Offline Preference Learning with Clustering and Active Data-Augmentation

Jingyuan Liu [1] **Fatemeh Ghaffari** [2] **Xuchuang Wang** [2] **Xutong Liu** [3] **Mohammad Hajiesmaili** [2]
Carlee Joe-Wong [4]

Abstract

Offline preference learning from pairwise feedback is an important problem in applications such as AI alignment and recommendations. Due to the static nature of offline data, most prior methods in this area suffer from poor coverage of the feature (i.e., context-action) distribution induced by the optimal policy for taking actions that a user most prefers. To address the sample restrictions and poor coverage challenges of offline preference learning, this work considers two complementary solutions. First, we exploit data from multiple users within a pure offline setting by learning user similarities. We design Off-C^{2}PL, which aggregates offline data from users with similar preferences to broaden the sample size. Our theoretical results show that this approach improves coverage and reduces policy suboptimality. Second, we consider a hybrid setting in which we can actively collect a small number of samples to augment the offline data. In this setting, we propose ADA-Off-C^{2}PL, which targets the least-covered directions of the offline data to alleviate poor coverage. Theoretical results demonstrate that this approach is particularly effective under highly imbalanced offline data, where the offline data provide good coverage for most feature dimensions but poor coverage for a few. Empirical results on synthetic and real-world datasets show that our methods outperform baselines by at least 57.5%.

1. Introduction

Learning human preferences is a fundamental component of modern AI systems, including aligning large language models (LLMs) with human values (Ouyang et al., 2022; Bai et al., 2022), recommendation systems (Yan et al., 2022; Aramayo et al., 2023), and personalized digital assistants (Musto et al., 2021; Stucke & Ezrachi, 2017). In these applications, systems must infer not only available actions but also which ones users prefer. Unlike supervised learning with explicit labels (Verma et al., 2021; Jiang et al., 2020), preference learning must handle subjective and heterogeneous human choices, and failures in capturing preferences can lead to misaligned outputs or poor user experience. A practical and widely adopted approach to eliciting preferences is *pairwise feedback*, where users indicate which of two options they prefer instead of providing absolute scores. Such comparisons are natural and reliable in practice (e.g., annotators can easily compare two LLM responses). This paradigm has been widely studied under the dueling bandits framework, which leverages sequences of pairwise comparisons to learn preference structures (Yue et al., 2012; Dudík et al., 2015; Saha, 2021; Saha & Krishnamurthy, 2022).

Prior work on preference learning has primarily relied on static offline datasets to learn user preferences (Zhu et al., 2023; Zhan et al., 2023a; Li et al., 2025b). While appealing for their safety and practicality, purely offline methods suffer from fundamental limitations in applications. In particular, these approaches analyze data from different users separately. This restriction significantly limits the effective sample size and often leads to poor coverage of the feature distribution induced by the optimal policy, degrading algorithm performance. Here, a feature refers to a mapping of a context–action pair (e.g., prompt–response pairs in RLHF). Poor coverage means the offline data contain sparse information along certain feature dimensions, with algorithm performance constrained by the least-covered dimension which causes high uncertainty. For example, effective alignment of LLMs can be difficult when feedback reflecting specific regional or cultural preferences is sparse. A natural way to address the poor coverage problem is to use online or active data collection, which can be directed towards collecting more "useful" (i.e., informative) data samples (Xiong et al., 2023; Das et al., 2024). However, fully online methods can be expensive or even infeasible in safety-critical domains such as medical treatments, where collecting online data may incur high costs or ethical concerns. Motivated by these

[1]University of California, Berkeley, California, USA [2]University of Massachusetts Amherst, Massachusetts, USA [3]University of Washington, Washington, USA [4]Carnegie Mellon University, Pennsylvania, USA. Correspondence to: Xutong Liu <xutongl@uw.edu>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

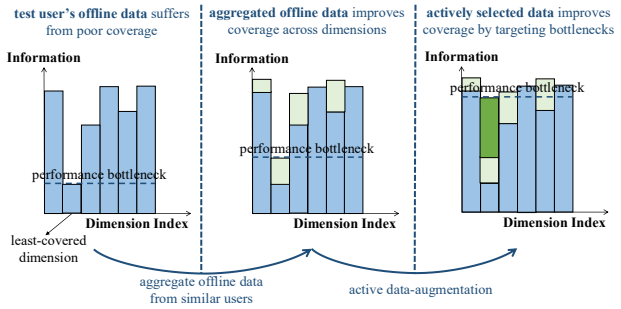


Figure 1. Illustration of how our methods address the coverage problem, where algorithm performance is limited by the least-covered dimension. The x-axis denotes feature dimensions and the y-axis denotes information extent (i.e., eigenvalue) for each feature dimension in the utilized data. Blue bars represent test user's data, light-green bars represent aggregated data from other users, and green bars represent active augmented data.

challenges, this work centers on the following question:

How to leverage the structure of offline data or a small amount of actively selected data to improve feature coverage and algorithm performance in offline preference learning?

To answer this question, we study two complementary settings. In the first, we consider a pure offline setting in which offline data are from multiple users organized into latent clusters, where users within the same cluster share the same preferences while preferences differ across clusters. The goal is to aggregate offline data from users with similar preferences in a principled manner to reduce bias, increase sample size, and improve coverage, thereby enhancing algorithm performance. Unlike prior work on clustering of bandits (Gentile et al., 2014; Wang et al., 2025; Li et al., 2025b) relying on sufficient information and strong coverage assumptions, the data in our setting are fixed and may provide poor coverage, making the cluster structure more difficult to learn. Appendix B introduces the specialization of our results to the classical clustering of bandits setting by imposing the coverage assumptions used in prior work.

In the second setting, we further consider a hybrid model that allows a small number of actively selected data points (e.g., querying for users' feedback in RLHF) on top of the aggregated offline data. Here, the goal is to select data that targets underrepresented directions in the offline dataset of the test user to further improve feature coverage compared to the pure offline setting. Unlike prior pure offline or fully active settings (Zhu et al., 2023; Das et al., 2024; Li et al., 2025a), our hybrid setting requires actively selecting samples to augment the offline data and mitigate poor coverage, making it necessary to select data conditioned on the offline dataset to make use of benefits from both data types. Appendix A provides detailed discussions of related work.

Contributions. (i) To the best of our knowledge, this work is the first to address poor coverage of the optimal policy's feature distribution in offline preference learning from pairwise feedback, using hindsight from clustering and active data-augmentation. The intuition behind how both methods address the coverage problem is illustrated in Figure 1, while comparisons of the theoretical results are in Table 1. For the first pure offline setting, the main challenge is to **identify users with similar preferences from limited and poorly covered data** so as to avoid large bias, and we propose Off-C^{2}PL in Section 3 to address this. The algorithm constructs confidence intervals for estimated preferences based on the minimum eigenvalue of the information matrix for each user u, which characterizes the coverage bottleneck of u's offline dataset. Using these intervals, Off-C^{2}PL identifies users with similar preferences to the test user and aggregates their data to improve the sample size and coverage. Theoretical results show that aggregation improves coverage by supplementing the test user's dataset with information from other users across multiple preference dimensions, while our theoretical upper bounds on algorithm performance are determined by the information in the least-covered dimension. Hence, the performance is substantially improved when the aggregated users' data are rich in the feature dimensions where the test user's data are sparse (e.g., different RLHF datasets emphasize different perspectives on human values).

(ii) For the second hybrid setting, the main challenge is how to **actively select samples based on poorly covered offline data** so that they are informative and can mitigate coverage deficiencies. Accordingly, we design ADA-Off-C^{2}PL in Section 4 to allocate the samples by targeting the least-covered dimensions of the aggregated offline information matrix. We analyze the theoretical benefits of this policy from two perspectives. First, unlike pure offline results that depend on the less transparent eigenvalue of the information matrix, the information gain from actively selected data depends directly on the number of samples, which is more straightforward and pronounced. Second, because the algorithm targets the least-covered dimensions, it is particularly effective when the offline dataset is *highly imbalanced* (i.e., poor coverage in a few dimensions but rich in others). In such cases, the actively selected samples directly correct the imbalance, leading to substantially improved coverage. Notably, each actively selected sample can be as informative as up to d offline samples under such scenarios, where d is the dimension of the preference vectors.

(iii) We evaluate our methods on both synthetic benchmarks and the Reddit TL;DR dataset in Section 5. Empirically, our algorithms outperform strong baselines by at least 57.5%, demonstrating their effectiveness in practical settings.

2. Setting

Notations. In this paper, we use $[s] = \{1, 2, \ldots, s\}$ to denote the set of integers from 1 to s. For any matrix $M \in \mathbb{R}^{d\times d}$, $\lambda_{\min}(M) = \lambda_1(M)$ denotes its smallest eigenvalue,

and $\lambda_i(M)$ denotes its i-th smallest eigenvalue. For vector norms, $\|\cdot\|_2$ denotes the Euclidean norm, and $\|\cdot\|_M$ denotes the Mahalanobis norm induced by matrix M.

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 $\boldsymbol{\theta}_u \in \Theta$, with $\Theta := \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta}\|_2 \leq 1\}$. To model preference heterogeneity, the users are partitioned into J clusters ($J \leq U$). 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 each cluster j share the same preference vector $\boldsymbol{\theta}^j$, [1] i.e., $\boldsymbol{\theta}_u = \boldsymbol{\theta}_{u'} = \boldsymbol{\theta}^j$ 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, each user $u \in \mathcal{U}$ is provided with an offline dataset $\mathcal{D}_u = \{(\boldsymbol{x}_u^i, \boldsymbol{a}_u^i, \boldsymbol{a'}_u^i, y_u^i)\}_{i=1}^{N_u}$ where N_u denotes the number of samples for each user. We define $N_{\mathcal{S}} = \sum_{u \in \mathcal{S}} N_u$ as the total number of samples from users in a set $\mathcal{S}$. Within each dataset $\mathcal{D}_u$, $\boldsymbol{x}_u^i \in \mathcal{X}$ represents a context for selecting actions (e.g., prompts in RLHF or user features in recommendation systems) drawn from the context set $\mathcal{X}$, and $\boldsymbol{a}_u^i, \boldsymbol{a'}_u^i \in \mathcal{A}$ represent a pair of actions (e.g., responses in RLHF or items in recommendation systems) 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 $\boldsymbol{a}_u^i$ over $\boldsymbol{a'}_u^i$ given context $\boldsymbol{x}_u^i$, whereas $y_u^i = 0$ implies the opposite. Preferences y_u^i are assumed to follow the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Debreu, 1960; Zhu et al., 2023):

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

where $r_u(\boldsymbol{x}, \boldsymbol{a}) = \boldsymbol{\theta}_u^\top \phi(\boldsymbol{x}, \boldsymbol{a})$ is a linear reward function parameterized by an unknown vector $\boldsymbol{\theta}_u$ and a known feature mapping $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ with $\|\phi(\boldsymbol{x}, \boldsymbol{a})\|_2 \leq 1$ for all $(\boldsymbol{x}, \boldsymbol{a}) \in \mathcal{X} \times \mathcal{A}$, and $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the sigmoid function. The interpretations of these concepts in applications are discussed in detail in Remark 2.1. Additionally, we define the feature difference $\boldsymbol{z}_u^i = \phi(\boldsymbol{x}_u^i, \boldsymbol{a}_u^i) - \phi(\boldsymbol{x}_u^i, \boldsymbol{a'}_u^i)$. A policy $\pi : \mathcal{X} \to \mathcal{A}$ 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:

$$\begin{aligned}&\mathrm{SubOpt}_{u_t}(\pi_{u_t}) := J_{u_t}(\pi^*_{u_t}) - J_{u_t}(\pi_{u_t}) \\ &= \mathbb{E}_{\boldsymbol{x} \sim \rho_p}\left[\boldsymbol{\theta}_{u_t}^\top \phi(\boldsymbol{x}, \pi^*_{u_t}(\boldsymbol{x})) - \boldsymbol{\theta}_{u_t}^\top \phi(\boldsymbol{x}, \pi_{u_t}(\boldsymbol{x}))\right],\end{aligned} \tag{1}$$

where $J_u(\pi) = \mathbb{E}_{\boldsymbol{x} \sim \rho_p}[r_u(\boldsymbol{x}, \pi(\boldsymbol{x}))]$ denotes the expected reward for user u under policy π, $\pi^*_u = \mathrm{argmax}_\pi J_u(\pi)$ is the optimal policy, and ρ_p denotes the context distribution. 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 fixed offline dataset $\mathcal{D}$, the learner actively selects N additional data points specifically for the test user u_t. The learner chooses a data tuple $(\mathring{\boldsymbol{x}}_{u_t}^n, \mathring{\boldsymbol{a}}_{u_t}^n, \mathring{\boldsymbol{a}}'^n_{u_t}) \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}$ and obtains preference feedback $\mathring{y}_{u_t}^n$ at each selection round $n \in [N]$. We use $\mathring{\mathcal{D}} = \left\{\left(\mathring{\boldsymbol{x}}_{u_t}^n, \mathring{\boldsymbol{a}}_{u_t}^n, \mathring{\boldsymbol{a}}'^n_{u_t}, \mathring{y}_{u_t}^n\right)\right\}_{n=1}^N$ to denote the actively selected dataset after N rounds. The objective is to minimize Equation (1) by leveraging both datasets $\mathcal{D} \cup \mathring{\mathcal{D}}$.

Remark 2.1 (Representative applications). Our framework is closely related to RLHF (Zhu et al., 2023; Das et al., 2024; Li et al., 2025a) and recommendations (Li et al., 2010; Aramayo et al., 2023). In RLHF, $\boldsymbol{x}_u^i$ represents a prompt to labeler u, $(\boldsymbol{a}_u^i, \boldsymbol{a'}_u^i)$ are two candidate responses, and y_u^i indicates the u's preference over them. The reward $r_u(\boldsymbol{x}, \boldsymbol{a})$ reflects the u's underlying evaluation, while $\phi(\boldsymbol{x}_u^i, \boldsymbol{a}_u^i)$ can be interpreted as the output of all but the final layer of a pre-trained language model and $\boldsymbol{\theta}_u$ as the weights in its final layer (Zhu et al., 2023; Park et al., 2024; Li et al., 2025a). In this view, the pure offline setting aims to aggregate offline pairwise 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 selecting prompt–response pairs based on the offline data. For instance, the learner may target prompts where the model's responses are more uncertain, and pair them with contrasting responses, so that the feedback provides additional information for refining the preference estimate. In recommendations, u denotes a user, $\boldsymbol{x}_u^i$ captures contextual feature (e.g., time, category, or interface variant), $(\boldsymbol{a}_u^i, \boldsymbol{a'}_u^i)$ are two candidate items (e.g., movies or products), and y_u^i indicates the preferred one. The pure offline case models cold-start recommendation, estimating u_t's preferences from historical interactions of similar users. The active data-augmented case extends this by querying the user with designed contextual features and item pairs, collecting feedback to estimate preference better.

Remark 2.2 (Poor coverage in the single-user setting). Prior work in offline preference learning primarily relies on the offline dataset of a single user (Zhu et al., 2023; Zhan et al., 2023a; Li et al., 2025b). However, their results suffer from a fundamental *poor coverage* issue. Specifically, when only

[1] Our algorithms remain robust in more general settings where users in a cluster have not exactly same preferences (e.g., people from similar backgrounds have minor differences), as discussed in Appendix C and verified in Section 5. For clarity and consistency with prior work (Gentile et al., 2014; Liu et al., 2025; Wang et al., 2025), we assume same preferences in each cluster in main text.

the offline data of the test user u_t are available, the best result achieved by existing algorithms (denoted by π_{p}) satisfies

$$\begin{aligned}&\mathrm{SubOpt}_{u_t}(\pi_{\mathrm{p}})\\ \le&\tilde{O}\big(\sqrt{d}\big\|\overline{\phi}(\pi^*_{u_t})-\boldsymbol{\omega}\big\|_{M_{u_t}^{-1}}\big)\le\tilde{O}\left(\sqrt{\frac{d}{\lambda_{\min}(M_{u_t})}}\right),\end{aligned}\tag{2}$$

where $\overline{\phi}(\pi)=\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x},\pi(\boldsymbol{x}))]$ denotes the feature expectation of π, $\boldsymbol{\omega}$ is an input reference vector, and $M_{u_t}=\lambda I+\sum_{i=1}^{N_{u_t}}\boldsymbol{z}^i_{u_t}(\boldsymbol{z}^i_{u_t})^\top$ is the regularized information matrix. Equation (2) shows that when the offline data of u_t fail to provide a good coverage of the optimal policy's feature expectation, the term $\big\|\overline{\phi}(\pi^*_{u_t})-\boldsymbol{\omega}\big\|_{M_{u_t}^{-1}}$ becomes large, resulting in high suboptimality. In the worst (most general) case, performance is limited by the minimum eigenvalue of M_{u_t} that captures the least-covered (bottleneck) dimension of data. Since the offline dataset is fixed, resolving this limitation requires utilizing other data. This motivates our approaches that improve coverage by either aggregating offline data from other users, or actively selecting a small amount of data to augment the offline data.

3. Algorithm with Pure Offline Data

In this section, we address the poor coverage challenge in Remark 2.2 using offline data collected across multiple, potentially heterogeneous users. We introduce *Offline Connection-based Clustering of Preference Learning* (Off-C^{2}PL) in Section 3.1, which leverages offline data to identify users with similar preferences and aggregates their data to broaden sample size and improve coverage. We then present theoretical guarantees in Section 3.2, demonstrating how this policy improves coverage under our framework.

3.1. Algorithm Design: Off-C^{2}PL

We present Off-C^{2}PL in Algorithm 1. The key idea is to identify similar users by accounting for both preference similarity and estimation uncertainty under limited and unevenly covered offline data. Since raw sample counts may be misleading when coverage is poor, Off-C^{2}PL constructs confidence intervals using the minimum eigenvalue of each user's information (Gramian) matrix, which reliably captures uncertainty under poor data coverage. The algorithm initializes a null graph and connects users only when their preferences are confidently similar, enabling conservative and safe data aggregation. To handle binary pairwise feedback under a logistic model, Off-C^{2}PL estimates preferences via regularized maximum likelihood estimation (MLE).

Input and Initialization. The inputs (line 1) include test user u_t, dataset $\mathcal{D}$, parameters $(\alpha,\lambda,\delta,\kappa,\hat{\gamma})$ explained later, and a reference vector $\boldsymbol{w}\in\mathbb{R}^d$ used for theoretical simplification (Zhu et al., 2023; Li et al., 2025a). The algorithm

Algorithm 1 Off-C^{2}PL

1: **Input** test user $u_t\in\mathcal{U}$; offline dataset $\mathcal{D}=\bigcup_{u\in\mathcal{U}}\mathcal{D}_u$; parameters $\alpha\ge1$, $\lambda>0$, $\delta>0$, $\kappa>0$, $\hat{\gamma}\ge0$; and reference vector $\boldsymbol{w}$.
2: **Initialize** a null graph $\mathcal{G}=(\mathcal{V},\emptyset)$ where $\mathcal{V}=\mathcal{U}$; for each user $u\in\mathcal{V}$, compute: $\ell^i_u(\boldsymbol{\theta})=y^i_u\log\sigma(\boldsymbol{\theta}^\top\boldsymbol{z}^i_u)+(1-y^i_u)\log\sigma(-\boldsymbol{\theta}^\top\boldsymbol{z}^i_u)$, $M_u=\frac{\lambda}{\kappa}I+\sum_{i=1}^{N_u}\boldsymbol{z}^i_u(\boldsymbol{z}^i_u)^\top$, $\hat{\boldsymbol{\theta}}_u=\mathrm{argmin}_{\boldsymbol{\theta}}\big[-\sum_{i=1}^{N_u}\ell^i_u(\boldsymbol{\theta})+\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\big]$, $\mathrm{CI}_u=\frac{\sqrt{\lambda\kappa}+2\sqrt{d\log\left(1+\frac{4\kappa N_u}{\lambda d}\right)+2\log\left(\frac{2U}{\delta}\right)}}{\kappa\sqrt{\lambda_{\min}(M_u)}}$.
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 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 $u\in\mathcal{V}$, aggregate data and update statistics: $\mathcal{V}_{\hat{\gamma}}(u)=\{v\mid(u,v)\in\mathcal{E}_{\hat{\gamma}}\}\cup\{u\}$, $\tilde{M}_u=\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)}M_v$, $\tilde{N}_u=\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)}N_v$, $\tilde{\boldsymbol{\theta}}_u=\mathrm{argmin}_{\boldsymbol{\theta}}\big[-\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)}\sum_{i=1}^{N_v}\ell^i_v(\boldsymbol{\theta})+\frac{\lambda|\mathcal{V}_{\hat{\gamma}}(u)|}{2}\|\boldsymbol{\theta}\|_2^2\big]$.
10: `// Policy Output`
11: Calculate pessimistic value estimate $\tilde{J}_{u_t}(\pi)$ for policy π as Equation (4), and output $\pi_{u_t}=\arg\max_\pi\tilde{J}_{u_t}(\pi)$.

initializes a null graph $\mathcal{G}$, with each user in $\mathcal{U}$ represented as an isolated node (line 2). It then initializes key statistics: $\ell^i_u(\boldsymbol{\theta})$ denotes the log-likelihood, M_u is the regularized Gramian matrix, $\hat{\boldsymbol{\theta}}_u$ estimates user preferences, and CI_u is a confidence interval based on the minimum eigenvalue of M_u to characterize the coverage of u's offline data.

Offline Cluster Learning. Unlike classical online clustering of bandits algorithms (Gentile et al., 2014; Li & Zhang, 2018; Wang et al., 2025), which start from a complete user graph and iteratively delete edges with online feedback, our method starts with a null graph $\mathcal{G}$ and incrementally connects users whose preferences are sufficiently similar. This connection-based strategy is more robust under offline data, where data sparsity and poor coverage makes edge deletion unreliable and biased (Liu et al., 2025). User similarity is estimated by the threshold $\hat{\gamma}$, which controls the clustering condition. Specifically, u_1 and u_2 are connected (line 5) if

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

where α controls clustering conservativeness. This criterion ensures only users with sufficiently similar preferences are connected (see Section 3.2), building a graph that accurately captures the underlying cluster structure from offline data.

Data Aggregation. Let $\mathcal{G}_{\hat{\gamma}}$ denote the graph obtained after learning the cluster. Based on $\mathcal{G}_{\hat{\gamma}}$, the algorithm aggregates data from users identified to have similar preferences (line 9).

Specifically, we define $\mathcal{V}_{\hat{\gamma}}(u)$ as the set containing u and its one-shot neighbors, indicating users estimated to share similar preferences with u. Accordingly, the algorithm constructs the aggregated Gramian matrix $\tilde{M}_u$ by combining samples from users in $\mathcal{V}_{\hat{\gamma}}(u)$. The preference estimate of u is then refined by applying MLE, yielding $\tilde{\boldsymbol{\theta}}_u$.

Policy Output. Finally, the algorithm computes a pessimistic value estimate (Jin et al., 2021; Rashidinejad et al., 2021; Li et al., 2022) for u_t and policy π that down-weights underrepresented dimensions to avoid overestimation:

$$\begin{aligned}\tilde{J}_{u_t}(\pi) =&(\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x},\pi(\boldsymbol{x}))] - \boldsymbol{w})^\top \tilde{\boldsymbol{\theta}}_{u_t} \\ &- \tilde{\beta}_{u_t}\|\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x},\pi(\boldsymbol{x}))] - \boldsymbol{w}\|_{\tilde{M}_{u_t}^{-1}},\end{aligned} \tag{4}$$

where $\tilde{\beta}_u = \big(2\sqrt{d\log\big(1+\frac{4\tilde{N}_u\kappa}{\lambda|\mathcal{V}_{\hat{\gamma}}(u)|d}\big)+2\log\big(\frac{2U}{\delta}\big)} + \sqrt{\lambda|\mathcal{V}_{\hat{\gamma}}(u)|\kappa}\big)/\kappa + \big(\hat{\gamma}\sqrt{dN_{\mathcal{W}_{\hat{\gamma}}(u)}}\big)/2$ is a confidence term capturing estimation uncertainty (line 11). The algorithm outputs policy π_{u_t} that maximizes $\tilde{J}_{u_t}(\pi)$.

3.2. Theoretical Results for Algorithm 1

We present the theoretical results for Algorithm 1 with proofs in Appendix E. Note the estimator $\tilde{\boldsymbol{\theta}}_u$ is obtained by aggregating data from users in the neighborhood $\mathcal{V}_{\hat{\gamma}}(u)$, which may include homogeneous and heterogeneous ones. We formally distinguish two types of neighbors as follows:

$$\begin{aligned}\mathcal{R}_{\hat{\gamma}}(u) &:= \{v \mid v\in\mathcal{V}_{\hat{\gamma}}(u),\ \boldsymbol{\theta}_v = \boldsymbol{\theta}_u\},\\ \mathcal{W}_{\hat{\gamma}}(u) &:= \{v \mid v\in\mathcal{V}_{\hat{\gamma}}(u),\ \boldsymbol{\theta}_v \neq \boldsymbol{\theta}_u\}.\end{aligned} \tag{5}$$

Here, set $\mathcal{R}_{\hat{\gamma}}(u)$ contains u and its *homogeneous neighbors* that share the same preference vector, whose data can be safely aggregated without bias. In contrast, $\mathcal{W}_{\hat{\gamma}}(u)$ denotes the set of *heterogeneous neighbors* with different preference vectors, whose data aggregation may induce bias and must be controlled. Therefore, characterizing the cardinalities of both sets is crucial for analysis. This is shown in Lemma 3.2.

Definition 3.1 (Heterogeneity gap)**.** For any two users u and v from different clusters (i.e. $j_u \neq j_v$), there exists γ s.t. the gap between their preferences satisfy $\|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2 \geq \gamma$.

Lemma 3.2 (Cardinality of $\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$)**.** *For any user u, let inputs in Algorithm 1 satisfy $\alpha \geq 1$, $\kappa = 1/(2+e^2+e^{-2})$, λ and δ satisfy $\lambda \leq 2\log(2U/\delta) + d\log\big(1+\frac{4\kappa\min_v\{N_v\}}{d\lambda}\big)$, $\delta \leq \frac{d\lambda}{4\kappa\min_v\{N_v\}+d\lambda}$. Then there exist some $\alpha_r \in \big(\frac{\kappa}{3(\alpha+1)\sqrt{2\max\{2,d\}\log(2U/\delta)}}, \frac{\kappa}{2(\alpha-1)\sqrt{2\log(2U/\delta)}}\big)$, $\alpha_w \in \big(0, \frac{\kappa}{2(\alpha-1)\sqrt{2\log(2U/\delta)}}\big)$ such that $\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$ can be characterized with probability at least $1-\delta$ as:*

$$\begin{aligned}\mathcal{R}_{\hat{\gamma}}(u) =& \{u\} \cup \Big\{v \,\Big|\, \boldsymbol{\theta}_u = \boldsymbol{\theta}_v \text{ and} \\ &\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha_r\hat{\gamma}\Big\},\end{aligned} \tag{6}$$

$$\begin{aligned}\mathcal{W}_{\hat{\gamma}}(u) =& \Big\{v \,\Big|\, \gamma \leq \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2 < \hat{\gamma} \text{ and} \\ &\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha_w(\hat{\gamma}-\gamma)\Big\}.\end{aligned} \tag{7}$$

In Lemma 3.2, $\lambda_{\min}(M_u)$ captures the information along the least-covered dimension of the information matrix M_u. The first condition in Equation (6) guarantees the homogeneity of users in $\mathcal{R}_{\hat{\gamma}}(u)$, while the second condition shows the least-covered dimension as the bottleneck for connecting homogeneous users. Notably, the right-hand side of the second condition scales with $\hat{\gamma}$, implying that increasing the clustering threshold $\hat{\gamma}$ relaxes the coverage requirement and includes more homogeneous neighbors. In contrast, $\mathcal{W}_{\hat{\gamma}}(u)$ represents heterogeneous neighbors that may introduce bias. The first condition in Equation (7) ensures that only users with preference differences below $\hat{\gamma}$ can be connected, while the second condition mirrors that in Equation (6), indicating that $\hat{\gamma}$ also controls the number of heterogeneous users included. Leveraging Lemma 3.2, we establish an upper bound on the suboptimality of Off-C^{2}PL in Theorem 3.3.

Theorem 3.3. *Under the same conditions as in Lemma 3.2, the suboptimality of Algorithm 1 for u_t can be bounded with probability at least $1-\delta$ as:*

$$\begin{aligned}&\text{SubOpt}_{u_t}(\pi_{u_t})\\ &\leq \tilde{O}\big(\sqrt{d}\big(1+\hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}\big)\big\|\overline{\phi}(\pi^*_{u_t}) - \boldsymbol{w}\big\|_{\tilde{M}_{u_t}^{-1}}\big) \qquad (8)\\ &\leq \tilde{O}\left(\frac{\sqrt{d}\big(1+\hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}\big)}{\sqrt{\lambda_{\min}(\tilde{M}_{u_t})}}\right), \qquad (9)\end{aligned}$$

where $\overline{\phi}(\pi) = \mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x},\pi(\boldsymbol{x}))]$ denotes the feature expectation of policy π, and $\tilde{M}_{u_t} = \sum_{v\in\mathcal{V}_{\hat{\gamma}}(u_t)} M_v$ denotes the information matrix constructed from the aggregated data of 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 3.2, i.e., $\mathcal{W}_{\hat{\gamma}}(u_t) = \emptyset$, and the bound simplifies to

$$\begin{aligned}&\text{SubOpt}_{u_t}(\pi_{u_t})\\ &\leq \tilde{O}\big(\sqrt{d}\big\|\overline{\phi}(\pi^*_{u_t}) - \boldsymbol{w}\big\|_{\tilde{M}_{u_t}^{-1}}\big) \leq \tilde{O}\left(\sqrt{\frac{d}{\lambda_{\min}(\tilde{M}_{u_t})}}\right).\end{aligned} \tag{10}$$

The suboptimality gap in Equation (8) is the product of two interpretable terms. The first term, $\sqrt{d}\big(1+\hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}\big)$, decomposes into a statistical sampling noise component $\sqrt{d}$ (up to logarithmic factors), which grows with the preference dimension, and a bias component that increases with $\hat{\gamma}$ and the number of samples from heterogeneous neighbors $N_{\mathcal{W}_{\hat{\gamma}}(u_t)}$. The second term in Equation (8), $\big\|\overline{\phi}(\pi^*_{u_t}) - \boldsymbol{w}\big\|_{\tilde{M}_{u_t}^{-1}}$, is the *concentratability coefficient*,

a standard concept in offline learning and policy evaluation (Jin et al., 2021; Zhu et al., 2023; Li et al., 2025a). It quantifies the mismatch between the optimal-policy feature distribution and that supported by the aggregated offline data from $\mathcal{V}_{\hat{\gamma}}(u_t)$; smaller values indicate better coverage. Choosing the reference vector $\boldsymbol{w}$ as a representative feature (e.g., the most frequent feature vector in data) aligns this term with the data-supported subspace and yields a tighter bound (as discussed in Remark 3.5 of Zhu et al. (2023)).

Remark 3.4 (How offline clustering improves coverage). As discussed in Remark 2.2, prior work on single-user offline preference learning derives suboptimality with concentratability coefficients depending solely on u_t's information matrix M_{u_t}. Since this matrix is constructed from limited single-user data, it exhibits poorer coverage than the aggregated information matrix $\tilde{M}_{u_t}$ in our approach, which enriches information across dimensions. In the worst case, their bound scales as $\tilde{O}\big(\sqrt{d/\lambda_{\min}(M_{u_t})}\big)$ (corresponding to setting $\hat{\gamma}=0$ in Off-C^{2}PL, where only u_t's offline data are used). By contrast, Theorem 3.3 features a larger denominator $\sqrt{\lambda_{\min}(\tilde{M}_{u_t})}$ by aggregating offline data from neighbors in $\mathcal{V}_{\hat{\gamma}}(u_t)$ instead of using only u_t's data. This aggregation directly mitigates the poor coverage inherent in single-user data. Concretely, by Weyl's inequality,

$$\lambda_{\min}(\tilde{M}_{u_t}) - \lambda_{\min}(M_{u_t}) \in \Big[\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u_t)\setminus\{u_t\}} \lambda_{\min}(M_v), \sum_{v\in\mathcal{V}_{\hat{\gamma}}(u_t)\setminus\{u_t\}} \lambda_{\max}(M_v)\Big]. \quad (11)$$

This implies that aggregating neighbors' data increases the minimum eigenvalue of the information matrix by at least the sum of neighbors' minimum eigenvalues. In the most favorable scenario, the increase can be as large as the sum of neighbors' maximum eigenvalues. Such offline data aggregation is particularly effective when neighbors possess rich information along u_t's less-covered dimensions, thereby improving coverage situation in the information matrix. The tradeoff is that aggregating data from heterogeneous neighbors may introduce an additional bias term in Theorem 3.3. This effect can be controlled by choosing a $\hat{\gamma}$ smaller than γ (defined in Definition 3.1) when a positive lower bound on γ is available as shown in Equation (10), or by adopting conservative, data-driven selection strategies as analyzed in prior work (Liu et al., 2025). We defer additional discussions on parameter selection and properties to Appendix D.

4. Algorithm with Active Data-Augmentation

In Section 3, we aggregate offline data from multiple users to improve the coverage. However, because offline data are fixed and static, it is still possible that all users exhibit poor coverage along certain feature dimensions. In such cases, the bound in Equation (11) may achieve its lower limit, and the aggregated information matrix $\tilde{M}_{u_t}$ can remain poorly covered along some dimensions. This limitation is difficult to fully resolve with purely offline data. However, in many applications it is feasible to actively collect a limited amount of data to complement offline data. By targeting the least-covered dimensions, such samples can more effectively mitigate coverage issues. Motivated by this, we extend the offline framework to the active data-augmented model as Section 2 and propose the *Active Data-Augmented Offline Connection-based Clustering of Preference Learning* (ADA-Off-C^{2}PL) in Section 4.1, with analysis in Section 4.2.

Algorithm 2 ADA-Off-C^{2}PL

1: **Input** test user $u_t \in \mathcal{U}$; offline dataset $\mathcal{D} = \bigcup_{u\in\mathcal{U}} \mathcal{D}_u$; online rounds N; graph $\mathcal{G}_{\hat{\gamma}}$; neighbor set $\mathcal{V}_{\hat{\gamma}}(u_t)$; aggregated Gramian matrix $\tilde{M}_{u_t}$; and initial preference estimate $\tilde{\boldsymbol{\theta}}_{u_t}$ from Algorithm 1.
2: **Initialize** $\tilde{M}^1_{u_t} \leftarrow \tilde{M}_{u_t}$ and $\tilde{\boldsymbol{\theta}}^1_{u_t} \leftarrow \tilde{\boldsymbol{\theta}}_{u_t}$.
3: `// Active-data Augmentation`
4: **for** $n = 1, \dots, N$ **do**
5: Select $\big(\mathring{\boldsymbol{x}}^n_{u_t}, \mathring{\boldsymbol{a}}^n_{u_t}, \mathring{\boldsymbol{a}}'^n_{u_t}\big)$ according to Equation (12).
6: Receive feedback $\mathring{y}^n_{u_t}$, compute $\mathring{\boldsymbol{z}}^n_{u_t} = \phi(\mathring{\boldsymbol{x}}^n_{u_t}, \mathring{\boldsymbol{a}}^n_{u_t}) - \phi(\mathring{\boldsymbol{x}}^n_{u_t}, \mathring{\boldsymbol{a}}'^n_{u_t})$ and $\mathring{\ell}^n_{u_t}(\boldsymbol{\theta}) = \mathring{y}^n_{u_t} \log\sigma\big(\boldsymbol{\theta}^\top \mathring{\boldsymbol{z}}^n_{u_t}\big) + (1-\mathring{y}^n_{u_t}) \log\sigma\big(-\boldsymbol{\theta}^\top \mathring{\boldsymbol{z}}^n_{u_t}\big)$. Then update $\tilde{M}^{n+1}_{u_t} = \tilde{M}^n_{u_t} + \mathring{\boldsymbol{z}}^n_{u_t}\big(\mathring{\boldsymbol{z}}^n_{u_t}\big)^\top$ and $\tilde{\boldsymbol{\theta}}^{n+1}_{u_t}$ as in Equation (13).
7: **end for**
8: `// Policy Output`
9: Construct $\overline{\boldsymbol{\theta}}_{u_t}$ as Equation (14).
10: **Output:** $\pi_{u_t}(\boldsymbol{x}) = \operatorname{argmax}_{\boldsymbol{a}\in\mathcal{A}} \phi(\boldsymbol{x}, \boldsymbol{a})^\top \overline{\boldsymbol{\theta}}_{u_t}$.

4.1. Algorithm Design: ADA-Off-C^{2}PL

Recall from Section 2 that under the active data-augmented model, the learner is allowed to actively select N informative samples for the test user u_t to complement the offline data by improving coverage of the feature space, thereby mitigating the poor coverage of the offline datasets (e.g., via additional dialogue rounds in conversational recommendation systems). Since the clustering structure is learned offline, the active data-augmentation is based on the aggregated Gramian matrix $\tilde{M}_{u_t}$, which summarizes information from the neighborhood of u_t. As shown in Remark 2.2 and Theorem 3.3, the suboptimality is governed by the minimum eigenvalue of $\tilde{M}_{u_t}$. Therefore, the goal of this phase is to actively collect new data that increase the minimum eigenvalue, ensuring sufficient coverage across all dimensions. The detailed procedure is presented in Algorithm 2.

Input and Initialization. The inputs and initialization use the results from Off-C^{2}PL. Specifically, the algorithm takes test user u_t, dataset $\mathcal{D}$, learned user graph $\mathcal{G}_{\hat{\gamma}}$, neighbor set $\mathcal{V}_{\hat{\gamma}}(u_t)$, Gramian matrix $\tilde{M}_{u_t}$ and preference estimate $\tilde{\boldsymbol{\theta}}_{u_t}$ as inputs (line 1) and initializes core parameters in line 2.

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

$$(\mathring{\boldsymbol{x}}_{u_t}^n, \mathring{\boldsymbol{a}}_{u_t}^n, \mathring{\boldsymbol{a}}'^n_{u_t}) = \underset{(\boldsymbol{x},\boldsymbol{a},\boldsymbol{a}')\in\mathcal{X}\times\mathcal{A}\times\mathcal{A}}{\operatorname{argmax}} \left\{ \left\| \phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}') \right\|_{(\tilde{M}_{u_t}^{n-1})^{-1}} \right\}. \tag{12}$$

After observing the feedback $\mathring{y}_{u_t}^n$, the feature difference $\mathring{\boldsymbol{z}}_{u_t}^n$ and log-likelihood $\mathring{\ell}_{u_t}^n(\boldsymbol{\theta})$ are computed. Then the Gramian matrix is updated and the preference estimate is refined by solving a MLE problem combining the aggregated offline data and actively selected data up to round n:

$$\tilde{\boldsymbol{\theta}}_{u_t}^{n+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \Bigg[-\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u_t)} \sum_{i=1}^{N_v} \ell_v^i(\boldsymbol{\theta}) - \sum_{i=1}^{n} \mathring{\ell}_{u_t}^i(\boldsymbol{\theta}) + \frac{\lambda|\mathcal{V}_{\hat{\gamma}}(u_t)|}{2} \|\boldsymbol{\theta}\|_2^2 \Bigg]. \tag{13}$$

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

$$\overline{\boldsymbol{\theta}}_{u_t} = \frac{d\lambda_{\min}(\tilde{M}_{u_t}^N)\tilde{\boldsymbol{\theta}}_{u_t}^N + \sum_{n=1}^N \tilde{\boldsymbol{\theta}}_{u_t}^n}{d\lambda_{\min}(\tilde{M}_{u_t}^N) + N}. \tag{14}$$

This weighting places more emphasis on the final estimate, extending prior approach in Das et al. (2024) which only uses a simple average for the pure active setting. The policy then selects the action that maximizes the expected reward.

4.2. Theoretical Results for Algorithm 2

We analyze the theoretical guarantees for ADA-Off-C^{2}PL and its effectiveness in improving the data coverage. Firstly, Theorem 4.1 gives its theoretical upper bound.

Theorem 4.1. *Under the same assumptions as in Lemma 3.2 and Theorem 3.3, the suboptimality of Algorithm 2 for u_t can be bounded with probability at least $1-\delta$ as:*

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

where $\tilde{M}_{u_t}^N = \frac{\lambda}{\kappa}|\mathcal{V}_{\hat{\gamma}}(u_t)|I + \sum_{v\in\mathcal{V}_{\hat{\gamma}}(u_t)}\sum_{i=1}^{N_v} \boldsymbol{z}_v^i(\boldsymbol{z}_v^i)^\top + \sum_{i=1}^N \mathring{\boldsymbol{z}}_u^i(\mathring{\boldsymbol{z}}_u^i)^\top$ denotes the final Gramian matrix combining the aggregated offline data and the actively collected data, we will show its relationship with $\tilde{M}_{u_t}$ and N in Lemma 4.3.

In Theorem 4.1, the numerator has the same form as in Theorem 3.3, capturing both sample noise and the bias from heterogeneous neighbors. The key difference lies in the denominator: the quantity inside the square root is increased by $N/d + \big(\lambda_{\min}(\tilde{M}_{u_t}^N) - \lambda_{\min}(\tilde{M}_{u_t})\big)$ compared to Theorem 3.3. First, the term N/d captures the information gain from the actively selected samples, yielding a direct improvement that depends on the sample count N rather than indirectly on eigenvalues as in pure offline settings. In the special case of a single user with no offline data ($\hat{\gamma} = 0$ and $\mathcal{D} = \emptyset$), Theorem 4.1 recovers the result of pure active setting in Das et al. (2024). Second, the term $\lambda_{\min}(\tilde{M}_{u_t}^N) - \lambda_{\min}(\tilde{M}_{u_t})$ captures the additional benefit of augmenting offline data with actively selected data. Since the minimum eigenvalue determines the learning bottleneck in the pure offline setting, the selected samples can directly alleviate this limitation by increasing $\lambda_{\min}(\tilde{M}_{u_t})$. To formalize this effect, we introduce Definition 4.2 and Lemma 4.3 to show the improvement.

Definition 4.2 ((d^*, N)-dimension imbalanced Gramian matrix)**.** A Gramian matrix M is called (d^*, N)-dimension imbalanced if d^* is the smallest value in $\{1, \cdots, d\}$ such that $\lambda_{d^*+1}(M) - \lambda_{\min}(M) \ge \lceil N/d^* \rceil$. By convention, any matrix is at least (d, N)-dimension imbalanced, since there are only d dimensions and we treat $\lambda_{d+1}(M)$ as $+\infty$.

Lemma 4.3 (Minimum eigenvalue improvement)**.** *Assume that the feature difference vector $\boldsymbol{z} = \phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')$ can span the entire Euclidean unit ball $\{\boldsymbol{z} \in \mathbb{R}^d : \|\boldsymbol{z}\|_2 \le 1\}$ for all $(\boldsymbol{x},\boldsymbol{a},\boldsymbol{a}') \in \mathcal{X}\times\mathcal{A}\times\mathcal{A}$, and $\tilde{M}_{u_t}$ is (d^*, N)-dimension imbalanced. Then, under Algorithm 2 with N rounds, it holds that $\lambda_{\min}(\tilde{M}_{u_t}^N) - \lambda_{\min}(\tilde{M}_{u_t}) \ge \lfloor N/d^* \rfloor$.*

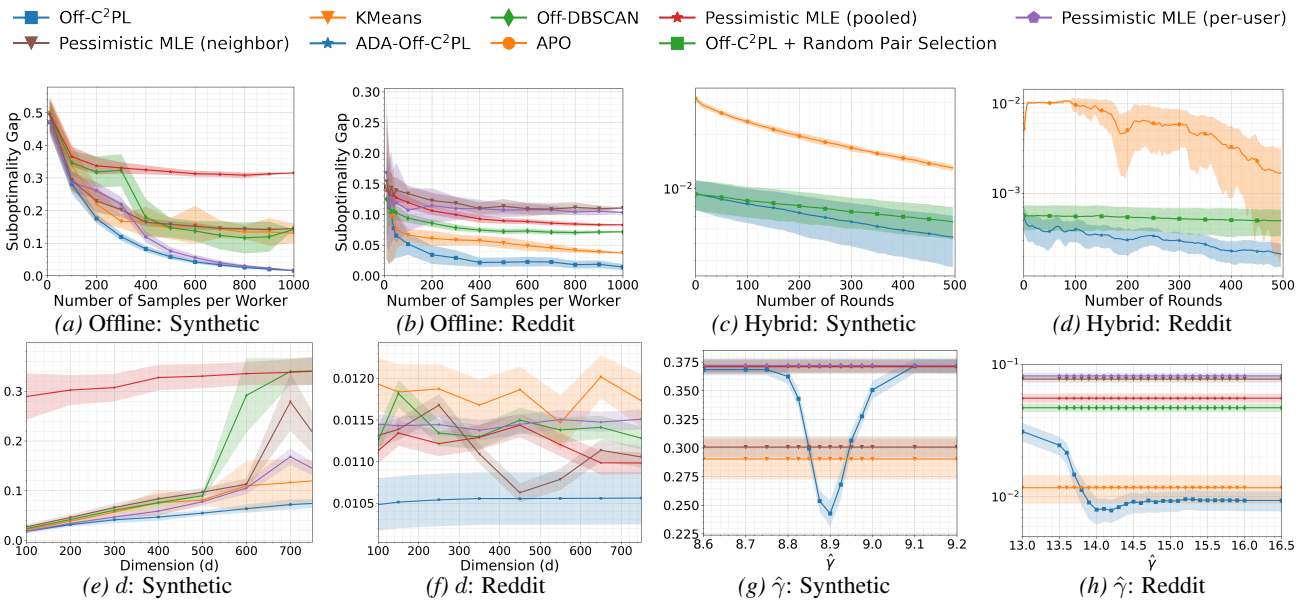
Definition 4.2 indicates a large gap in sample sufficiency between the least-covered dimension and the (d^*+1)-th dimension under N samples. Lemma 4.3 further shows that Algorithm 2 can effectively increase the minimum eigenvalue by selectively targeting these d^* underrepresented dimensions. By combining Theorem 4.1 with Lemma 4.3, we obtain Corollary 4.4, which characterizes suboptimality improvements under dimension-imbalanced data.

Corollary 4.4. *With same assumptions as Lemma 4.3, Theorem 4.1 can be rewritten as:*

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

For $\hat{\gamma} \le \gamma$, it can be simplified to $\tilde{O}\left(\sqrt{\frac{d}{\lambda_{\min}(\tilde{M}_{u_t}) + N/d^*}}\right)$.

Compared to Theorem 4.1, the square-root term in the denominator depends directly on the minimum eigenvalue of the pure offline matrix $\tilde{M}_{u_t}$ instead of the hybrid information matrix $\tilde{M}_{u_t}^N$, and further improves the N/d term to N/d^*. This refinement more clearly quantifies the benefit of active data-augmentation under ADA-Off-C^{2}PL by introducing an additional N/d^* term in the denominator's square root compared to the pure offline result in Theorem 3.3. Intuitively, the actively selected samples only need



(a) Offline: Synthetic *(b)* Offline: Reddit *(c)* Hybrid: Synthetic *(d)* Hybrid: Reddit

(e) d: Synthetic *(f)* d: Reddit *(g)* $\hat{\gamma}$: Synthetic *(h)* $\hat{\gamma}$: Reddit

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

to be allocated across d^* dimensions, rather than all d dimensions (with $d^* \leq d$). Hence, for a (d^*, N)-dimension imbalanced matrix $\tilde{M}_{u_t}$, a single actively selected sample can be as effective as d/d^* samples, leading to a strictly improved suboptimality bound, which is also illustrated by Figure 1. Overall, this result highlights how active data-augmentation can effectively improve the coverage situation of data by reinforcing the less-covered feature dimensions.

5. Experiments

In this section, we evaluate the performance of Off-C^{2}PL and ADA-Off-C^{2}PL using real-world and synthetic data. All experiments are averaged over 20 independent rounds. More details on baselines and datasets are in Appendix G.

Experiment 1: Performance under pure offline model. We compare Off-C^{2}PL to several baselines. On synthetic data (Figure 2a), Off-C^{2}PL attains the smallest suboptimality throughout, improving by 88.1% over KMeans, 89.1% over Off-DBSCAN, and 95.1%, 89.2%, 3.39% over Pessimistic MLE (pooled), (neighbor), and (per-user). The per-user MLE becomes competitive only after $\gtrsim 80\%$ of the samples and remains clearly worse in low-sample regimes. On Reddit (Figure 2b), no baseline matches Off-C^{2}PL: with ≈ 400 pairs per user it reaches near-zero suboptimality, improving by at least 61.5% over the rest baselines.

Experiment 2: Performance under active data-augmented model. We compare ADA-Off-C^{2}PL to APO and to an Off-C^{2}PL-initialized variant that replaces our active augmentation with *random* pair selection. We use 20% of the data offline, then run 500 active rounds. On Reddit, ADA-Off-C^{2}PL improves by 87.6% over APO and 57.5% over random selection; on synthetic, the gains are 58.7% and 18.0%. Figures 2c and 2d show why: APO starts with a large gap (no offline warm start) and remains worse even after active rounds, while random selection shares ADA-Off-C^{2}PL's initial gap but makes little progress. In contrast, ADA-Off-C^{2}PL steadily reduces the gap throughout the active phase and achieves the best performance.

Experiment 3: The impact of dimension d. We vary d from 100 to 800 on synthetic data and from 100 to 768 on Reddit; for Reddit we use PCA to obtain lower-dimensional features. On synthetic (Figure 2e), the suboptimality increases with d, reflecting higher estimation complexity, while Off-C^{2}PL degrades the slowest. On Reddit (Figure 2f), performance shows no clear dependence on d, consistent with PCA retaining the dominant variance directions and discarding low-variance components that contribute little to performance.

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-C^{2}PL recovers the correct cluster structure and substantially reduces the suboptimality gap, demonstrating that accurate control of cluster connectivity is crucial when data is scarce.

6. Conclusion

This paper addresses the poor coverage problem in offline preference learning, where fixed and limited data may inadequately cover different feature dimensions. We propose

two complementary approaches. The first aggregates offline data across users with similar preferences via clustering, improving data sufficiency and coverage across dimensions, with performance driven by gains in the least-covered dimension. The second augments offline data with a small amount of actively selected data, directly addressing coverage bottlenecks and further improving performance. Future work includes extending beyond the BTL feedback model and considering nonstationary preference dynamics.

Acknowledgements

This work was supported in part by the National Science Foundation under Grant No. 2533813.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Aramayo, N., Schiappacasse, M., and Goic, M. A multi-armed bandit approach for house ads recommendations. *Marketing Science*, 42(2):271–292, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

Bengs, V., Busa-Fekete, R., El Mesaoudi-Paul, A., and Hüllermeier, E. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Bu, J., Simchi-Levi, D., and Xu, Y. Online pricing with offline data: Phase transition and inverse square law. In *International Conference on Machine Learning*, pp. 1202–1210. PMLR, 2020.

Cai, T. T. and Guo, Z. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. 2017.

Chen, X., Wang, Y., and Zhou, Y. Dynamic assortment optimization with changing contextual information. *Journal of machine learning research*, 21(216):1–44, 2020.

Chen, X., Zhong, H., Yang, Z., Wang, Z., and Wang, L. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pp. 3773–3793. PMLR, 2022.

Cheung, W. C. and Lyu, L. Leveraging (biased) information: Multi-armed bandits with offline data. *arXiv preprint arXiv:2405.02594*, 2024.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

Dai, X., Wang, Z., Xie, J., Liu, X., and Lui, J. C. Conversational recommendation with online learning and clustering on misspecified users. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):7825–7838, 2024.

Das, N., Chakraborty, S., Pacchiano, A., and Chowdhury, S. R. Active preference optimization for sample efficient rlhf. *arXiv preprint arXiv:2402.10500*, 2024.

Debreu, G. Individual choice behavior: A theoretical analysis, 1960.

Duan, Y. and Wang, K. Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5):2015–2039, 2023.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

Gentile, C., Li, S., and Zappella, G. Online clustering of bandits. In *International conference on machine learning*, pp. 757–765. PMLR, 2014.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. On context-dependent clustering of bandits. In *International Conference on machine learning*, pp. 1253–1262. PMLR, 2017.

Gui, L., Gârbacea, C., and Veitch, V. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *arXiv preprint arXiv:2406.00832*, 2024.

Jang, J., Kim, S., Lin, B. Y., Wang, Y., Hessel, J., Zettlemoyer, L., Hajishirzi, H., Choi, Y., and Ammanabrolu, P. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Jiang, T., Gradus, J. L., and Rosellini, A. J. Supervised machine learning: a brief primer. *Behavior therapy*, 51 (5):675–687, 2020.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kim, D., Lee, K., Shin, J., and Kim, J. Spread preference annotation: Direct preference judgment for efficient llm alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. 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.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pp. 45–73. Springer, 2012.

Lee, J. and Oh, M.-h. Nearly minimax optimal regret for multinomial logistic bandit. *Advances in Neural Information Processing Systems*, 37:109003–109065, 2024.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, G., Ma, C., and Srebro, N. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *Advances in Neural Information Processing Systems*, 35:20974–20987, 2022.

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

Li, L.-F., Qian, Y.-Y., Zhao, P., and Zhou, Z.-H. Provably efficient rlhf pipeline: A unified view from contextual bandits. *arXiv preprint arXiv:2502.07193*, 2025a.

Li, S. and Zhang, S. Online clustering of contextual cascading bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Li, S., Chen, W., and Leung, K.-S. Improved algorithm on online clustering of bandits. *arXiv preprint arXiv:1902.09162*, 2019.

Li, X., Zhou, R., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024.

Li, Z., Yang, Z., and Wang, M. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*, 2023.

Li, Z., Liu, M., Dai, X., and Lui, J. Demystifying online clustering of bandits: Enhanced exploration under stochastic and smoothed adversarial contexts. *arXiv preprint arXiv:2501.00891*, 2025b.

Liu, J., Zhang, Z., Wang, X., Liu, X., Lui, J., Hajiesmaili, M., and Joe-Wong, C. Offline clustering of linear bandits: Unlocking the power of clusters in data-limited environments. *arXiv preprint arXiv:2505.19043*, 2025.

Liu, P., Shi, C., and Sun, W. W. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.

Liu, X., Zhao, H., Yu, T., Li, S., and Lui, J. C. Federated online clustering of bandits. In *Uncertainty in Artificial Intelligence*, pp. 1221–1231. PMLR, 2022.

McQueen, J. B. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pp. 281–297, 1967.

Musto, C., Narducci, F., Polignano, M., De Gemmis, M., Lops, P., and Semeraro, G. Myrrorbot: A digital assistant based on holistic user models for personalized access to online services. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–34, 2021.

Oh, M.-h. and Iyengar, G. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Park, C., Liu, M., Kong, D., Zhang, K., and Ozdaglar, A. Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Poddar, S., Wan, Y., Ivison, H., Gupta, A., and Jaques, N. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

Ramesh, S. S., Hu, Y., Chaimalas, I., Mehta, V., Sessa, P. G., Bou Ammar, H., and Bogunovic, I. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137, 2024.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Rosset, C., Cheng, C.-A., Mitra, A., Santacroce, M., Awadallah, A., and Xie, T. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Saha, A. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021.

Saha, A. and Krishnamurthy, A. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pp. 968–994. PMLR, 2022.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

Shivaswamy, P. and Joachims, T. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pp. 1046–1054. PMLR, 2012.

Singla, A., Rafferty, A. N., Radanovic, G., and Heffernan, N. T. Reinforcement learning for education: Opportunities and challenges. *arXiv preprint arXiv:2107.08828*, 2021.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Stucke, M. E. and Ezrachi, A. How digital assistants can harm our economy, privacy, and democracy. *Berkeley Technology Law Journal*, 32(3):1239–1300, 2017.

Tiapkin, D., Belomestny, D., Calandriello, D., Moulines, E., Naumov, A., Perrault, P., Valko, M., and Menard, P. Regularized rl. *arXiv preprint arXiv:2310.17303*, 2023.

Verma, R., Nagar, V., and Mahapatra, S. Introduction to supervised learning. *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 1–34, 2021.

Völske, M., Potthast, M., Syed, S., and Stein, B. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, 2017.

Wachi, A., Tran, T., Sato, R., Tanabe, T., and Akimoto, Y. Stepwise alignment for constrained language model

policy optimization. *Advances in Neural Information Processing Systems*, 37:104471–104520, 2024.

Wang, L., Krishnamurthy, A., and Slivkins, A. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 766–774. PMLR, 2024.

Wang, Y., Liu, Q., and Jin, C. Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:76006–76032, 2023a.

Wang, Z., Xie, J., Liu, X., Li, S., and Lui, J. Online clustering of bandits with misspecified user models. *Advances in Neural Information Processing Systems*, 36:3785–3818, 2023b.

Wang, Z., Xie, J., Yu, T., Li, S., and Lui, J. Online corrupted user detection and regret minimization. *Advances in Neural Information Processing Systems*, 36:33262–33287, 2023c.

Wang, Z., Sun, J., Kong, M., Xie, J., Hu, Q., Lui, J., and Dai, Z. Online clustering of dueling bandits. *arXiv preprint arXiv:2502.02079*, 2025.

Xiao, C., Wu, Y., Mei, J., Dai, B., Lattimore, T., Li, L., Szepesvari, C., and Schuurmans, D. On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*, pp. 11362–11371. PMLR, 2021.

Xiao, J., Li, Z., Xie, X., Getzen, E., Fang, C., Long, Q., and Su, W. J. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*, 2024.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

Xu, Y., Wang, R., Yang, L., Singh, A., and Dubrawski, A. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020.

Yan, C., Han, H., Zhang, Y., Zhu, D., and Wan, Y. Dynamic clustering based contextual combinatorial multi-armed bandit for online recommendation. *Knowledge-Based Systems*, 257:109927, 2022.

Ye, C., Xiong, W., Zhang, Y., Dong, H., Jiang, N., and Zhang, T. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37: 81773–81807, 2024.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline reinforcement learning with human feedback. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023a.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. How to query human feedback efficiently in rl? 2023b.

Zhang, C., Agarwal, A., Daumé III, H., Langford, J., and Negahban, S. N. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301*, 2019.

Zhang, C.-H. and Zhang, S. S. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242, 2014.

Zhong, H., Deng, Z., Su, W. J., Wu, Z. S., and Zhang, L. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*, 2024.

Zhu, B., Jordan, M., and Jiao, J. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023.

Appendix

The appendix is organized as follows. Appendix A reviews related work and comparisons of main results. Appendix B bridges our results with the clustering of bandits literature by specializing to the item regularity assumption, which enforces minimum coverage and is commonly adopted in prior work (Gentile et al., 2014; Li & Zhang, 2018; Li et al., 2019; Liu et al., 2025; Wang et al., 2025; Li et al., 2025b). Appendix C discusses the robustness of our algorithms when users within the same cluster are not identical and may exhibit small preference gaps. Appendix D provides additional discussion of our results. Detailed proofs of the theoretical guarantees are given in Appendix E, with technical lemmas collected in Appendix F. Finally, Appendix G contains omitted experimental details from Section 5.

A. Related Works and Comparisons of Main Results

Offline RL and Bandit. Offline statistical learning (Zhang & Zhang, 2014; Cai & Guo, 2017) primarily focuses on parameter estimation, while offline reinforcement learning (batch RL) extends the scope to sequential decision-making problems using fixed offline datasets (Lange et al., 2012; Levine et al., 2020; Jin et al., 2021; Rashidinejad et al., 2021; Xiao et al., 2021; Xie et al., 2021) and has found wide applications in diverse domains such as dialogue generation (Jaques et al., 2019), autonomous driving (Yurtsever et al., 2020), educational technologies (Singla et al., 2021) and personal recommendations (Li et al., 2010; Bottou et al., 2013). Within this landscape, offline bandits—a special case of offline RL—extend the multi-armed bandit framework to learning solely from pre-collected data (Shivaswamy & Joachims, 2012). Prior studies have considered settings where the offline distributions align with the online reward distributions (Bu et al., 2020; Banerjee et al., 2022) or where distribution shift arises between them (Zhang et al., 2019; Cheung & Lyu, 2024). Among them, studies on offline contextual linear bandits (Li et al., 2022; Wang et al., 2024) are most closely related to our setting. However, beyond the standard contextual linear bandits formulation, our work studies pairwise feedback modeled by a logistic function. Moreover, we explicitly leverage the clustering structure and a hybrid setting with active data augmentation of offline user preference data to address coverage limitations existing in prior algorithms and improve learning efficiency, yielding a more general and realistic setting than prior work.

Preference Learning from Pairwise Feedback. Theoretical studies of preference learning from pairwise feedback trace back to the dueling bandit problem (Yue et al., 2012; Saha, 2021; Bengs et al., 2021) and its extension, the contextual dueling bandit problem (Dudík et al., 2015). These ideas extend naturally to preference-based reinforcement learning (Xu et al., 2020; Chen et al., 2022; Wang et al., 2023a; Zhan et al., 2023b). Recent work has emphasized offline preference-based RL, often motivated by reinforcement learning with human feedback (RLHF). Approaches include pessimism-driven methods (Zhu et al., 2023; Zhan et al., 2023a; Li et al., 2023) and KL-regularized formulations (Tiapkin et al., 2023; Xiong et al., 2023; Xiao et al., 2024). For instance, Xiong et al. (2023) study active context selection under strong coverage assumptions, deriving sample-dependent bounds. Beyond RLHF, researchers have explored general preference structures (Rosset et al., 2024; Gui et al., 2024; Ye et al., 2024), pure active preference learning without offline datasets (Das et al., 2024), safety-constrained alignment (Wachi et al., 2024), and sample-efficient learning under limited data (Kim et al., 2025). Our work departs from these above mentioned works by explicitly incorporating offline clustering into pairwise preference learning and combining it with active data augmentation to deal with the coverage issue existing in prior works. 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 these challenges through algorithm design and theoretical analysis that reveal the benefits of improving coverage via offline clustering and active-data augmentation in both pure offline and hybrid settings.

Heterogeneous Preference Learning. Heterogeneous preference learning has been widely studied under the clustering of bandits (Gentile et al., 2014; Li & Zhang, 2018; Li et al., 2019) and multi-task learning (Duan & Wang, 2023), where data from users with distinct preference vectors could be used to accelerate learning. Later works investigate privacy (Liu et al., 2022), model misspecification (Wang et al., 2023b), and robustness to corrupted users (Wang et al., 2023c). More recent studies by Liu et al. (2025) provide offline algorithms for clustering of contextual bandits, and Wang et al. (2025) considers online setting of preference learning from pairwise feedback. With growing interest in RLHF, recent efforts have addressed scenarios involving users with diverse preferences, which are often referred to as personalized RLHF (Kirk et al., 2023; Li et al., 2024; Conitzer et al., 2024; Jang et al., 2023; Poddar et al., 2024; Ramesh et al., 2024). Theoretically, Liu et al. (2024) study heterogeneous user rationality, Zhong et al. (2024) focus on meta-learning and social welfare aggregation, and Park et al. (2024) analyze representation-based aggregation under assumptions on uniqueness, diversity, and concentrability.

Table 1. Summary of main and additional theoretical results.

Comparisons of Algorithms for Pure Offline Model				
	Algorithm	**Setting**	**Condition**	**Suboptimality**
Previous	P-MLE (Zhu et al., 2023) PDC (Li et al., 2025a)	Pure Offline Single User	—	$\tilde{O}\left(\sqrt{\frac{d}{\lambda_1}}\right)$
Main Result 1 (Theorem 3.3)	Off-C^{2}PL (Algorithm 1)	Pure Offline Multiple Users	—	$\tilde{O}\left(\frac{\sqrt{d}\left(1+\hat{\gamma}\sqrt{N_1}\right)}{\sqrt{\lambda_2}}\right)$
Additional Result 1 (Equation (10))	Off-C^{2}PL (Algorithm 1)	Pure Offline Multiple Users	Lower Threshold $\hat{\gamma} \leq \gamma$ (Definition 3.1)	$\tilde{O}\left(\sqrt{\frac{d}{\lambda_2}}\right)$
Comparisons of Algorithms for Active Data-Augmented Model				
	Algorithm	**Setting**	**Condition**	**Suboptimality**
Previous	APO (Das et al., 2024)	Pure Active Single User	—	$\tilde{O}\left(\sqrt{\frac{d}{N/d}}\right)$
Main Result 2 (Theorem 4.1)	ADA-Off-C^{2}PL (Algorithm 2)	Hybrid (Offline + Active) Multiple Users	—	$\tilde{O}\left(\frac{\sqrt{d}\left(1+\hat{\gamma}\sqrt{N_1}\right)}{\sqrt{\lambda_3+N/d}}\right)$
Additional Result 2 (Corollary 4.4)	ADA-Off-C^{2}PL (Algorithm 2)	Hybrid (Offline + Active) Multiple Users	Lower Threshold $\hat{\gamma} \leq \gamma$ (Definition 3.1) + Imbalanced Dataset (Definition 4.2)	$\tilde{O}\left(\sqrt{\frac{d}{\lambda_2+N/d^*}}\right)$

Here, d denotes the dimension of each user's preference vector, and $d^* \leq d$ denotes the number of sparsely informed dimensions, as defined in Definition 4.2. N_1 denotes the number of heterogeneous offline samples included. 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.

Compared to prior work, our contributions differ in three key aspects. First, we extend prior online settings and study offline clustering of bandits under pairwise feedback, which more naturally captures applications in recommendation systems and RLHF. Second, we remove the dependence of strong coverage assumptions (e.g., Assumption B.1 in Appendix B, commonly adopted in prior work (Gentile et al., 2014; 2017; Li et al., 2019; Wang et al., 2025; Li et al., 2025b; Liu et al., 2025)) and instead explicitly handle poor coverage through the minimum eigenvalue of the information matrix. Third, while existing works focus on either purely online or purely offline settings, we consider both offline aggregation and a hybrid offline with active augmentation framework.

B. Bridging to Classical Clustering of Bandits under Coverage Assumptions

In this appendix, we introduce the *item regularity* assumption, which is widely adopted in prior work on online and offline clustering of bandits (Gentile et al., 2014; 2017; Li et al., 2019; Wang et al., 2025; Li et al., 2025b; Liu et al., 2025). This assumption enforces a minimum level of coverage in the offline data $\mathcal{D}_u$ for each user u. We then specialize our theoretical results under this assumption to connect our analysis with existing clustering of bandits literature and to illustrate how our guarantees simplify when such coverage conditions are imposed. Detailed proofs of the results in this appendix are provided in Section E. We begin by formally stating the item regularity in Assumption B.1.

Assumption B.1 (Item Regularity)**.** Let ρ be a distribution over $\{(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{a}') \in \mathcal{X} \times \mathcal{A} \times \mathcal{A} : \|\phi(\boldsymbol{x}, \boldsymbol{a})\|_2 \leq 1,\ \|\phi(\boldsymbol{x}, \boldsymbol{a}')\|_2 \leq 1\}$ where coveriance matrix $\mathbb{E}_{(\boldsymbol{x},\boldsymbol{a},\boldsymbol{a}')\sim\rho_a}\left[(\phi(\boldsymbol{x}, \boldsymbol{a}) - \phi(\boldsymbol{x}, \boldsymbol{a}'))(\phi(\boldsymbol{x}, \boldsymbol{a}) - \phi(\boldsymbol{x}, \boldsymbol{a}'))^\top\right]$ is full rank with minimum eigenvalue $\lambda_a > 0$. For any fixed unit vector $\boldsymbol{\theta} \in \mathbb{R}^d$, the random variable $\left(\boldsymbol{\theta}^\top(\phi(\boldsymbol{x}, \boldsymbol{a}) - \phi(\boldsymbol{x}, \boldsymbol{a}'))\right)^2$, with $(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{a}') \sim \rho$, has sub-Gaussian tails with variance upper bounded by σ^2. Each context-action pair $(\boldsymbol{x}_u^i, \boldsymbol{a}_u^i, \boldsymbol{a}'^i_u)$ 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* $\tilde{\lambda}_a = \int_0^{\lambda_a} \left(1 - e^{-\frac{(\lambda_a - x)^2}{2\sigma^2}}\right)^S \mathrm{d}x$ is known to the algorithm.

Assumption B.1 ensures that the data distribution is sufficiently rich to provide informative samples in all directions of the preference vector $\boldsymbol{\theta}_u$, thereby offering a good coverage of the optimal policy. 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 (Dudík et al., 2015; Wang et al., 2025). 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 directly with the amount of offline data rather than depending solely

on the minimum eigenvalue itself. We then discuss the corresponding theoretical results under Assumption B.1 for the pure offline setting.

B.1. Results and Comparisons for Pure Offline Setting under Assumption B.1

First, Lemma B.2 summarizes the modified clustering conditions and resulting characterizations.

Lemma B.2 (Extension of Lemma 3.2)**.** *Under Assumption B.1, replace the confidence interval by* $\mathrm{CI}_u = \left(\sqrt{\lambda\kappa} + 2\sqrt{d\,\log\left(1+\frac{4\kappa N_u}{\lambda d}\right) + 2\log\left(\frac{2U}{\delta}\right)}\right) \Big/ \left(\kappa\sqrt{\tilde{\lambda}_a N_u/2}\right)$*, and adjust the condition in Equation (3) to:*

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

where $N_{\min} = \frac{16}{\tilde{\lambda}_a^2}\log\left(\frac{8Ud}{\tilde{\lambda}_a^2\delta}\right)$*. All other conditions remain as in Lemma 3.2. 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:*

$$\mathcal{R}_{\hat{\gamma}}(u) = \begin{cases} \left\{v \,\middle|\, \boldsymbol{\theta}_u = \boldsymbol{\theta}_v,\ \frac{1}{\sqrt{N_u}} + \frac{1}{\sqrt{N_v}} < \alpha_r'\hat{\gamma},\ N_v \geq N_{\min}\right\} \cup \{u\}, & N_u \geq N_{\min} \\ \{u\}, & \textit{otherwise} \end{cases}, \tag{15}$$

$$\mathcal{W}_{\hat{\gamma}}(u) = \begin{cases} \left\{v \,\middle|\, \gamma \leq \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2 < \hat{\gamma},\ \frac{1}{\sqrt{N_u}} + \frac{1}{\sqrt{N_v}} < \alpha_w'\varepsilon\right\}, & N_u \geq N_{\min} \\ \emptyset, & \textit{otherwise} \end{cases}. \tag{16}$$

The expressions above show that, under Assumption B.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 (Liu et al., 2025). Below we present Corollary B.3, which characterizes the suboptimality of our algorithm under Assumption B.1.

Corollary B.3. *Under the same conditions as in Lemma B.2, the suboptimality of the algorithm is bounded with probability at least* $1-\delta$ *as:*

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

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 B.3 takes a form similar to the suboptimality bounds in classical offline clustering of bandits (Liu et al., 2025). 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 Section D.1. Finally, the scaling factor $\sqrt{d/\tilde{\lambda}_a}$ arises from Assumption B.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.

B.2. Results and Comparisons for Active-data Augmented Setting under Assumption B.1

Next, we present special-case results for those with the active-data augmented model under the item regularity assumption (Assumption B.1) and the condition that $\tilde{M}_{u_t}$ is (d^*, N)-dimension imbalanced, which illustrates the benefit of active-data augmentation even in a traditional clustering of bandits context:

Corollary B.4. *Suppose Assumption B.1 holds and that $\tilde{M}_{u_t}$ is (d^*, N)-dimension imbalanced. Following the proof of Corollary B.3, it holds that*

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

Corollary B.4 can be interpreted in terms of *noise* (the first term) and *bias* (the second term). Importantly, under Assumption B.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. (2023b)). This advantage arises because active samples offer better coverage through the active selection rule than the coverage offered by Assumption B.1 for offline samples. Consequently, this result strengthens Corollary B.3, yielding a strictly better suboptimality bound by reducing both noise and bias.

C. Robustness of Algorithms for Non-identical Within-cluster Users

Sometimes the assumption that users within the same cluster have exactly identical preferences can be restrictive in realistic scenarios, since small discrepancies may still arise even among users with similar backgrounds. For example, annotators or users from similar backgrounds may still exhibit slight differences in their preferences because of individual values or experiences. In this section, we show that our proposed algorithms (Algorithm 1 and 2) remain robust under a weaker setting that allows such small within-cluster gaps, requiring only minor modifications to the theoretical guarantees and no changes to the algorithms themselves. We formalize this setting as follows:

Non-identical Within-cluster Users. Still consider U users indexed by $\mathcal{U} = [U]$, partitioned into J clusters. Let $\mathcal{V}(j_u)$ denote the cluster containing user u. For each cluster j, let $\boldsymbol{\theta}^j$ denote its *common preference vector*, which represents the center of the user preferences in that cluster, and let $\boldsymbol{\theta}_u$ denote the *user preference vector* of user u which represents its own preference. We assume that the common preference vectors of different clusters are separated by at least γ (Definition 3.1), that is, $\|\boldsymbol{\theta}^{j_1} - \boldsymbol{\theta}^{j_2}\|_2 \geq \gamma$ for all $j_1 \neq j_2$. In addition, for each user u, the deviation between the user preference vector and the corresponding cluster center is at most ζ (where $\zeta \ll \gamma$), namely $\|\boldsymbol{\theta}_u - \boldsymbol{\theta}^{j_u}\|_2 < \zeta$.

Compared with the identical within-cluster user setting considered in the main body, the present setting allows a small within-cluster gap ζ among users in the same cluster. When $\zeta = 0$, it reduces to the setting studied in the main body. To state the corresponding results for the non-identical clustering setting, we first modify the set definitions in Equation (5) as follows:

$$\mathcal{R}_{\hat{\gamma}}(u) \coloneqq \{v \mid v \in \mathcal{V}_{\hat{\gamma}}(u) \cap \mathcal{V}(j_u)\},\; \mathcal{W}_{\hat{\gamma}}(u) \coloneqq \{v \mid v \in \mathcal{V}_{\hat{\gamma}}(u) \backslash \mathcal{R}_{\hat{\gamma}}(u)\}. \tag{17}$$

These two sets retain the similar interpretation as in the main setting: they distinguish between users connected to u in Algorithm 1 who belong to the same cluster and those who belong to different clusters. However, their formal definitions must be adjusted to account for the within-cluster gap ζ. We next characterize the cardinalities of both sets, similar to those in Lemma 3.2.

Lemma C.1 (Cardinality of $\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$ under Non-identical Setting)**.** *For any user u, let inputs in Algorithm 1 satisfy $\alpha \geq 1$, $\kappa = 1/(2 + e^2 + e^{-2})$, λ and δ satisfy $\lambda \leq 2\log(2U/\delta) + d\log\left(1 + \frac{4\kappa \min_v\{N_v\}}{d\lambda}\right)$, $\delta \leq \frac{d\lambda}{4\kappa \min_v\{N_v\} + d\lambda}$. Then there exist some α''_r and α''_w both in $\left(0, \frac{\kappa}{2(\alpha-1)\sqrt{2\log(2U/\delta)}}\right)$ such that $\mathcal{R}_{\hat{\gamma}}(u)$ and $\mathcal{W}_{\hat{\gamma}}(u)$ can be characterized with probability at least $1 - \delta$ as:*

$$\mathcal{R}_{\hat{\gamma}}(u) = \{u\} \cup \left\{v \,\middle|\, \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2 < \min\{\hat{\gamma}, \zeta\} \text{ and } \frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha''_r\hat{\gamma}\right\}, \tag{18}$$

$$\mathcal{W}_{\hat{\gamma}}(u) = \left\{v \,\middle|\, \gamma \leq \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2 < \hat{\gamma} \text{ and } \frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha''_w(\hat{\gamma} - \gamma)\right\}. \tag{19}$$

Lemma C.1 also has the similar form as Lemma 3.2, except for the first condition in the definition of $\mathcal{R}_{\hat{\gamma}}(u)$, which is modified to account for non-identical within-cluster users. This change is important because users in $\mathcal{R}_{\hat{\gamma}}(u)$ are no longer perfectly homogeneous. As a result, the small within-cluster gap ζ appears explicitly in the theoretical guarantees below, quantifying the additional error introduced by within-cluster heterogeneity.

Theorem C.2. *Under the same conditions as in Lemma C.1, the suboptimality of Algorithm 1 in the non-identical setting can be bounded with probability at least $1-\delta$ as follows:*

$$\begin{aligned}\text{SubOpt}_{u_t}(\pi_{u_t}) &\le \tilde{O}\big(\sqrt{d}\big(1+\zeta\sqrt{N_{\mathcal{R}_{\hat{\gamma}}(u_t)}}+\hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}\big)\big\|\overline{\phi}(\pi^*_{u_t})-\boldsymbol{w}\big\|_{\tilde{M}_{u_t}^{-1}}\big)\\ &\le \tilde{O}\left(\frac{\sqrt{d}\big(1+\zeta\sqrt{N_{\mathcal{R}_{\hat{\gamma}}(u_t)}}+\hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}\big)}{\sqrt{\lambda_{\min}(\tilde{M}_{u_t})}}\right).\end{aligned}$$

Similarly, Theorem 4.1 and Corollary 4.4 can be extended to the non-identical setting by replacing $1+\hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}$ with $1+\zeta\sqrt{N_{\mathcal{R}_{\hat{\gamma}}(u_t)}}+\hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u_t)}}$ in their numerators.

Lemma C.1 and Theorem C.2 together show that the algorithms proposed in the main body extend naturally to the non-identical within-cluster users setting, without requiring any modification to the algorithms themselves. At the theoretical level, the resulting guarantees differ only through an additional term involving the within-cluster gap ζ in the numerator. In practice, ζ is expected to be much smaller than the inter-cluster gap γ (resulting $\zeta \ll \gamma$ as discussed before), so this extra term should have only a limited effect on the suboptimality bound. This observation highlights the robustness of our approach to mild within-cluster heterogeneity, which is likely to arise in realistic applications.

Small within-cluster preference gaps have also been studied in some prior work on online clustering of bandits (Wang et al., 2023b; Dai et al., 2024), where the authors adapt edge-deletion strategies (Gentile et al., 2014; Li & Zhang, 2018) to handle such heterogeneity. A key difference is that those approaches require an a prior knowledge of the upper bound on the within-cluster gap and must explicitly incorporate that quantity into the algorithm design. By contrast, our connection-based approach does not require prior knowledge of ζ and remains effective under this more general setting. This distinction highlights an additional practical advantage of our method in non-identical settings, especially in applications where the degree of within-cluster heterogeneity is unknown in advance.

In the main body, we adopt the simplifying assumption that users within the same cluster share identical preference vectors, rather than the more general formulation with a small within-cluster gap ζ. This choice helps keep the presentation and the main theoretical results clear and streamlined, while also maintaining consistency with the primary prior literature on clustering of bandits (Gentile et al., 2014; Li & Zhang, 2018; Liu et al., 2025; Wang et al., 2025). Furthermore, the extension developed in this appendix shows that this simplification does not limit the applicability of our methods to more general settings that allow a nonzero within-cluster gap.

Finally, our synthetic experiments in Section 5 already introduce mild gaps for users within the same cluster, where each user is generated as $\boldsymbol{\theta}_u = \boldsymbol{\theta}^{j_u} + \boldsymbol{\epsilon}_u$ (as detailed in Appendix G.2). The real-world Reddit dataset in Section 5 also inherently reflects the preferences of real users and therefore does not satisfy a perfectly identical within-cluster user model. The fact that our methods still achieve strong improvements in both settings provides empirical support that they remain effective even under the more general non-identical within-cluster users setting.

D. Additional Discussions

D.1. Selection of $\hat{\gamma}$

This appendix elaborates practical policies for choosing the clustering threshold $\hat{\gamma}$. Our treatment closely follows the guidance in Liu et al. (2025); we include their ideas here for completeness and refer readers there for additional discussion. As shown in Lemma 3.2, 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 $\boldsymbol{\theta}_u$, whereas a larger $\mathcal{W}_{\hat{\gamma}}(u)$ may introduce greater bias due to the inclusion of heterogeneous neighbors (as analyzed in Theorem 3.3). Therefore, careful selection of $\hat{\gamma}$ is crucial. Notably, Equation (10) shows that choosing $\hat{\gamma} \le \gamma$ 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. Below we discuss two cases for selecting $\hat{\gamma}$ under known or unknown γ.

Case 1: Known γ. When the minimum heterogeneity gap γ (defined in Definition 3.1) is known, a natural choice is $\hat{\gamma} = \gamma$, which exactly separates users across clusters.

Remark D.1 (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 3.2 and Equation (10) 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.

Case 2: Unknown γ. When γ is unknown, the threshold $\hat{\gamma}$ must be estimated from the offline data. We define

$$\Gamma(u,v) = \|\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_v\|_2 - \alpha(\mathrm{CI}_u + \mathrm{CI}_v), \qquad M(u) = \{v \in \mathcal{U} \setminus \{u\} \,:\, \Gamma(u,v) > 0\}, \tag{20}$$

where CI_u is given in line 2 of Algorithm 1. For $\alpha \geq 1$, $\Gamma(u,v) \leq \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2$ 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 D.2 (Underestimation Policy)**.** The underestimation policy is defined as:

$$\hat{\gamma} = \mathbb{I}\{M(u_t) \neq \emptyset\} \cdot \min_{v \in M(u_t)} \Gamma(u_t, v). \tag{21}$$

Theorem D.3 (Effect of the Underestimation Policy)**.** *With $\hat{\gamma}$ chosen by Equation (21) 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 3.2 also satisfies*

$$\frac{1}{\sqrt{\lambda_{\min}(M_{u_t})}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} \;\geq\; \alpha'_w \, \|\boldsymbol{\theta}_{u_t} - \boldsymbol{\theta}_v\|_2.$$

Remark D.4 (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 D.5 (Overestimation Policy)**.** The overestimation policy is defined as:

$$\hat{\gamma} = \mathbb{I}\{M(u_t) \neq \emptyset\} \cdot \min_{v \in M(u_t)} \tilde{\Gamma}(u_t, v), \tag{22}$$

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

Theorem D.6 (Effect of the overestimation policy)**.** *Under the policy in Definition D.5, if $M(u_t) \neq \emptyset$ then $\hat{\gamma} \geq \gamma$.*

Remark D.7 (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 D.3 and D.6, see Liu et al. (2025).

D.2. 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.,

$$\min_{(\boldsymbol{x},\boldsymbol{a},\boldsymbol{a}') \in \mathcal{X}\times\mathcal{A}\times\mathcal{A},\; \boldsymbol{\theta}\in\Theta} \nabla\sigma\left(\phi(\boldsymbol{x},\boldsymbol{a})^\top\boldsymbol{\theta} - \phi(\boldsymbol{x},\boldsymbol{a}')^\top\boldsymbol{\theta}\right) \;\geq\; \kappa > 0. \tag{23}$$

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 3.3). This is because we assume $\|\boldsymbol{\theta}_u\|_2 \leq 1$ and $\|\phi(\boldsymbol{x},\boldsymbol{a})\|_2 \leq 1$, following prior works on contextual logistic bandits (Chen et al., 2020; Oh & Iyengar, 2019; Lee & Oh, 2024) and clustering of bandits literature (Gentile et al.,

2014; Wang et al., 2023b; 2025; Liu et al., 2025). In more general scenarios where the ℓ_2-norm of either $\boldsymbol{\theta}_u$ or $\phi(\boldsymbol{x}, \boldsymbol{a})$ 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 F.2 and Section E.2 (proof of Theorem 3.3), 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 $1/\sqrt{\kappa}$ (Li et al., 2025a), which is argued to be tight (Das et al., 2024; Li et al., 2025a). Extending the dependence on $\sqrt{\kappa}$ to our heterogeneous multi-user setting with clustering remains an interesting open problem.

E. Detailed Proofs

E.1. Proof of Lemma 3.2

Proof. In order to prove Lemma 3.2, it suffices to show the following statement: under the same conditions as in Lemma 3.2, both sets can be characterized as

$$\mathcal{R}_{\hat{\gamma}}(u) = \left\{ v \,\middle|\, \boldsymbol{\theta}_u = \boldsymbol{\theta}_v \text{ and } \frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha_r \hat{\gamma} \right\} \cup \{u\},$$
$$\mathcal{W}_{\hat{\gamma}}(u) = \left\{ v \,\middle|\, \gamma \leq \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2 < \hat{\gamma} \text{ and } \frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}} < \alpha_w(\hat{\gamma} - \gamma) \right\}$$

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 F.1 and a union bound, we have that the event

$$\mathcal{E} := \bigcap_{u \in \mathcal{U}} \left\{ \|\hat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_u\|_2 \leq \mathrm{CI}_u \right\}$$

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

Recall that the connection condition in Algorithm 1 is given by

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

which implies

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

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 $\boldsymbol{\theta}_u = \boldsymbol{\theta}_v$. To prove the claim for $\mathcal{R}_{\hat{\gamma}}(u)$ in Lemma 3.2, 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

$$\begin{aligned}
&(\alpha+1)\left(\mathrm{CI}_u + \mathrm{CI}_v\right) && (24)\\
&\leq \frac{3(\alpha+1)\sqrt{2\log(2U/\delta) + d\log(1+4N_u\kappa/(d\lambda))}}{\kappa\sqrt{\lambda_{\min}(M_u)}} + \frac{3(\alpha+1)\sqrt{2\log(2U/\delta) + d\log(1+4N_v\kappa/(d\lambda))}}{\kappa\sqrt{\lambda_{\min}(M_v)}}\\
&\leq \frac{3(\alpha+1)\sqrt{2\max\{2,d\}\log(2U/\delta)}}{\kappa}\left(\frac{1}{\sqrt{\lambda_{\min}(M_u)}} + \frac{1}{\sqrt{\lambda_{\min}(M_v)}}\right) < \hat{\gamma}, && (25)
\end{aligned}$$

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

$$\left\|\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_v\right\|_2 \leq \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2 + \mathrm{CI}_u + \mathrm{CI}_v \overset{(a)}{=} \mathrm{CI}_u + \mathrm{CI}_v \overset{(b)}{\leq} \hat{\gamma} - \alpha(\mathrm{CI}_u + \mathrm{CI}_v),$$

where (a) uses $\boldsymbol{\theta}_u = \boldsymbol{\theta}_v$, and (b) follows from (24). 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

$$\begin{aligned}
(\alpha-1)\left(\mathrm{CI}_u + \mathrm{CI}_v\right) &\geq \frac{2(\alpha-1)}{\kappa}\sqrt{\frac{2\log(2U/\delta)}{\lambda_{\min}(M_u)}} + \frac{2(\alpha-1)}{\kappa}\sqrt{\frac{2\log(2U/\delta)}{\lambda_{\min}(M_v)}}\\
&\geq \hat{\gamma}. && (26)
\end{aligned}$$

Therefore, it follows that

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

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 (26), we have $(\alpha-1)(\mathrm{CI}_u + \mathrm{CI}_v) \geq (\hat{\gamma} - \gamma)$. This yields

$$(\hat{\gamma} - \gamma) - \alpha(\mathrm{CI}_u + \mathrm{CI}_v) \leq -(\mathrm{CI}_u + \mathrm{CI}_v) \leq \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_v\|_2 - (\mathrm{CI}_u + \mathrm{CI}_v) - \gamma \leq \|\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_v\|_2 - \gamma,$$

which implies

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

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

□

E.2. Proof of Theorem 3.3

Proof. By Lemmas F.1 and F.2, we have

$$\left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u\right\|_{\tilde{M}_u} \leq \tilde{\beta}_u \tag{27}$$

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

For simplicity, let $u = u_t$ denote the test user. Define $J'_u(\pi) = J_u(\pi) - \langle \boldsymbol{\theta}_u, \boldsymbol{w} \rangle$. 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, according to Appendix D.2 of Zhu et al. (2023), it holds that $\tilde{J}_u(\pi_u) - J'_u(\pi_u) \leq 0$.

Finally, for the first term:

$$\begin{aligned}J'_u(\pi_u^*) - \tilde{J}_u(\pi_u^*) &= \left(\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u\right)^\top \left(\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x}))] - \boldsymbol{w}\right) + \tilde{\beta}_u \left\|\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x}))] - \boldsymbol{w}\right\|_{\tilde{M}_u^{-1}} \\ &\leq \left(\left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u\right\|_{\tilde{M}_u} + \tilde{\beta}_u\right) \left\|\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x}))] - \boldsymbol{w}\right\|_{\tilde{M}_u^{-1}} \\ &\leq 2\tilde{\beta}_u \left\|\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x}))] - \boldsymbol{w}\right\|_{\tilde{M}_u^{-1}}.\end{aligned}$$

Putting everything together, we obtain:

$$\begin{aligned}\text{SubOpt}_u(\pi_u) &\leq 2\tilde{\beta}_u \left\|\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x}))] - \boldsymbol{w}\right\|_{\tilde{M}_u^{-1}} \\ &= \tilde{O}\left(\sqrt{d}\left(1 + \hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u)}}\right) \left\|\mathbb{E}_{\boldsymbol{x}\sim\rho_p}[\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x}))] - \boldsymbol{w}\right\|_{\tilde{M}_u^{-1}}\right),\end{aligned}$$

which concludes the proof of Theorem 3.3. □

E.3. Proof of Theorem 4.1

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

$$\begin{aligned}&\text{SubOpt}_u(\pi_u, \boldsymbol{x}) := \boldsymbol{\theta}_u^\top \left(\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right), \\ &\overline{\beta}_u^n := \frac{2\sqrt{d\,\log\left(1 + \frac{4\kappa(\tilde{N}_u + n)}{\lambda d}\right) + 2\log(2U/\delta)} + \sqrt{\lambda\kappa}}{\kappa}.\end{aligned}$$

First, note that by Lemma F.2 and Lemma F.3, since the cardinality of the heterogeneous neighbor set $\mathcal{W}_{\hat{\gamma}}(u)$ remains unchanged during the online phase, we have

$$\left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^n\right\|_{\tilde{M}_u^n} \leq \overline{\beta}_u^n + \frac{\hat{\gamma}}{2}\sqrt{d\,N_{\mathcal{W}_{\hat{\gamma}}(u)}} \quad \text{for each } n \in [N], \tag{28}$$

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 $\text{SubOpt}_u(\pi_u, \boldsymbol{x})$. It holds that

$$\text{SubOpt}_u(\pi_u, \boldsymbol{x}) \tag{29}$$

$$
\begin{aligned}
&= \boldsymbol{\theta}_u^\top \left(\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right) \\
&\leq \boldsymbol{\theta}_u^\top \left(\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right) + \overline{\boldsymbol{\theta}}_u^\top \left(\phi(\boldsymbol{x}, \pi_u(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x}))\right) \\
&= \left(\boldsymbol{\theta}_u - \overline{\boldsymbol{\theta}}_u\right)^\top \left(\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right) \\
&= \left(\boldsymbol{\theta}_u - \frac{1}{d\,\lambda_{\min}\left(\tilde{M}_u^N\right) + N}\left(d\,\lambda_{\min}\left(\tilde{M}_u^N\right)\,\tilde{\boldsymbol{\theta}}_u^N + \sum_{n=1}^N \tilde{\boldsymbol{\theta}}_u^n\right)\right)^\top \left(\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right) \\
&= \frac{1}{d\,\lambda_{\min}\left(\tilde{M}_u^N\right) + N}\left(d\,\lambda_{\min}\left(\tilde{M}_u^N\right)\left(\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^N\right)^\top + \sum_{n=1}^N \left(\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^n\right)^\top\right)\left(\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right). \qquad (30)
\end{aligned}
$$

Next, for the first term in (30), we have:

$$
\begin{aligned}
&\left(\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^N\right)^\top \left(\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right) \\
&\overset{(a)}{\leq} \left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^N\right\|_2 \left\|\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right\|_2 \\
&\overset{(b)}{\leq} 2\frac{\left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^N\right\|_{\tilde{M}_u^N}}{\sqrt{\lambda_{\min}\left(\tilde{M}_u^N\right)}} \\
&\overset{(c)}{\leq} \frac{2\,\overline{\beta}_u^N + \hat{\gamma}\sqrt{d\,N_{\mathcal{W}_{\hat{\gamma}}(u)}}}{\sqrt{\lambda_{\min}\left(\tilde{M}_u^N\right)}}. \qquad (31)
\end{aligned}
$$

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 (28).

For the summation term in (30), we have:

$$
\begin{aligned}
&\sum_{n=1}^N \left(\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^n\right)^\top \left(\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right) \\
&\leq \sum_{n=1}^N \left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^n\right\|_{\tilde{M}_u^n} \left\|\phi(\boldsymbol{x}, \pi_u^*(\boldsymbol{x})) - \phi(\boldsymbol{x}, \pi_u(\boldsymbol{x}))\right\|_{(\tilde{M}_u^n)^{-1}} \\
&\overset{(a)}{\leq} \sum_{n=1}^N \left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u^n\right\|_{\tilde{M}_u^n} \left\|\phi(\mathring{\boldsymbol{x}}_u^n, \mathring{\boldsymbol{a}}_u^n) - \phi(\mathring{\boldsymbol{x}}_u^n, \mathring{\boldsymbol{a}}'^n_u)\right\|_{(\tilde{M}_u^n)^{-1}} \\
&\overset{(b)}{\leq} \sum_{n=1}^N \left(2\,\overline{\beta}_u^n + \hat{\gamma}\sqrt{d\,N_{\mathcal{W}_{\hat{\gamma}}(u)}}\right) \left\|\phi(\mathring{\boldsymbol{x}}_u^n, \mathring{\boldsymbol{a}}_u^n) - \phi(\mathring{\boldsymbol{x}}_u^n, \mathring{\boldsymbol{a}}'^n_u)\right\|_{(\tilde{M}_u^n)^{-1}} \\
&\overset{(c)}{\leq} \left(2\,\overline{\beta}_u^N + \hat{\gamma}\sqrt{d\,N_{\mathcal{W}_{\hat{\gamma}}(u)}}\right) \sum_{n=1}^N \left\|\phi(\mathring{\boldsymbol{x}}_u^n, \mathring{\boldsymbol{a}}_u^n) - \phi(\mathring{\boldsymbol{x}}_u^n, \mathring{\boldsymbol{a}}'^n_u)\right\|_{(\tilde{M}_u^n)^{-1}} \\
&\overset{(d)}{\leq} \left(2\,\overline{\beta}_u^N + \hat{\gamma}\sqrt{d\,N_{\mathcal{W}_{\hat{\gamma}}(u)}}\right)\sqrt{N}\sqrt{\sum_{n=1}^N \left\|\phi(\mathring{\boldsymbol{x}}_u^n, \mathring{\boldsymbol{a}}_u^n) - \phi(\mathring{\boldsymbol{x}}_u^n, \mathring{\boldsymbol{a}}'^n_u)\right\|^2_{(\tilde{M}_u^n)^{-1}}} \\
&\overset{(e)}{\leq} \left(2\,\overline{\beta}_u^N + \hat{\gamma}\sqrt{d\,N_{\mathcal{W}_{\hat{\gamma}}(u)}}\right)\sqrt{2dN\,\log\left(1 + \frac{4\kappa N}{\lambda d}\right)}. \qquad (32)
\end{aligned}
$$

Here, (a) holds by the active data augmentation rule in line 4 of Algorithm 2; (b) uses the ellipsoid bound (28); (c) holds because $\overline{\beta}_u^n$ is non-decreasing in n; (d) applies the Cauchy–Schwarz inequality; and (e) follows from the elliptical potential lemma (Lemma F.4).

Combining Equation (30), Equation (31), and Equation (32) yields:

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

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

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

which completes the proof of Theorem 4.1. □

E.4. Proof of Lemma 4.3

Proof. According to Lemma F.7, under the active data augmentation rule in Equation (12), 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 $i \in \left\{1, \cdots, \lfloor \frac{N}{d^*} \rfloor\right\}$,

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

Therefore, we have:

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

where we define $\lambda_{\min}\left(\tilde{M}_{u_t}^0\right) = \lambda_{\min}\left(\tilde{M}_{u_t}\right)$ to be the minimum eigenvalue of the Gramian matrix constructed from the aggregated offline data. This completes the proof of Lemma 4.3. □

E.5. Proof of Corollary 4.4

Combining Theorem 4.1 and Lemma 4.3, and noting that $d^* \leq d$, we obtain Corollary 4.4.

E.6. Proof of Lemma B.2

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 F.1, Lemma J.1 in Wang et al. (2023b) and Lemma 7 in (Li & Zhang, 2018), 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

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

with probability at least $1-\delta$.

Finally, by following the same argument used in the proof of Lemma 3.2, but replacing $\lambda_{\min}(M_u)$ with the explicit bound on N_u under Assumption B.1, we obtain the desired result in Lemma B.2. □

E.7. Proof of Corollary B.3

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

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

Here the first inequality follows directly from Theorem 3.3, while the second inequality applies Lemma J.1 in Wang et al. (2023b) and Lemma 7 in Li & Zhang (2018). This completes the proof of Corollary B.3.

□

E.8. Proof of Lemma C.1

Proof. The proof for $\mathcal{W}_{\hat{\gamma}}(u)$ in Lemma C.1 is identical to that of Lemma 3.2. Likewise, the bound on the size of $\mathcal{R}_{\hat{\gamma}}(u)$ can be treated as a special case of the proof for $\mathcal{W}_{\hat{\gamma}}(u)$ in Lemma C.1 by setting $\gamma = 0$. □

E.9. Proof of Theorem C.2

Proof. We first prove an adjusted version of Lemma F.2 for the non-identical setting, showing that

$$\left\|\tilde{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_u\right\|_{\tilde{M}_u} \leq \tilde{O}\left(\sqrt{d}\left(\frac{1}{\kappa} + \zeta\sqrt{N_{\mathcal{R}_{\hat{\gamma}}(u)}} + \hat{\gamma}\sqrt{N_{\mathcal{W}_{\hat{\gamma}}(u)}}\right)\right). \tag{33}$$

The first part of the proof closely follows that of Lemma F.2, except that we replace Equation (45) with

$$\begin{aligned}
&\frac{1}{\kappa^2}\left\|\tilde{G}_u(\tilde{\boldsymbol{\theta}}_u)\right\|_{\tilde{M}_u^{-1}} \\
&\leq \left(\frac{1}{\kappa}\left\|\sum_v\sum_i \varepsilon_v^i \boldsymbol{z}_v^i\right\|_{\tilde{M}_u^{-1}} + \frac{1}{\kappa}\left\|\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)}\sum_i\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\right\|_{\tilde{M}_u^{-1}}\right)^2,
\end{aligned} \tag{34}$$

where we have

$$\begin{aligned}
&\left\|\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)}\sum_i\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\right\|_{\tilde{M}_u^{-1}} \\
\leq &\left\|\sum_{v\in\mathcal{R}_{\hat{\gamma}}(u)}\sum_i\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\right\|_{\tilde{M}_u^{-1}} + \left\|\sum_{v\in\mathcal{W}_{\hat{\gamma}}(u)}\sum_i\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\right\|_{\tilde{M}_u^{-1}} \\
\leq &\frac{\zeta}{2}\sqrt{dN_{\mathcal{R}_{\hat{\gamma}}(u)}} + \frac{\hat{\gamma}}{2}\sqrt{dN_{\mathcal{W}_{\hat{\gamma}}(u)}}.
\end{aligned} \tag{35}$$

Therefore, by following the same procedure as in Lemma F.2, we establish Equation (33). Furthermore, by following the same procedure as in Appendix E.2, Theorem C.2 can be proved. □

F. Technical Lemmas

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

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

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

$$G_u(\boldsymbol{\theta}_s) := \sum_{i=1}^{N_u} \left(\sigma\left(\boldsymbol{\theta}_s^\top \boldsymbol{z}_u^i\right) - \sigma\left(\boldsymbol{\theta}_u^\top \boldsymbol{z}_u^i\right)\right)\boldsymbol{z}_u^i + \lambda\boldsymbol{\theta}_s.$$

By the mean value theorem, for any two parameter vectors $\boldsymbol{\theta}_{s_1}$ and $\boldsymbol{\theta}_{s_2}$, we have

$$G_u(\boldsymbol{\theta}_{s_1}) - G_u(\boldsymbol{\theta}_{s_2}) = \left(\sum_{i=1}^{N_u} \nabla\sigma\left(\boldsymbol{\theta}_{\bar{s}}^\top \boldsymbol{z}_u^i\right)\boldsymbol{z}_u^i(\boldsymbol{z}_u^i)^\top + \lambda I\right)(\boldsymbol{\theta}_{s_1} - \boldsymbol{\theta}_{s_2}) = W_u(\boldsymbol{\theta}_{s_1} - \boldsymbol{\theta}_{s_2}),$$

where we define

$$W_u := \sum_{i=1}^{N_u} \nabla\sigma\left(\boldsymbol{\theta}_{\bar{s}}^\top \boldsymbol{z}_u^i\right)\boldsymbol{z}_u^i(\boldsymbol{z}_u^i)^\top + \lambda I \quad \text{and} \quad \boldsymbol{\theta}_{\bar{s}} = \xi\boldsymbol{\theta}_{s_1} + (1-\xi)\boldsymbol{\theta}_{s_2},\ \xi \in [0,1].$$

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 $\boldsymbol{\theta}_{\bar{u}} = \xi_u\boldsymbol{\theta}_u + (1-\xi_u)\hat{\boldsymbol{\theta}}_u$.

Furthermore, we define

$$W_u := \sum_{i=1}^{N_u} \nabla\sigma\left(\boldsymbol{\theta}_{\bar{u}}^\top \boldsymbol{z}_u^i\right)\boldsymbol{z}_u^i(\boldsymbol{z}_u^i)^\top + \lambda I.$$

Recall that

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

By Equation (23), we have $W_u \succeq \kappa M_u$ and $M_u^{-1} \succeq \kappa W_u^{-1}$ since $\nabla\sigma(\boldsymbol{\theta}_{\bar{u}}^\top \boldsymbol{z}_u^i) \geq \kappa$. 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

$$\begin{aligned}
\left\|G_u(\hat{\boldsymbol{\theta}}_u) - \lambda\boldsymbol{\theta}_u\right\|_{M_u^{-1}}^2 &= \left\|G_u(\hat{\boldsymbol{\theta}}_u) - G_u(\boldsymbol{\theta}_u)\right\|_{M_u^{-1}}^2 = \left\|W_u(\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u)\right\|_{M_u^{-1}}^2 \\
&= (\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u)^\top W_u M_u^{-1} W_u(\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u) \\
&\overset{(a)}{\geq} \kappa\,(\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u)^\top W_u(\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u) \\
&\overset{(b)}{\geq} \kappa^2\,(\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u)^\top M_u(\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u) = \kappa^2\left\|\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u\right\|_{M_u}^2,
\end{aligned} \tag{36}$$

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

Moreover, observe that

$$\|\lambda\boldsymbol{\theta}_u\|_{M_u^{-1}} = \lambda\sqrt{\boldsymbol{\theta}_u^\top M_u^{-1}\boldsymbol{\theta}_u} \leq \sqrt{\lambda\kappa}\|\boldsymbol{\theta}_u\|_2 \leq \sqrt{\lambda\kappa}, \tag{37}$$

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

Combining these results, we have

$$\begin{aligned}\left\|\boldsymbol{\theta}_u - \hat{\boldsymbol{\theta}}_u\right\|_{M_u} &\overset{(a)}{\leq} \frac{1}{\kappa}\left\|G_u(\hat{\boldsymbol{\theta}}_u) - \lambda\boldsymbol{\theta}_u\right\|_{M_u^{-1}} \\ &\overset{(b)}{\leq} \frac{1}{\kappa}\left\|G_u(\hat{\boldsymbol{\theta}}_u)\right\|_{M_u^{-1}} + \frac{1}{\kappa}\|\lambda\boldsymbol{\theta}_u\|_{M_u^{-1}} \\ &\overset{(c)}{\leq} \frac{1}{\kappa}\left\|G_u(\hat{\boldsymbol{\theta}}_u)\right\|_{M_u^{-1}} + \sqrt{\frac{\lambda}{\kappa}},\end{aligned} \tag{38}$$

where (a) follows from (36), (b) uses the triangle inequality, and (c) applies (37).

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

$$\begin{aligned}\left\|G_u(\hat{\boldsymbol{\theta}}_u)\right\|_{M_u^{-1}} &= \left\|\sum_{i=1}^{N_u}\left(\sigma(\hat{\boldsymbol{\theta}}_u^\top \boldsymbol{z}_u^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_u^i)\right)\boldsymbol{z}_u^i + \lambda\hat{\boldsymbol{\theta}}_u\right\|_{M_u^{-1}} \\ &= \left\|\sum_{i=1}^{N_u}\left(\sigma(\hat{\boldsymbol{\theta}}_u^\top \boldsymbol{z}_u^i) - (y_u^i - \varepsilon_u^i)\right)\boldsymbol{z}_u^i + \sum_{i=1}^{N_u}\varepsilon_u^i \boldsymbol{z}_u^i + \lambda\hat{\boldsymbol{\theta}}_u\right\|_{M_u^{-1}} \\ &\overset{(a)}{\leq} \left\|\sum_{i=1}^{N_u}\varepsilon_u^i \boldsymbol{z}_u^i\right\|_{M_u^{-1}},\end{aligned} \tag{39}$$

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

$$\hat{\boldsymbol{\theta}}_u = \arg\min_{\boldsymbol{\theta}}\left[-\sum_{i=1}^{N_u}\left(y_u^i \log\sigma(\boldsymbol{\theta}^\top \boldsymbol{z}_u^i) + (1-y_u^i)\log\sigma(-\boldsymbol{\theta}^\top \boldsymbol{z}_u^i)\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right], \tag{40}$$

and thus its gradient satisfies

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

Therefore, it follows from (39) that

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

Next, let $V = \frac{\lambda}{\kappa} I$. Since ε_u^i is 2-subgaussian, we apply Theorem 1 in Abbasi-Yadkori et al. (2011) to obtain

$$\left\|\sum_{i=1}^{N_u}\varepsilon_u^i \boldsymbol{z}_u^i\right\|_{M_u^{-1}}^2 \leq 8\log\left(\frac{\det(M_u)^{1/2}}{\delta\,\det(V)^{1/2}}\right) \tag{41}$$

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

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

Putting everything together, we conclude that

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

which follows from combining (36), (38), (39), and (41). □

Lemma F.2 (Confidence Ellipsoid of $\tilde{\boldsymbol{\theta}}_u$)**.** *For any user u, under the data aggregation step of Algorithm 1 and the same conditions as in Lemma 3.2, it holds with probability at least $1-\delta$ that*

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

Proof. First, we define

$$\tilde{G}_u(\boldsymbol{\theta}_s) = \sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \left(\sigma\left(\boldsymbol{\theta}_s^\top \boldsymbol{z}_v^i\right) - \sigma\left(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i\right)\right) \boldsymbol{z}_v^i + \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\boldsymbol{\theta}_s, \quad \forall\, \boldsymbol{\theta}_s \in \mathbb{R}^d.$$

By the mean value theorem, for any $\boldsymbol{\theta}_{s_1}$ and $\boldsymbol{\theta}_{s_2}$, we have

$$\tilde{G}_u(\boldsymbol{\theta}_{s_1}) - \tilde{G}_u(\boldsymbol{\theta}_{s_2}) = \left(\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \nabla\sigma\left(\boldsymbol{\theta}_{\overline{s}}^\top \boldsymbol{z}_v^i\right) \boldsymbol{z}_v^i \boldsymbol{z}_v^{i\top} + \lambda|\mathcal{V}_{\hat{\gamma}}(u)|I\right)\left(\boldsymbol{\theta}_{s_1} - \boldsymbol{\theta}_{s_2}\right),$$

for some intermediate point $\boldsymbol{\theta}_{\overline{s}} = \xi\boldsymbol{\theta}_{s_1} + (1-\xi)\boldsymbol{\theta}_{s_2}$ 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 $\boldsymbol{\theta}_{\overline{u}} = \xi_u\boldsymbol{\theta}_u + (1-\xi_u)\tilde{\boldsymbol{\theta}}_u$.

We further define

$$\tilde{W}_u = \sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \nabla\sigma\left(\boldsymbol{\theta}_{\overline{u}}^\top \boldsymbol{z}_v^i\right) \boldsymbol{z}_v^i \boldsymbol{z}_v^{i\top} + \lambda|\mathcal{V}_{\hat{\gamma}}(u)|I \quad \text{and} \quad \tilde{M}_u = \sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \boldsymbol{z}_v^i \boldsymbol{z}_v^{i\top} + \frac{\lambda}{\kappa}|\mathcal{V}_{\hat{\gamma}}(u)|I.$$

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

$$\begin{aligned}
\left\|\tilde{G}_u(\tilde{\boldsymbol{\theta}}_u) - \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\boldsymbol{\theta}_u\right\|_{\tilde{M}_u^{-1}}^2 &= \left\|\tilde{G}_u(\tilde{\boldsymbol{\theta}}_u) - \tilde{G}_u(\boldsymbol{\theta}_u)\right\|_{\tilde{M}_u^{-1}}^2 = \left\|\tilde{W}_u(\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u)\right\|_{\tilde{M}_u^{-1}}^2 \\
&= (\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u)^\top \tilde{W}_u \tilde{M}_u^{-1} \tilde{W}_u (\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u) \\
&\overset{(a)}{\geq} \kappa\,(\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u)^\top \tilde{W}_u (\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u) \\
&\overset{(b)}{\geq} \kappa^2\,(\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u)^\top \tilde{M}_u (\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u) = \kappa^2 \left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u\right\|_{\tilde{M}_u}^2,
\end{aligned} \tag{42}$$

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 $\tilde{M}_u \succeq \frac{\lambda}{\kappa}|\mathcal{V}_{\hat{\gamma}}(u)|I$, we have

$$\begin{aligned}
\left\|\lambda|\mathcal{V}_{\hat{\gamma}}(u)|\boldsymbol{\theta}_u\right\|_{\tilde{M}_u^{-1}} &= \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\sqrt{\boldsymbol{\theta}_u^\top \tilde{M}_u^{-1}\boldsymbol{\theta}_u} \leq \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\sqrt{\boldsymbol{\theta}_u^\top \left(\frac{\kappa}{\lambda|\mathcal{V}_{\hat{\gamma}}(u)|}I\right)\boldsymbol{\theta}_u} \\
&= \sqrt{\lambda|\mathcal{V}_{\hat{\gamma}}(u)|\kappa}\,\left\|\boldsymbol{\theta}_u\right\|_2 \leq \sqrt{\lambda|\mathcal{V}_{\hat{\gamma}}(u)|\kappa}.
\end{aligned} \tag{43}$$

Hence, we obtain

$$\left\|\boldsymbol{\theta}_u - \tilde{\boldsymbol{\theta}}_u\right\|_{\tilde{M}_u} \overset{(a)}{\leq} \frac{1}{\kappa}\left\|\tilde{G}_u(\tilde{\boldsymbol{\theta}}_u) - \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\boldsymbol{\theta}_u\right\|_{\tilde{M}_u^{-1}}$$

$$\overset{(b)}{\leq} \frac{1}{\kappa}\left\|\tilde{G}_u(\tilde{\boldsymbol{\theta}}_u)\right\|_{\tilde{M}_u^{-1}} + \frac{1}{\kappa}\left\|\lambda|\mathcal{V}_{\hat{\gamma}}(u)|\boldsymbol{\theta}_u\right\|_{\tilde{M}_u^{-1}}$$
$$\overset{(c)}{\leq} \frac{1}{\kappa}\left\|\tilde{G}_u(\tilde{\boldsymbol{\theta}}_u)\right\|_{\tilde{M}_u^{-1}} + \sqrt{\frac{\lambda|\mathcal{V}_{\hat{\gamma}}(u)|}{\kappa}}, \tag{44}$$

where (a) follows from Equation (42), (b) applies the triangle inequality, and (c) uses the bound in Equation (43).

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

$$\begin{aligned}
&\frac{1}{\kappa^2}\left\|\tilde{G}_u(\tilde{\boldsymbol{\theta}}_u)\right\|^2_{\tilde{M}_u^{-1}} \\
&\overset{(a)}{=} \frac{1}{\kappa^2}\Big\|\sum_{v\in\mathcal{V}_{\hat{\gamma}}(u)}\sum_{i=1}^{N_v}\left(\sigma(\tilde{\boldsymbol{\theta}}_u^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i + \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\tilde{\boldsymbol{\theta}}_u\Big\|^2_{\tilde{M}_u^{-1}} \\
&= \frac{1}{\kappa^2}\Big\|\sum_v\sum_i\left(\sigma(\tilde{\boldsymbol{\theta}}_u^\top \boldsymbol{z}_v^i) - y_v^i + y_v^i - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i + \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\tilde{\boldsymbol{\theta}}_u\Big\|^2_{\tilde{M}_u^{-1}} \\
&= \frac{1}{\kappa^2}\Big\|\sum_v\sum_i\left(\sigma(\tilde{\boldsymbol{\theta}}_u^\top \boldsymbol{z}_v^i) - y_v^i\right)\boldsymbol{z}_v^i + \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\tilde{\boldsymbol{\theta}}_u + \sum_v\sum_i\left(y_v^i - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\Big\|^2_{\tilde{M}_u^{-1}} \\
&\overset{(b)}{=} \frac{1}{\kappa^2}\Big\|\sum_v\sum_i\left(y_v^i - \sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) + \sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\Big\|^2_{\tilde{M}_u^{-1}} \\
&= \frac{1}{\kappa^2}\Big\|\underbrace{\sum_v\sum_i\varepsilon_v^i\boldsymbol{z}_v^i}_{\text{noise}} + \underbrace{\sum_v\sum_i\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i}_{\text{bias}}\Big\|^2_{\tilde{M}_u^{-1}} \\
&\overset{(c)}{\leq} \left(\frac{1}{\kappa}\Big\|\sum_v\sum_i\varepsilon_v^i\boldsymbol{z}_v^i\Big\|_{\tilde{M}_u^{-1}} + \frac{1}{\kappa}\Big\|\sum_v\sum_i\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\Big\|_{\tilde{M}_u^{-1}}\right)^2 \\
&\overset{(d)}{=} \left(\frac{1}{\kappa}\Big\|\sum_v\sum_i\varepsilon_v^i\boldsymbol{z}_v^i\Big\|_{\tilde{M}_u^{-1}} + \frac{1}{\kappa}\Big\|\sum_{v\in\mathcal{W}_{\hat{\gamma}}(u)}\sum_i\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\Big\|_{\tilde{M}_u^{-1}}\right)^2.
\end{aligned} \tag{45}$$

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

$$\sum_v\sum_i\left(\sigma(\tilde{\boldsymbol{\theta}}_u^\top \boldsymbol{z}_v^i) - y_v^i\right)\boldsymbol{z}_v^i + \lambda|\mathcal{V}_{\hat{\gamma}}(u)|\tilde{\boldsymbol{\theta}}_u = 0;$$

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

Next, we bound the term

$$\Big\|\sum_{v\in\mathcal{W}_{\hat{\gamma}}(u)}\sum_{i=1}^{N_v}\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\Big\|_{\tilde{M}_u^{-1}}.$$

By the triangle inequality, we have

$$\begin{aligned}
&\Big\|\sum_{v\in\mathcal{W}_{\hat{\gamma}}(u)}\sum_{i=1}^{N_v}\left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right)\boldsymbol{z}_v^i\Big\|_{\tilde{M}_u^{-1}} \\
&\leq \sum_{v\in\mathcal{W}_{\hat{\gamma}}(u)}\sum_{i=1}^{N_v}\left|\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right|\left\|\boldsymbol{z}_v^i\right\|_{\tilde{M}_u^{-1}} \\
&\overset{(a)}{\leq} \sum_v\sum_i\frac{1}{4}\left|\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i - \boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i\right|\left\|\boldsymbol{z}_v^i\right\|_{\tilde{M}_u^{-1}} \leq \frac{\hat{\gamma}}{4}\sum_v\sum_i\|\boldsymbol{z}_v^i\|_2\|\boldsymbol{z}_v^i\|_{\tilde{M}_u^{-1}}
\end{aligned}$$

$$\overset{(b)}{\leq} \frac{\hat{\gamma}}{2} \sum_{v \in \mathcal{W}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \|\boldsymbol{z}_v^i\|_{\tilde{M}_u^{-1}}, \tag{46}$$

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

Furthermore, observe that

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

By applying Cauchy–Schwarz inequality, we get

$$\sum_v \sum_i \|\boldsymbol{z}_v^i\|_{\tilde{M}_u^{-1}} \leq \sqrt{\left(\sum_v N_v\right)\left(\sum_v \sum_i \|\boldsymbol{z}_v^i\|_{\tilde{M}_u^{-1}}^2\right)} \leq \sqrt{d \cdot N_{\mathcal{W}_{\hat{\gamma}}(u)}}. \tag{47}$$

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

Therefore, by applying Equation (46) and (47), we obtain

$$\Big\| \sum_{v \in \mathcal{W}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \left(\sigma(\boldsymbol{\theta}_v^\top \boldsymbol{z}_v^i) - \sigma(\boldsymbol{\theta}_u^\top \boldsymbol{z}_v^i)\right) \boldsymbol{z}_v^i \Big\|_{\tilde{M}_u^{-1}} \leq \frac{\hat{\gamma}}{2} \sqrt{d\, N_{\mathcal{W}_{\hat{\gamma}}(u)}}, \tag{48}$$

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

Furthermore, for the noise term in Equation (45), by applying Theorem 1 in Abbasi-Yadkori et al. (2011) with $V = \frac{\lambda}{\kappa} I$, we have

$$\Big\| \sum_{v \in \mathcal{V}_{\hat{\gamma}}(u)} \sum_{i=1}^{N_v} \varepsilon_v^i \boldsymbol{z}_v^i \Big\|_{\tilde{M}_u^{-1}} \leq 2\sqrt{2\log\left(\frac{\det(\tilde{M}_u)^{1/2}}{\delta\, \det(V)^{1/2}}\right)} \leq 2\sqrt{2\log\left(\frac{1}{\delta}\right) + d\,\log\left(1 + \frac{4\, N_{\mathcal{V}_{\hat{\gamma}}(u)}\kappa}{d\lambda|\mathcal{V}_{\hat{\gamma}}(u)|}\right)} \tag{49}$$

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

Combining Equation (44), Equation (45), Equation (48), and (49), we finally have

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

which holds for all $u \in \mathcal{U}$ with probability at least $1-\delta$. This completes the proof of Lemma F.2. □

Lemma F.3 (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*

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

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

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

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

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

Proof. Lemma F.3 comes from Theorem 2 in Abbasi-Yadkori et al. (2011). □

Lemma F.4 (Elliptic Potential Lemma)**.** *Let* $\{\boldsymbol{z}_s\}_{s=1}^n$ *be a sequence of vectors in* $\mathbb{R}^d$ *such that* $\|\boldsymbol{z}_s\| \leq L$ *for any* $s \in [t]$*. Let* $V_t = \sum_{s=1}^{t-1} \boldsymbol{z}_s \boldsymbol{z}_s^\top + \lambda I$*. Then,*

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

Proof. Lemma F.4 comes from Lemma C.2 in Das et al. (2024). □

Lemma F.5 (Lower Bound on the Minimum Eigenvalue)**.** *Let* $\boldsymbol{a}_s, n \geq 1$ *be generated sequentially from a random distribution such that* $||\boldsymbol{a}||_2 \leq 1$ *and* $\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top]$ *is full rank with minimal eigenvalue* $\lambda_a > 0$*. Let* $M_n = \sum_{s=1}^n \boldsymbol{a}_s \boldsymbol{a}_s^\top$*. 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,*

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

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

Proof. Lemma F.5 comes from Lemma 7 in Li & Zhang (2018) and Lemma B.2 in Wang et al. (2025). □

Lemma F.6 (One-step Update on the Euclidean Unit Ball)**.** *Let* $M \in \mathbb{R}^{d \times d}$ *be symmetric positive semidefinite with eigenvalues* $\lambda_1(M) \leq \lambda_2(M) \leq \cdots \leq \lambda_d(M)$*, and corresponding orthonormal eigenvectors* $q_1, \ldots, q_d$*. Let*

$$z^* := \arg\max_{\|z\|_2 \leq 1} \ z^\top M^{-1} z, \tag{50}$$

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

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

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

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

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

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

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 $\lambda_1 + 1,\ \lambda_2, \ldots, \lambda_d$, 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 F.7 (Multi-step Update on the Euclidean Unit Ball)**.** *Let $M \in^{d \times d}$ 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 $s \in \{1, 2, \cdots, d-1\}$ such that

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

Perform s greedy rank-one updates

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

Then

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

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 F.6, 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

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

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 $\lambda_1\big(M^{(k)}\big)$. It follows that

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

as claimed. □

G. Experiments

G.1. Baselines

We compare Off-C^{2}PL with both enhanced versions of traditional clustering algorithms and prior methods for contextual logistic bandits. Specifically, we adapt classical clustering algorithms such as KMeans (McQueen, 1967) (with $\sqrt{\#\text{ of users}}$ as cluster number) and DBSCAN (Schubert et al., 2017) 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 (Zhu et al., 2023) 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 $\boldsymbol{\theta}$. For evaluating ADA-Off-C^{2}PL, we compare against the pure offline algorithm Off-C^{2}PL trained on randomly generated offline samples and the pure active learning algorithm Active Preference Optimization (APO) from Das et al. (2024) that operates without any offline data.

G.2. Datasets

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 $\boldsymbol{\theta}^j \in \mathbb{R}^d$ with $d = 768$, matching the dimensionality of the real-world embeddings used in our experiments. For a user u in cluster c, we set $\boldsymbol{\theta}_u = \boldsymbol{\theta}^j + \boldsymbol{\epsilon}_u$, where $\boldsymbol{\epsilon}_u \sim \mathcal{N}(0, s^2 I_d)$. 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 $\boldsymbol{z} \sim \mathcal{N}(0, I_d)$, the preferred item is sampled with probability $\sigma\left(\beta\, \boldsymbol{\theta}_u \cdot \boldsymbol{z}\right)$, 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 (Völske et al., 2017) alongside human preferences collected by Stiennon et al. (2020). 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.