$. Therefore the eigenvalues of M^+ are $\lambda_1 + 1, \lambda_2, \dots, \lambda_d$, and so

$$\lambda_{\min}(M^+) = \min\{\lambda_1 + 1, \lambda_2\}.$$

Subtracting $\lambda_{\min}(M) = \lambda_1$ gives

$$\lambda_{\min}(M^+) - \lambda_1 = \min\{\lambda_1 + 1, \lambda_2\} - \lambda_1 = \min\{1, \lambda_2 - \lambda_1\},$$

as claimed. ■

Lemma 10 (Multi-step Update on the Euclidean Unit Ball). *Let $M \in \mathbb{R}^{d \times d}$ be symmetric positive semidefinite with eigenvalues*

$$\lambda_1(M) \leq \lambda_2(M) \leq \cdots \leq \lambda_d(M).$$

Suppose that there exists an integer $s \in \{1, 2, \dots, d-1\}$ such that

$$\lambda_{s+1}(M) \geq \lambda_1(M) + 1.$$

Perform s greedy rank-one updates

$$M^{(0)} = M, \quad z_t = \arg \max_{\|z\| \leq 1} z^\top (M^{(t-1)})^{-1} z, \quad M^{(t)} = M^{(t-1)} + z_t z_t^\top, \quad t = 1, \dots, s.$$

Then

$$\lambda_1(M^{(s)}) \geq \lambda_1(M) + 1.$$

Proof. Let k be the largest index such that

$$\lambda_k(M) < \lambda_1(M) + 1,$$

so that $1 \leq k \leq s$, and by definition, $\lambda_{k+1}(M) \geq \lambda_1(M) + 1$. By Lemma 9, each rank-one update increases the eigenvalue of the currently smallest dimension by 1; in particular, the smallest eigenvalue itself increases by 1 if the second-smallest eigenvalue is at least 1 larger. In our case, since $\lambda_{k+1}(M) \geq \lambda_1(M) + 1$, the condition of the lemma is satisfied. Thus, after applying the first k updates (each to a direction aligned with the corresponding eigenvector), we have

$$\lambda_1(M^{(k)}) \geq \lambda_1(M) + 1.$$

For any $i > k$, the original eigenvalue $\lambda_i(M)$ already satisfies $\lambda_i(M) \geq \lambda_{k+1}(M) \geq \lambda_1(M) + 1$, and rank-one updates can only increase or leave unchanged the eigenvalues. Therefore, the remaining $s - k$ updates (if any) cannot decrease $\lambda_1(M^{(k)})$. It follows that

$$\lambda_1(M^{(s)}) \geq \lambda_1(M^{(k)}) \geq \lambda_1(M) + 1,$$

as claimed. ■