DIRECTIONS THAT MATTER: EFFICIENT ONLINE RLHF VIA PREFERENCE GEOMETRY SKETCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Online Reinforcement Learning from Human Feedback (RLHF) enables models to continually collect preference feedback and refine their reward models during interaction. However, scaling existing online RLHF methods to large language models (LLMs) remains computationally challenging. Their one-pass estimators rely on high-dimensional local curvature to preserve statistical efficiency, leading to substantial computation and memory costs that grow at least quadratically with the feature dimension. We address this challenge by proposing a geometry-sketched reward modeling framework that reduces the dependence on the feature dimension to linear. The key idea is to approximate the geometry of accumulated preference data using *matrix sketching*: we retain the dominant directions of preference geometry without explicitly maintaining the full curvature matrices. We establish theoretical efficiency–utility trade-offs for representative online RLHF settings with different data-collection strategies and forms of comparison feedback. Extensive experiments show that our method provides an effective efficiency–utility trade-off, consistently improves over existing approximation baselines, and scales to larger LLMs with limited additional computational overhead.

1 INTRODUCTION

Preference-based reinforcement learning provides a general framework for learning complex objectives from comparative feedback. It has been applied to robotic control (Christiano et al., 2017), recommender systems (Xue et al., 2023), and healthcare (Barata et al., 2023). Its most prominent recent application is reinforcement learning from human feedback (RLHF), which enables large language models to learn from preference comparisons when explicit scalar rewards are unavailable or difficult to specify (Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). RLHF has consequently become an important component of LLM training, helping align model behavior with human preferences and intended objectives (OpenAI, 2023; Llama Team, 2023; Qwen Team, 2024).

Traditional RLHF typically relies on static preference datasets, which limits its ability to adapt to changing user preferences and deployment environments. Online RLHF addresses this limitation by iteratively collecting new preference feedback and updating the model, enabling continual alignment beyond fixed datasets (Zhu et al., 2023; Dong et al., 2024; Cen et al., 2025; Xie et al., 2025; Das et al., 2025). This capability extends LLM alignment from offline post-training to deployment-time adaptation, making online RLHF particularly relevant for agentic systems where models must incorporate environmental feedback while interacting, reasoning, and making decisions (Suhr & Artzi, 2023; Silver & Sutton, 2025). It is also closely related to test-time learning (Hardt & Sun, 2024) and provides a potential pathway toward personalized LLMs that adapt to individual user preferences through continual interaction (Qu et al., 2026). From a theoretical perspective, a central challenge in online RLHF is to achieve both computational and statistical efficiency. Under pairwise-comparison feedback, Li et al. (2025) proposed a *one-pass* reward-modeling framework based on an online mirror descent (OMD) estimator. This one-pass design is well-suited to online RLHF because it avoids repeatedly re-optimizing the reward model over an ever-growing preference dataset. Lee et al. (2025) further extended this framework to multi-comparison feedback, demonstrating that additional comparisons provide richer preference information and lead to improved sample efficiency.

Motivation. Despite achieving one-pass processing, existing online RLHF methods can still incur substantial per-round computational and memory costs. Their statistical guarantees rely on OMD

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

estimators defined with carefully designed local norms (Zhang & Sugiyama, 2023; Li et al., 2025; Lee et al., 2025). Implementing these estimators requires $O(d^3)$ time and $O(d^2)$ space per round, where d denotes the feature dimension of the reward model. In LLM-based reward modeling, such features are typically derived from hidden-state representations, whose dimension is large and tends to increase with model size. For example, it reaches $d = 3584$ for `Qwen2.5-7B` and $d = 5120$ for `Qwen2.5-32B` (Qwen Team, 2024). These dimension-dependent costs present a major obstacle to scaling theoretically guaranteed online RLHF methods to larger LLMs.

To alleviate this computational burden, practical implementations often resort to iterative approximation methods such as conjugate gradient (CG) (Pearlmutter, 1994; Li et al., 2025). However, these approximation-based implementations typically rely on empirically chosen approximation parameters and lack theoretical control over the induced preference-geometry approximation error. Their performance depends on empirically chosen approximation parameters, and existing analyses do not characterize how such errors affect the statistical guarantees of online RLHF. Moreover, these approximations are not readily extended to more general feedback settings such as multi-comparison feedback. Therefore, it remains unclear whether one can achieve scalable computation while preserving the statistical efficiency of online RLHF. This gap motivates the following question:

Can we reduce the dimension-dependent complexity of online RLHF while preserving statistical efficiency under general comparison feedback?

Contributions. We answer this question affirmatively with GeoSRM, a geometry-sketched reward modeling framework for online RLHF. GeoSRM exploits the geometry of accumulated preference data by sketching the local curvature and retaining its dominant directions, thereby reducing complexity while preserving statistical guarantees. Our main contributions are summarized below.

- Efficient one-pass reward modeling via sketching. We extend matrix sketching techniques from streaming algorithms (Liberty, 2013; Ghashami et al., 2016) into online reward modeling and derive approximation guarantees for the accumulated preference-curvature matrix. Our analysis establishes how the resulting sketching error propagates to the online estimation error. It shows that retaining the dominant curvature directions can reduce the computational and memory costs of online RLHF while maintaining provable statistical efficiency.
- Provably efficient online RLHF under general comparison feedback. We instantiate GeoSRM in two representative settings: active preference optimization and deployment-time adaptation. The deployment setting considers pairwise-comparison feedback, whereas the active setting supports the more general multi-comparison feedback. We derive a unified approximate online estimator for both settings and establish corresponding statistical guarantees. We further reduces the computational cost of each action selection from $O(d^2)$ to $O(d)$ by exploiting the low-rank structure.
- Empirical studies. We empirically show that preference curvature exhibits strong spectral concentration in realistic LLM representations, supporting the effectiveness of sketch-based approximation. Through extensive evaluations across LLMs of different sizes, we demonstrate that our method achieves a favorable efficiency-utility trade-off and scales up effectively to larger LLMs.

2 Preliminaries and Problem Setup

In this section, we introduce the necessary preliminaries and problem setup used throughout.

2.1 Online RLHF

Following recent works (Zhu et al., 2023; Xiong et al., 2024; Lee et al., 2025), we formulate online RLHF as a contextual bandit problem. Let $\mathcal{X}$ denote a context space and $\mathcal{A} := \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_N\}$ denote a finite set of candidate actions. At each interaction, the learner constructs a query $(\boldsymbol{x}, S)$, where $\boldsymbol{x} \in \mathcal{X}$ is a context and $S \subseteq \mathcal{A}$ contains several actions, and presents S to a human labeler. The labeler returns a ranking $\omega = (\omega_1, \dots, \omega_{|S|})$, which is a permutation of the actions in S, where ω_i denotes the i-th most preferred action. The Plackett–Luce (PL) model (Plackett, 1975; Luce, 2012) defines the ranking probability for $2 \le |S| \le K$ using a latent reward function r as follows

$$\mathbb{P}(\omega \mid \boldsymbol{x}, S) = \prod_{i=1}^{|S|} \frac{\exp(r(\boldsymbol{x}, \omega_i))}{\sum_{j=i}^{|S|} \exp(r(\boldsymbol{x}, \omega_j))} \quad \text{where} \quad r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}.$$

When $K = 2$, it reduces to the pairwise comparison considered in the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952). Following prior works on the theoretical analysis of RLHF (Zhu et al., 2023; Cen et al., 2025), we further assume that the latent reward function is linearly realizable:

Assumption 1 (Linear realizability)**.** *There exists an unknown parameter $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ such that $r(\boldsymbol{x}, \boldsymbol{a}) = \phi(\boldsymbol{x}, \boldsymbol{a})^\top \boldsymbol{\theta}^\star$, where $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a known and fixed feature map. We assume that $\|\phi(\boldsymbol{x}, \boldsymbol{a})\|_2 \le 1$ for all $(\boldsymbol{x}, \boldsymbol{a}) \in \mathcal{X} \times \mathcal{A}$ and that $\boldsymbol{\theta}^\star \in \Theta := \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta}\|_2 \le B\}$.*

This assumption provides a standard and analytically tractable model for reward learning. In LLM-based reward modeling, ϕ can be interpreted as a fixed representation produced by a pretrained LLM, while $\boldsymbol{\theta}^\star$ corresponds to the weights of a linear reward head. Our analysis focuses on the realizable linear setting, while extensions to model misspecification (Jin et al., 2023) and neural function approximation (Du et al., 2024; Verma et al., 2025a) have been studied in prior work.

General Pipeline. Online RLHF alternates between collecting preference feedback and updating the model (Dong et al., 2024; Li et al., 2025). At iteration t, the general pipeline proceeds as follows:

Step 1. *Data collection*: Observe a prompt $\boldsymbol{x}_t$ and select an assortment of responses S_t, query the preference oracle for ranking feedback ω_t, expand the dataset $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(\boldsymbol{x}_t, S_t, \omega_t)\}$.

Step 2. *Online reward modeling*: Train the reward model using augmented preference data $\mathcal{D}_{t+1}$.

Step 3. *Policy optimization*: Updates the policy from π_t to π_{t+1} using the current reward model.

The two online RLHF settings studied in this paper instantiate the above pipeline differently (see Section 4), primarily in how the query $(\boldsymbol{x}_t, S_t)$ is constructed and how the policy model is updated.

2.2 Online Parameter Estimation

Traditional online RLHF methods estimate the reward model in *Step 2*. using maximum likelihood estimation (MLE) over all preference data collected (Saha, 2021; Zhu et al., 2023; Das et al., 2025). As the preference dataset grows, repeatedly solving the MLE problem requires iterative optimization, leading to a per-round computational cost that increases with the number of interactions. Recent work avoids this repeated re-optimization by replacing MLE with an online mirror descent estimator (Li et al., 2025; Lee et al., 2025). The resulting OMD estimator processes each new observation once, so its per-round cost does not depend on the size of the accumulated dataset.

More precisely, for multi-comparison feedback, the ranking ω_t is first converted into $q_t = \binom{|S_t|}{2}$ pairwise observations using rank-breaking (RB) (Ouyang et al., 2022; Lee et al., 2025). For each $p \in [q_t]$, let $\boldsymbol{a}^+$ and $\boldsymbol{a}^-$ denote the higher and lower-ranked actions, and define the signed feature difference $\boldsymbol{z}_{t,p} := \phi(\boldsymbol{x}_t, \boldsymbol{a}^+) - \phi(\boldsymbol{x}_t, \boldsymbol{a}^-)$. The pairwise logistic loss, gradients, and Hessians are

$$\ell_{t,p}(\boldsymbol{\theta}) = -\log \mu(\boldsymbol{z}_{t,p}^\top \boldsymbol{\theta}), \;\; g_{t,p}(\boldsymbol{\theta}) = (\mu(\boldsymbol{z}_{t,p}^\top \boldsymbol{\theta}) - 1)\boldsymbol{z}_{t,p}, \;\; \boldsymbol{H}_{t,p}(\boldsymbol{\theta}) = \dot{\mu}(\boldsymbol{z}_{t,p}^\top \boldsymbol{\theta})\boldsymbol{z}_{t,p}\boldsymbol{z}_{t,p}^\top, \tag{1}$$

where $\dot{\mu}(w) = \mu(w)(1 - \mu(w))$ and $\mu(w) = 1/(1 + \exp(-w))$ is the sigmoid function. Let $\widetilde{\boldsymbol{\theta}}_{t,0} := \widetilde{\boldsymbol{\theta}}_t$ denote the estimator at the beginning of round t. Given a step size $\eta > 0$, the OMD estimator is updated sequentially for each comparison $p \in [q_t]$ as

$$\widetilde{\boldsymbol{\theta}}_{t,p} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \left\{ \left\langle g_{t,p}(\widetilde{\boldsymbol{\theta}}_{t,p-1}), \boldsymbol{\theta} \right\rangle + \frac{1}{2\eta} \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_{t,p-1} \right\|_{\widetilde{\boldsymbol{\mathcal{H}}}_{t,p}}^2 \right\}, \tag{2}$$

where $\widetilde{\boldsymbol{\mathcal{H}}}_{t,p} = \boldsymbol{\mathcal{H}}_t + \eta \sum_{j=1}^{p} \boldsymbol{H}_{t,j}(\widetilde{\boldsymbol{\theta}}_{t,j-1})$ is the curvature-aware local norm after incorporating the first p comparisons in round t, and $\boldsymbol{\mathcal{H}}_t = \lambda \boldsymbol{I}_d + \sum_{i=1}^{t-1} \sum_{j=1}^{q_i} \boldsymbol{H}_{i,j}(\widetilde{\boldsymbol{\theta}}_{i,j})$ collects the curvature accumulated before round t. After processing all q_t comparisons, we set $\widetilde{\boldsymbol{\theta}}_{t+1} := \widetilde{\boldsymbol{\theta}}_{t,q_t}$ and pairwise feedback can be recovered as the special case $q_t = 1$. Eq. (2) can be expressed as a projection under the local norm. It processes each pairwise observation once, without revisiting the historical data. Related local-norm OMD estimators have also been studied for logistic bandits (Faury et al., 2022; Zhang et al., 2025) and multinomial logit bandits (Zhang & Sugiyama, 2023; Lee & Oh, 2024).

Although the OMD estimator avoids revisiting the full preference history during each update, its computational and memory costs remain substantial. Specifically, the optimization problem in Eq. (2) can be implemented as a projected gradient step and requires $O(d^3)$ time (Mhammedi et al.,

2019). Since RB produces $q_t = O(K^2)$ pairwise comparisons, the resulting per-round computational cost is $O(K^2d^3)$. In LLM applications, d is typically much larger than K, making the dependence on d the dominant computational bottleneck. The curvature matrices can be maintained incrementally, but explicitly storing them still requires $O(d^2)$ space. These quadratic dependencies on the feature dimension limit the scalability of theoretically guaranteed online RLHF to LLMs.

2.3 Matrix sketching

Matrix sketching provides compact summaries of large matrices while approximately preserving their action on vectors and other matrices (Cormode & Yi, 2020). In this paper, we focus on Frequent Directions (FD) (Liberty, 2013), a deterministic algorithm designed for streaming matrices. Given a streaming matrix $\boldsymbol{X}_t = [\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_t^\top]^\top \in \mathbb{R}^{t\times d}$, FD maintains a sketch $\boldsymbol{B}_t \in \mathbb{R}^{l\times d}$ with size $l \ll d$. To process row $\boldsymbol{x}_t$, we first replace the last row of $\boldsymbol{B}_t$ with $\boldsymbol{x}_t$. Then, FD performs SVD on $\boldsymbol{B}_t$, i.e., $\boldsymbol{B}_t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t$. Let $\boldsymbol{\Sigma}_t = \mathrm{diag}(\sigma_1, \dots, \sigma_d)$, where σ_l is the l-th largest singular value. Subsequently, FD sets $\widehat{\boldsymbol{\Sigma}}_{t+1} = \mathrm{diag}(\sqrt{\sigma_1^2 - \sigma_l^2}, \dots, \sqrt{\sigma_l^2 - \sigma_l^2})$ and $\boldsymbol{B}_{t+1} = \widehat{\boldsymbol{\Sigma}}_{t+1}\boldsymbol{V}_t$.

The complete procedure is provided in Appendix B.1. Implementing FD requires $O(dl)$ space and $O(dl)$ amortized time per update. It has been used to reduce the computational and memory costs of several large-scale learning problems (Luo et al., 2016; Feinberg et al., 2023; Yang et al., 2025; Wen et al., 2026b). But directly applying FD to online RLHF is nontrivial: its guarantee controls the matrix approximation error, whereas the online estimator requires control of the estimation error induced by approximating accumulated local curvature. Establishing this connection thus requires a sketching construction that preserves the preference geometry relevant to the online estimator.

3 Algorithms

In this section, we propose a geometry-sketched reward modeling framework for online RLHF. We theoretically proves that our method possesses both statistical and computational efficiency.

3.1 Preference Geometry Sketching

The main computational bottleneck of existing online RLHF arises from explicitly maintaining and operating on the local curvature matrix $\widetilde{\boldsymbol{\mathcal{H}}}_{t,p} \in \mathbb{R}^{d\times d}$ in Eq. (2). Our goal is therefore to approximate this curvature without materializing the full matrix. Prior work suggests that high-dimensional learning and preference representations often admit useful low-dimensional structure (Marteau-Ferey et al., 2019; Luo et al., 2025). This motivates a geometry-aware compression of the local curvature: rather than preserving the full matrix, we retain its dominant directions through matrix sketching.

Selective Sketching. Directly applying previous covariance sketching (Luo et al., 2016; Kuzborskij et al., 2019), however, does not exploit the structure of OMD. Recall that $\widetilde{\boldsymbol{\mathcal{H}}}_{t,p}$ consists of two parts: the look-ahead curvature accumulated from previous rounds and a second-order correction generated by the comparisons in the current round. We therefore sketch only the historical curvature while maintaining the current-round correction exactly. Specifically, we maintain a FD sketch $\boldsymbol{B}_t \in \mathbb{R}^{l\times d}$ such that $\boldsymbol{B}_t^\top \boldsymbol{B}_t$ approximates the accumulated historical $\boldsymbol{\mathcal{H}}_t$ along the estimate, and define

$$\widehat{\boldsymbol{\mathcal{H}}}_{t,p} \;=\; \underbrace{\lambda \boldsymbol{I}_d + \boldsymbol{B}_t^\top \boldsymbol{B}_t}_{\text{historical curvature sketch}} \;+\; \underbrace{\eta \sum\nolimits_{j=1}^{p} \boldsymbol{H}_{t,j}\left(\widehat{\boldsymbol{\theta}}_{t,j-1}\right)}_{\text{second-order correction}}. \tag{3}$$

The first term provides an approximation to the historical curvature, whereas the second term retains the curvature from the first p comparisons. Since each $\boldsymbol{H}_{t,j}(\widehat{\boldsymbol{\theta}}_{t,j-1})$ is rank one, the current-round correction can be updated exactly as new comparisons arrive. The following proposition shows that the approximation error is controlled by the spectral tail of the accumulated historical Hessians.

Proposition 1 (Claim 1 of Liberty (2013))**.** *Let $\boldsymbol{A}_T$ be the matrix whose rows are $\sqrt{\dot{\mu}(\boldsymbol{z}_{i,j}^\top \widehat{\boldsymbol{\theta}}_{i,j})}\, \boldsymbol{z}_{i,j}^\top$, so that $\boldsymbol{A}_T^\top \boldsymbol{A}_T = \sum_{i=1}^{T-1}\sum_{j=1}^{q_i} \boldsymbol{H}_{i,j}(\widehat{\boldsymbol{\theta}}_{i,j})$. Let $\boldsymbol{A}_{T,[k]}$ denote best rank-k approximation, we have*

$$\left\|\boldsymbol{A}_T^\top \boldsymbol{A}_T - \boldsymbol{B}_T^\top \boldsymbol{B}_T\right\|_2 \le \min_{0\le k<l} \frac{\left\|\boldsymbol{A}_T - \boldsymbol{A}_{T,[k]}\right\|_F^2}{l-k} := \Delta_l. \tag{4}$$

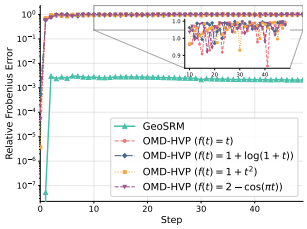


Figure 1: Different Strategy.

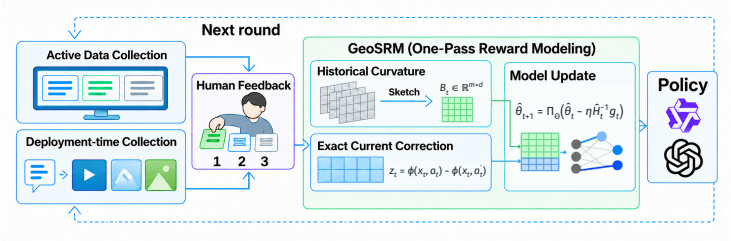


Figure 2: Geometry-sketched modeling framework for online RLHF.

Comparison with Heuristic Approximation. Previous work (Li et al., 2025) makes online RLHF computationally feasible for LLMs through Hessian-vector products (HVP). In its practical implementation, the historical curvature $\boldsymbol{\mathcal{H}}_t$ is further approximated by a time-varying isotropic matrix $\lambda_t \boldsymbol{I}_d$, where $\lambda_t = \lambda_0 \min\{1, f(t/T)\}$ and $f(\cdot)$ is a monotonically increasing function. This isotropic approximation compresses the historical curvature into a single scalar and therefore discards its directional structure. Figure 1 compares its approximation error with that of FD sketching (green line) under several commonly used choices of f. The relative Frobenius error of our method is at least two orders of magnitude smaller than that of the isotropic approximations. This result shows the benefit of preserving the dominant directions of the preference geometry. Section 5 further provides empirical evidence that the spectral tail term Δ_l is relatively small in practice.

3.2 Geometry-Sketched Reward Modeling Framework

Building on the preference geometry sketching, we now present GeoSRM, a Geometry-Sketched Reward Modeling framework for efficient online RLHF. Following the general pipeline in Figure 2, we focuses on making reward modeling both one-pass and efficient in the feature dimension d.

Data Structure. At round t, GeoSRM receives a new preference observation $(\boldsymbol{x}_t, S_t, \omega_t)$, where S_t contains at most K actions and ω_t denotes their observed preference ranking. The construction of the query $(\boldsymbol{x}_t, S_t)$ depends on the specific online RLHF setting and will be discussed in Section 4. We apply RB to (S_t, ω_t) to obtain q_t pairwise comparisons, indexed according to the ranking order.

Sketched Estimator. Following the sequential rank-breaking update of Lee et al. (2025), we maintain a sketched OMD estimator that processes the q_t pairwise comparisons sequentially at round t. The key difference is that we replace the exact historical curvature with its FD sketch, while retaining the current-round correction exactly as in Eq. (3). This preserves the sequential local-norm OMD structure while using a compact representation of the historical preference geometry. Specifically, for each $p \in [q_t]$ at round t, we update the estimator as follows

$$\widehat{\boldsymbol{\theta}}_{t,p} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\{ \left\langle g_{t,p}\left(\widehat{\boldsymbol{\theta}}_{t,p-1}\right), \boldsymbol{\theta} \right\rangle + \frac{1}{2\eta} \left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{t,p-1} \right\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_{t,p}} \right\}, \tag{5}$$

where $\eta > 0$ is the step size and $\widehat{\boldsymbol{\mathcal{H}}}_{t,p}$ is defined in Eq. (3). As with standard OMD updates (Orabona, 2019), Eq. (5) can be equivalently written as a preconditioned step followed by a projection,

$$\boldsymbol{\xi}_{t,p} = \widehat{\boldsymbol{\theta}}_{t,p-1} - \eta \widehat{\boldsymbol{\mathcal{H}}}^{-1}_{t,p} \cdot g_{t,p}\left(\widehat{\boldsymbol{\theta}}_{t,p-1}\right) \quad \text{and} \quad \widehat{\boldsymbol{\theta}}_{t,p} = \arg\min_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\theta} - \boldsymbol{\xi}_{t,p}\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_{t,p}}. \tag{6}$$

A direct implementation of Eq. (6) would require operating on the $d \times d$ matrix $\widehat{\boldsymbol{\mathcal{H}}}^{-1}_{t,p}$, which would eliminate the computational benefit of sketching. We therefore exploit the low-rank structure of Eq. (3) and avoid explicitly forming this inverse. For the historical curvature, we maintain $\boldsymbol{B}_t \in \mathbb{R}^{l\times d}$ together with $\boldsymbol{M}_t := (\boldsymbol{B}_t\boldsymbol{B}_t^\top + \lambda \boldsymbol{I}_l)^{-1} \in \mathbb{R}^{l\times l}$. The required inverse operations can then be reduced to low-dimensional ones using the following Woodbury identity (Woodbury, 1950).

Fact 1 (Woodbury Identity). *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be invertible and $\boldsymbol{U} \in \mathbb{R}^{d\times r}$. Then*

$$\left(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{U}^\top\right)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\left(\boldsymbol{I}_r + \boldsymbol{U}^\top\boldsymbol{A}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^\top\boldsymbol{A}^{-1},$$

where the case $r = 1$ recovers the Sherman–Morrison update (Sherman & Morrison, 1950). Applying Fact 1 to the inverse of historical curvature sketch in Eq. (3) gives

$$\left(\lambda \boldsymbol{I}_d + \boldsymbol{B}_t^\top \boldsymbol{B}_t\right)^{-1} = \frac{1}{\lambda}\left(\boldsymbol{I}_d - \boldsymbol{B}_t^\top \boldsymbol{M}_t \boldsymbol{B}_t\right). \tag{7}$$

Algorithm 1 GeoSRM: **Geo**metry-**S**ketched **R**eward **M**odeling Framework

Input: Sketch size l, step size η, regularization parameter λ, maximum assortment size K

Output: Reward model $\widehat{\boldsymbol{\theta}}_t$ and sketch matrix $\boldsymbol{B}_t$, $\boldsymbol{M}_t$ at round t

1: Initialize estimator $\widehat{\boldsymbol{\theta}}_1 \in \Theta$ and sketch matrices $\boldsymbol{B}_1 = \mathbf{0}_{l\times d}$ and $\boldsymbol{M}_1 = \frac{1}{\lambda}\boldsymbol{I}_l$.
2: **for** $t = 1$ **to** T **do**
3: Observe preference data $\{(\boldsymbol{x}_t, S_t, \omega_t)\}$ and generate q_t pair comparisons by rank breaking
4: Set $\widehat{\boldsymbol{\theta}}_{t,0} = \widehat{\boldsymbol{\theta}}_t$, $\boldsymbol{B}^+_{t,0} = \boldsymbol{B}_t$ and $\boldsymbol{M}^+_{t,0} = \boldsymbol{M}_t$
5: **for** $p = 1$ **to** q_t **do**
6: Define pairwise logistic loss $\ell_{t,p}(\boldsymbol{\theta})$, gradients $g_{t,p}(\boldsymbol{\theta})$, and Hessians $\boldsymbol{H}_{t,p}(\boldsymbol{\theta})$ as Eq. (1)
7: Construct $\widehat{\boldsymbol{\mathcal{H}}}_{t,p}$ as in Eq. (3) and compute $\widehat{\boldsymbol{\theta}}_{t,p}$ by Eq. (5)
8: Update $\left(\boldsymbol{B}^+_{t,p}, \boldsymbol{M}^+_{t,p}\right) = \mathsf{FD}_l\left(\boldsymbol{B}^+_{t,p-1}, \boldsymbol{M}^+_{t,p-1}, \sqrt{\dot{\mu}\left(\boldsymbol{z}_{t,p}^\top \widehat{\boldsymbol{\theta}}_{t,p}\right)}\boldsymbol{z}_{t,p}\right)$ by Algorithm 3
9: **end for**
10: Commit $\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_{t,q_t}$, $\boldsymbol{B}_{t+1} = \boldsymbol{B}^+_{t,q_t}$ and $\boldsymbol{M}_{t+1} = \boldsymbol{M}^+_{t,q_t}$
11: **end for**

The current-round correction in Eq. (3) is a sum of rank-one Hessians. Hence, after incorporating the historical sketch through Eq. (7), we further apply rank-one instances of Fact 1 sequentially for the first p comparisons. This allows all related operations to be carried out using only the sketch and low-rank updates, without materializing any $d \times d$ curvature matrix or its inverse.

Complexity. The complete procedure is summarized in Algorithm 1. Its per-round computational cost consists of two parts: updating the sketched OMD estimator in Line 7 and updating the FD sketch in Line 8. By Fact 1, each estimator update can be implemented using matrix-vector products involving sketch matrices $\boldsymbol{B}_t \in \mathbb{R}^{l\times d}$ and $\boldsymbol{M}_t \in \mathbb{R}^{l\times l}$, with cost $O(ld + l^2)$. Since rank breaking produces at most $q_t = O(K^2)$ pairwise comparisons per round, the estimator updates require $O(K^2ld + K^2l^2)$ time in total. Each FD update has an amortized cost of $O(ld)$ in its batched version, so the q_t sketch updates contribute an additional $O(K^2ld)$ per round. Therefore, the overall amortized per-round computational cost can be reduced from $O(K^2d^3)$ to $O(K^2ld)$. Moreover, we only maintains the sketch $\boldsymbol{B}_t$ and the auxiliary matrix $\boldsymbol{M}_t$, resulting in $O(ld + l^2)$ total space.

Theoretical Guarantee. We next establish an estimation error bound for the sketched online estimator, which connects the curvature approximation error to the resulting statistical error. Building on the local-norm OMD analysis of Zhang & Sugiyama (2023), we obtain the following bound

Lemma 1. *Let $\delta \in (0, 1]$ and set $\eta = \frac{1}{2}\log 2 + B + 1$, $\lambda = \max\{8\eta, 4B\eta, 1 + 2C_{\mathrm{int}}\}$. Then*

$$\Pr[\forall\, t \geq 1, \boldsymbol{\theta}^\star \in \mathcal{C}_t] \geq 1-\delta \text{ and } \mathcal{C}_t := \{\boldsymbol{\theta} \in \Theta \mid \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\|_{\widehat{\boldsymbol{\mathcal{H}}}_t} \leq \widehat{\beta}_t \triangleq O(\sqrt{l\log(tK/\delta) + d\Delta_l})\},$$

where C_{int} is coupling coefficient, spectral error Δ_l is defined in Eq. (4), and $\widehat{\boldsymbol{\mathcal{H}}}_t := \lambda\boldsymbol{I}_d + \boldsymbol{B}_t^\top\boldsymbol{B}_t$.

Remark 1. For the exact OMD estimator in Eq. (2), previous work (Lee et al., 2025) establishes an estimation error of order $\widetilde{O}(\sqrt{d\log(tK/\delta)})$ under the full local geometry. Lemma 1 shows that replacing the full historical curvature with an l-dimensional sketch changes the corresponding term from d to $l + d\Delta_l$. The additional term $d\Delta_l$ quantifies the cost of sketching approximation and decreases as the sketch better preserves the spectrum of the accumulated preference geometry.

4 Applications

In this section, we apply our GeoSRM to two representative online RLHF settings, active preference optimization and deployment-time adaptation, and establish corresponding theoretical guarantees.

4.1 Active Preference Optimization with Multi-Comparison Feedback

Active learning reduces labeling costs by selectively querying informative samples for annotation (Settles, 2009; Das et al., 2025; Verma et al., 2025b). We consider active preference optimization with multi-comparison feedback, where the learner jointly selects a context and several actions from the context set $\mathcal{X}$ and action set $\mathcal{A}$. Each query contain up to K actions and returns ranking feedback, allowing the learner to extract multiple preference comparisons from a single interaction.

Data Collection. At round t, the learner selects a query $(\boldsymbol{x}_t, S_t)$ with $|S_t| \leq K$. Following the uncertainty-based strategy in Lee et al. (2025), we favor queries with large average uncertainty under the current reward model. For a candidate context $\boldsymbol{x}$ and assortment S, define

$$f_t(\boldsymbol{x}, S) := \frac{1}{2|S|} \sum_{\boldsymbol{a},\boldsymbol{a}' \in S} \dot{\mu}\left(\phi(\boldsymbol{x},\boldsymbol{a})^\top \widehat{\boldsymbol{\theta}}_t - \phi(\boldsymbol{x},\boldsymbol{a}')^\top \widehat{\boldsymbol{\theta}}_t\right) \|\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}, \tag{8}$$

where $\widehat{\boldsymbol{\mathcal{H}}}_t = \lambda \boldsymbol{I}_d + \boldsymbol{B}_t^\top \boldsymbol{B}_t$ and $\widehat{\boldsymbol{\theta}}_t$ is the estimator at round t. For each context $\boldsymbol{x} \in \mathcal{X}$, we initialize $S(\boldsymbol{x})$ with the pair of actions that maximizes Eq. (8) and then greedily expand it by selecting

$$\boldsymbol{a}^\star = \operatorname*{arg\,max}_{\boldsymbol{a} \in \mathcal{A} \setminus S(\boldsymbol{x})} \left\{ f_t\left(\boldsymbol{x}, S(\boldsymbol{x}) \cup \{\boldsymbol{a}\}\right) - f_t\left(\boldsymbol{x}, S(\boldsymbol{x})\right) \right\}.$$

If the resulting marginal gain is non-negative, we add $\boldsymbol{a}^\star$ to $S(\boldsymbol{x})$ and repeat this step until no non-negative gain remains or $|S(\boldsymbol{x})| = K$. The query at round t is then chosen as

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x} \in \mathcal{X}} f_t(\boldsymbol{x}, S(\boldsymbol{x})), \qquad S_t = S(\boldsymbol{x}_t). \tag{9}$$

After T rounds, the learner returns the policy $\widehat{\pi}_T(\boldsymbol{x}) = \arg\max_{a \in \mathcal{A}} \phi(\boldsymbol{x},\boldsymbol{a})^\top \widehat{\boldsymbol{\theta}}_T$. The complete procedure is deferred to Appendix B. We next establish its worst-case sub-optimality guarantee.

Theorem 1 (Sub-optimality Gap of GeoSRM-A)**.** *Setting the parameters as in Lemma 1, with probability at least $1-\delta$, Algorithm 4 ensures the following worst-case sub-optimality gap guarantee*

$$\mathrm{SubOpt}_T = \sup_{\boldsymbol{x} \in \mathcal{X}} \left[r(\boldsymbol{x}, \pi^\star(\boldsymbol{x})) - r(\boldsymbol{x}, \widehat{\pi}_T(\boldsymbol{x})) \right] \leq \widetilde{O}\left(\frac{l + d\Delta_l}{T} \sqrt{\sum_{t=1}^{T} \frac{1+\Delta_l}{|S_t|}} \right),$$

where $\pi^\star$ is the optimal policy, Δ_l is defined in Proposition 1 and $|S_t|$ is the assortment size.

Remark 2. We match the data-dependent sub-optimality gap $\widetilde{O}\big(\frac{d}{T}\sqrt{\sum_{t=1}^{T} 1/|S_t|}\big)$ of Lee et al. (2025) whenever the preference geometry has light-tailed spectra. In particular, Theorem 1 retains the rigorous theoretical guarantee that larger subsets directly improve sample efficiency while replacing the full preference geometry with a compact sketch. The sketch also reduces the per-pair cost of evaluating the uncertainty term in Eq. (8) from $O(d^2)$ to $O(ld)$. The same preference sketching mechanism can be applied to related pairwise active preference optimizations (Das et al., 2025; Li et al., 2025), since their uncertainty computation is based on the same local preference geometry.

4.2 Deployment-Time Adaptation with Pairwise-Comparison Feedback

We then consider deployment-time adaptation with pairwise-comparison feedback, where user contexts arrive sequentially and the learner generates responses while collecting preference feedback to continually improve the model (Saha et al., 2023; Li et al., 2024b; 2025). At round t, the learner observes the context $\boldsymbol{x}_t$ and selects a pair of actions $S_t = \{\boldsymbol{a}_t, \boldsymbol{a}_t'\}$. The learner must balance two objectives: providing responses with high predicted rewards to maintain good deployment performance, while collecting informative comparisons for continual reward-model improvement.

Data Collection. Given the current reward model $\widehat{\boldsymbol{\theta}}_t$, we select one action greedily according to the estimated reward and the other using an optimistic criterion that balances reward and uncertainty:

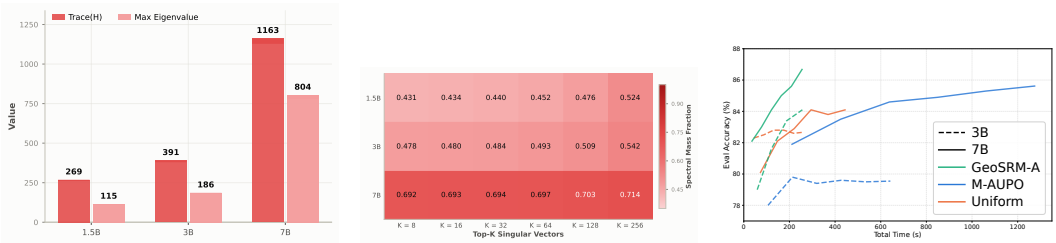
$$\boldsymbol{a}_t = \operatorname*{arg\,max}_{\boldsymbol{a} \in \mathcal{A}} \widehat{r}_t(\boldsymbol{a}), \qquad \boldsymbol{a}_t' = \operatorname*{arg\,max}_{\boldsymbol{a}' \in \mathcal{A}} \left\{ \widehat{r}_t(\boldsymbol{a}') + \widehat{\beta}_t \|\phi(\boldsymbol{x}_t, \boldsymbol{a}') - \phi(\boldsymbol{x}_t, \boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \right\}, \tag{10}$$

where $\widehat{r}_t(\boldsymbol{a}) := \phi(\boldsymbol{x}_t, \boldsymbol{a})^\top \widehat{\boldsymbol{\theta}}_t$, and $\widehat{\beta}_t$, $\widehat{\boldsymbol{\mathcal{H}}}_t$ are defined in Lemma 1. The first action $\boldsymbol{a}_t$ exploits the current reward estimate, whereas the second action $\boldsymbol{a}_t'$ additionally favors actions that are informative relative to $\boldsymbol{a}_t$ under the sketched preference geometry. The algorithm is deferred to Appendix B for brevity. We next show that this deployment strategy satisfies the following regret guarantee.

Theorem 2 (Regret Bound of GeoSRM-D)**.** *For any $\delta \in (0, 1]$, set the parameters as in Lemma 1. Then with probability at least $1-\delta$, Algorithm 5 ensure the following regret guarantee*

$$\mathrm{Regret}_T = \sum_{t=1}^{T} \left(r\left(\boldsymbol{x}_t, \pi^\star(\boldsymbol{x}_t)\right) - \frac{1}{2}\left(r(\boldsymbol{x}_t, \boldsymbol{a}_t) + r(\boldsymbol{x}_t, \boldsymbol{a}_t')\right) \right) \leq \widetilde{O}\left((l + d\Delta_l)\sqrt{T} \right),$$

where $\pi^\star$ is the optimal policy, l is sketch size and Δ_l is the spectral error defined in Proposition 1.



(a) Eigenvalue concentration (b) Spectral mass (c) Accuracy vs. Time

Figure 3: (a), (b): Spectral concentration of the preference geometry across `Qwen2.5-Instruct` with 1.5B, 3B, and 7B. (c): Pareto frontiers for accuracy vs. time in active setting w.r.t. model size.

Remark 3. Unlike the preceding sub-optimality guarantee for the final policy, Theorem 2 controls cumulative reward loss throughout online adaptation, reflecting the need to maintain response quality during deployment. The non-sketched method achieves $\widetilde{O}(d\sqrt{T})$ regret (Li et al., 2025). Relative to this full-geometry rate, sketching introduces the approximation term $l + d\Delta_l$, explicitly characterizing the statistical price of geometry compression. At the same time, the sketched geometry reduces the per-action cost of evaluating the exploration bonus in Eq. (10) from $O(d^2)$ to $O(ld)$.

5 Experiments

In this section, we empirically evaluate the performance of proposed method. We first answer this question: *Is preference curvature spectrally compressible in realistic LLM representations?*

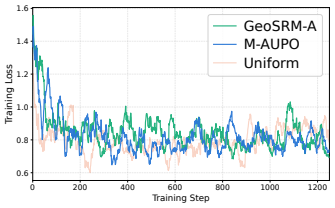
Spectral Analysis. We investigate whether preference curvature in realistic LLM representations exhibits spectral structure that supports compact approximation. Figure 3a shows that the intrinsic dimension, $\mathrm{Tr}(\boldsymbol{H})/\lambda_{\max}(\boldsymbol{H})$, consistently decreases as the backbone model size increases. This suggests that larger LLMs exhibit increasingly concentrated curvature. Figure 3b provides complementary evidence: the top 256 directions capture 52.4%, 54.2%, and 71.4% of the spectral mass, respectively. For the 7B backbone, the leading eight directions alone account for 69.2%. Although a non-negligible spectral tail remains, these results show that a fixed number of directions captures an increasing fraction of preference curvature at larger model scales. This concentration provides empirical support for compactly approximating the dominant geometry via matrix sketching, as Lemma 1 bounds the approximation error by the data-dependent spectral tail.

5.1 Active Preference Optimization

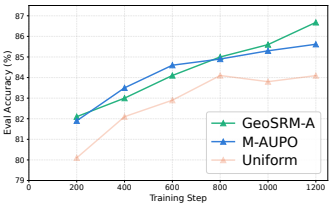
Experimental Setup. We conduct experiments using real-world dataset Nectar (Zhu et al., 2024), which provides seven ranked responses per prompt. Using frozen `Qwen2.5-Instruct` backbones (Qwen Team, 2024) of 3B and 7B parameters, we train linear reward heads over four passes through a pool of $5,000$ prompts. We compare GeoSRM-A with the active learning baseline M-AUPO (Lee et al., 2025) and a random-context baseline (Uniform). GeoSRM-A and M-AUPO select a prompt from each candidate minibatch, whereas Uniform randomly selects prompt. All methods then select assortment of up to seven responses and update the reward head using pairwise comparisons induced by the observed ranking. Appendix F.1 details the implementation settings.

Active Learning Efficacy. Figures 4a, 4b show the training loss and evaluation accuracy over iterations. GeoSRM-A and M-AUPO achieve comparable training losses, while our method attains higher evaluation accuracy in the later stages and the best final performance. Besides, GeoSRM-A consistently outperforms Uniform, showing the benefit of informative preference selection.

Scalable Accuracy–Time Trade-off. Figure 3c compares the accuracy–time trade-off as the model scales from 3B to 7B. While M-AUPO remains competitive in accuracy, its computational cost grows substantially with model size: on the 7B model, it requires over $1{,}200$ seconds to reach an evaluation accuracy of about 85.6%. Uniform incurs substantially lower computational cost, but its accuracy saturates at a noticeably lower level. In contrast, our method attains a comparable accuracy within only 260 seconds, yielding roughly a $5\times$ reduction in algorithm time compared with M-AUPO. This



(a) Training loss

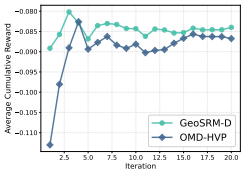


(b) Evaluation accuracy

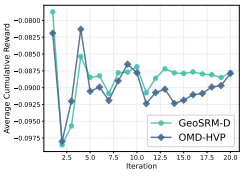
Method	ACC (%)	Time (s/step)
Ours ($K = 2$)	63.46	0.05±0.01
Ours ($K = 5$)	72.20	0.11±0.03
M-AUPO ($K = 7$)	85.62	1.06±0.02
Ours ($K = 7$)	**86.68**	0.21±0.08

(c) Effects of assortment size

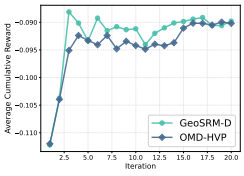
Figure 4: Online RLHF with active preference optimization in `Qwen2.5-7B-Instruct`. We report (a) training loss, (b) evaluation accuracy, and (c) effects of maximum assortment size K.



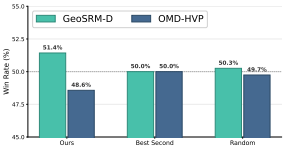
(a) `Qwen2.5-1.5B`



(b) `Qwen2.5-3B`



(c) `Qwen2.5-7B`



(d) Win rate

Figure 5: Online RLHF with deployment-time adaptation in `Qwen2.5-Instruct`. (a), (b), (c): Average cumulative reward w.r.t. model size; (d): Win rate w.r.t. different action selection strategies.

gap is substantially larger than that observed at 3B, indicating that GeoSRM-A preserves a strong accuracy–efficiency trade-off while scales more favorably in LLMs applications.

Effects of K. We further examine the effect of the maximum assortment size. Increasing K consistently improves accuracy, while the per-step update time grows only moderately. In realistic application, K is much smaller than the feature dimension d. Thus, larger assortments yield substantial accuracy gains for GeoSRM-A without the steep increase in computation observed for M-AUPO.

5.2 Deployment-time Adaptation

Experimental Setup. We evaluate our method on UltraFeedback (Cui et al., 2024) datasets over 20 deployment iterations. We use `Llama-3.2-1B-Instruct` (Llama Team, 2024) as the policy model and `Qwen2.5-Instruct` backbones of 1.5B, 3B, and 7B parameters for reward modeling. Each iteration fits the reward head and updates the policy with PPO (Schulman et al., 2017), then generates responses and collects preference feedback using a fixed reward oracle. We compare GeoSRM-D with the HVP-based OMD baseline (Li et al., 2025), which also approximates curvature.

Deployment-time Learning Efficacy. Effective deployment-time adaptation should learn from accumulating feedback while maintaining strong performance. Figures 5a, 5b, and 5c report the average cumulative reward across deployment rounds. GeoSRM-D generally improves its performance following the initial fluctuations, with the clearest upward trend observed at 7B. Across all three model sizes, our method achieves higher rewards than OMD-HVP over most deployment rounds.

Effects of Selection Strategy. We compare our strategy with (1): random selection and (2): select the best and second best action; using win rate to measure how often a method selects a pair with higher mean oracle reward. Figure 5d shows that GeoSRM-D matches or outperforms OMD-HVP across all selection strategies. The results favor sketching-based approximation for deployment-time adaptation, particularly when learning requires balancing immediate reward with exploration.

6 Conclusion

We present GeoSRM, a geometry-sketched framework for efficient one-pass online RLHF, with estimation and theoretical guarantee that account for sketching error. Experiments demonstrate favorable efficiency and utility in active preference optimization and deployment-time adaptation, supporting the value of preserving dominant preference-curvature directions at larger model scales.

AI USE STATEMENT

We used generative AI tools for language polishing, proof checking, and coding assistance. Language-related use focused on improving grammar, clarity, fluency, and consistency of presentation, while coding assistance supported implementation and code refinement. All AI-assisted content and code were reviewed and verified by the authors. The authors take full responsibility for the final manuscript, including its scientific claims, implementation, experimental results, and references.

ETHICS STATEMENT

Our experiments use existing preference datasets and pretrained models, with deployment feedback supplied by a fixed reward-model oracle. These resources may contain biases that propagate to learned policies. Improved computational efficiency and reward scores do not guarantee fairness or safety; real-world deployment requires additional evaluation beyond the benchmarks studied here.

REPRODUCIBILITY STATEMENT

To support reproducibility, we present the problem formulation, assumptions, and algorithms in Sections 2–4, with detailed pseudocode and proofs provided in the appendix. Section 5 and Appendix F.1 describe the datasets, backbone models, data selection procedures, hyperparameters, and evaluation protocols. Our code is available in an anonymous repository at `https://anonymous.4open.science/r/GeoSRM`.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24 (NeurIPS)*, pp. 2312–2320, 2011.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3691–3699, 2021.

Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the 29th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 274–281, 2001.

Francis R. Bach. Sharp analysis of low-rank kernel matrix approximations. In Shai Shalev-Shwartz and Ingo Steinwart (eds.), *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, volume 30, pp. 185–209, 2013.

Catarina Barata, Veronica Rotemberg, Noel C. F. Codella, Philipp Tschandl, Christoph Rinner, Bengu Nisa Akay, Zoe Apalla, Giuseppe Argenziano, Allan Halpern, Aimilios Lallas, Caterina Longo, Josep Malvehy, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. A reinforcement learning model for AI-based decision support in skin cancer. *Nature Medicine*, 29(8):1941–1946, 2023.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pp. 1764–1786, 2022.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

Nicolò Campolongo and Francesco Orabona. Temporal variability in implicit online learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach

to online and offline RLHF. In *Proceedings of the the 13th International Conference on Learning Representations (ICLR)*, 2025.

Cheng Chen, Luo Luo, Weinan Zhang, Yong Yu, and Yijiang Lian. Efficient and robust high-dimensional linear contextual bandits. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4259–4265, 2020.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 4299–4307, 2017.

Michael B. Cohen, Cameron Musco, and Jakub Pachocki. Online row sampling. *Theory Comput.*, 16:1–25, 2020.

Graham Cormode and Ke Yi. *Small Summaries for Big Data*. Cambridge University Press, 2020.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient RLHF. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 16017, pp. 96–112, 2025.

Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. In *Proceedings of the the 12th International Conference on Learning Representations (ICLR)*, 2024.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF workflow: From reward modeling to online RLHF. *Trans. Mach. Learn. Res.*, 2024, 2024.

Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R. Srikant. Exploration-driven policy optimization in RLHF: theoretical insights on efficient data utilization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 11830–11887, 2024.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, volume 40, pp. 563–587, 2015.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 546–580, 2022.

Vladimir Feinberg, Xinyi Chen, Y. Jennifer Sun, Rohan Anil, and Elad Hazan. Sketchy: Memory-efficient adaptive regularization with frequent directions. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Mach. Learn.*, 28(2-3):133–168, 1997.

Alan M. Frieze, Ravi Kannan, and Santosh S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041, 2004.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5):1762–1792, 2016.

Çaglar Gülçehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Siyuan Guo, Yali Du, Hechang Chen, Yi Chang, and Jun Wang. Cascade: Case-based continual adaptation for large language models during deployment. *arXiv preprint arXiv:2605.06702*, 2026.

Moritz Hardt and Yu Sun. Test-time training on nearest neighbors for large language models. In *Proceedings of the the 12th International Conference on Learning Representations (ICLR)*, 2024.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the the 10th International Conference on Learning Representations (ICLR)*, 2022.

Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(10):1936–1949, 2014.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. *Math. Oper. Res.*, 48(3):1496–1521, 2023.

Ilja Kuzborskij, Leonardo Cella, and Nicolò Cesa-Bianchi. Efficient linear bandits through matrix sketching. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 177–185, 2019.

Joongkyu Lee and Min-hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Joongkyu Lee, Seouh-won Yi, and Min-hwan Oh. Preference-based reinforcement learning beyond pairwise comparisons: Benefits of multiple options. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Long-Fei Li, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Provably efficient reinforcement learning with multinomial logit function approximation. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024a.

Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. Provably efficient online RLHF with one-pass reward modeling. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Xuheng Li, Heyang Zhao, and Quanquan Gu. Feel-good thompson sampling for contextual dueling bandits. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 29406–29426, 2024b.

Edo Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 581–588, 2013.

Llama Team. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation, 2012.

Feng Luo, Rui Yang, Hao Sun, Chunyuan Deng, Jiarui Yao, Jingyan Shen, Huan Zhang, and Hanjie Chen. Rethinking diverse human preference learning through principal component analysis. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 19857–19870, 2025.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pp. 902–910, 2016.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis R. Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Proceedings of the 32nd Annual Conference on Learning Theory (COLT)*, volume 99, pp. 2294–2340, 2019.

Zakaria Mhammedi, Wouter M. Koolen, and Tim van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Proceedings of the 32nd Annual Conference on Learning Theory (COLT)*, pp. 2490–2511, 2019.

Jayadev Misra and David Gries. Finding repeated elements. *Science of computer programming*, 2 (2):143–152, 1982.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 36577–36590, 2024.

Seiji Nabeya. Absolute moments in 2-dimensional normal distribution. *Annals of the Institute of Statistical Mathematics*, 3(1):1–1, 1951.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP): Findings*, pp. 1043–1067, 2024.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Comput.*, 6(1):147–160, 1994.

Mert Pilanci and Martin J. Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *J. Mach. Learn. Res.*, 17:53:1–53:38, 2016.

Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1975.

Zikun Qu, Min Zhang, Mingze Kong, Xiang Li, Zhiwei Shang, Zhiyong Wang, Yikun Ban, Shuang Qiu, Yao Shu, and Zhongxiang Dai. T-POP: Test-time personalization with online preference feedback. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 30050–30062, 2021.

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling RL: reinforcement learning with trajectory preferences. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206, pp. 6263–6289, 2023.

Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 143–152, 2006.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Burr Settles. Active learning literature survey. *Technical Report*, 2009.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the 20th European Conference on Computer Systems (EuroSys)*, pp. 1279–1297, 2025.

Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

Benwei Shi and Jeff M. Phillips. A deterministic streaming sketch for ridge regression. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 586–594, 2021.

David Silver and Richard S Sutton. Welcome to the era of experience. *Google AI*, 1, 2025.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, volume 33, 2020.

Alane Suhr and Yoav Artzi. Continual learning for instruction following from realtime feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Mirac Suzgun, Mert Yüksekgönül, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 7080–7106, 2026.

Arun Verma, Zhongxiang Dai, Xiaoqiang Lin, Patrick Jaillet, and Bryan Kian Hsiang Low. Neural dueling bandits: Preference-based optimization with human feedback. In *Proceedings of the the 13th International Conference on Learning Representations (ICLR)*, 2025a.

Arun Verma, Xiaoqiang Lin, Zhongxiang Dai, Daniela Rus, and Bryan Kian Hsiang Low. Active human feedback collection via neural contextual dueling bandits. *arXiv preprint arXiv:2504.12016*, 2025b.

Jing Wang, Miao Yu, Peng Zhao, and Zhi-Hua Zhou. Learning with adaptive resource allocation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 52099–52116, 2024.

Dongxie Wen, Lu Yi, Fengtai Li, Shaoqiu Zhang, Xiao Zhang, and Zhewei Wei. Certified streaming updates for machine learning models under limited throughput. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2026a.

Dongxie Wen, Hanyan Yin, Xiao Zhang, Peng Zhao, Lijun Zhang, and Zhewei Wei. Revisiting matrix sketching in linear bandits: Achieving sublinear regret via dyadic block sketching. In *Proceedings of the the 14th International Conference on Learning Representations (ICLR)*, 2026b.

Max A. Woodbury. Inverting modified matrices. Technical report, Statistical Research Group, Princeton University, 1950.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

Runzhe Wu and Wen Sun. Making RL with preference-based feedback efficient via randomization. In *Proceedings of the the 12th International Conference on Learning Representations (ICLR)*, 2024.

Tengyang Xie, Dylan J. Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Hassan Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient RLHF. In *Proceedings of the the 13th International Conference on Learning Representations (ICLR)*, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under kl-constraint. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 54715–54754, 2024.

Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. Prefrec: Recommender systems with human preferences for reinforcing long-term user engagement. In *Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2874–2884, 2023.

Sifan Yang, Yuanyu Wan, Peijia Li, Yibo Wang, Xiao Zhang, Zhewei Wei, and Lijun Zhang. Dimension-free adaptive subgradient methods with frequent directions. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556, 2012.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *Proceedings of the the 14th International Conference on Learning Representations (ICLR)*, 2026.

Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized linear bandits: Almost optimal regret with one-pass update. In *Advances in Neural Information Processing Systems 39 (NeurIPS)*, 2025.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 61121–61143, 2024.

Zhi-Hua Zhou. Learnability with time-sharing computational resource concerns. *National Science Review*, 11(10):204, 2024.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 43037–43067, 2023.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *Proceedings of the First Conference on Language Modeling (COLM)*, 2024.

CONTENTS

1 Introduction **1**

2 Preliminaries and Problem Setup **2**

2.1 Online RLHF 2

2.2 Online Parameter Estimation 3

2.3 Matrix sketching 4

3 Algorithms **4**

3.1 Preference Geometry Sketching 4

3.2 Geometry-Sketched Reward Modeling Framework 5

4 Applications **6**

4.1 Active Preference Optimization with Multi-Comparison Feedback 6

4.2 Deployment-Time Adaptation with Pairwise-Comparison Feedback 7

5 Experiments **8**

5.1 Active Preference Optimization 8

5.2 Deployment-time Adaptation 9

6 Conclusion **9**

A More Related Works **17**

B Omitted Algorithms **18**

B.1 Pseudo-code of Frequent Directions 18

B.2 Pseudocode of GeoSRM-A and GeoSRM-D 19

C Omitted Proofs for Section 3 **20**

C.1 Main Proof of Lemma 1 21

C.2 Proofs of the Auxiliary Lemmas 24

D Omitted Proofs for Section 4 **27**

D.1 Proof of Theorem 1 27

D.2 Proof of Theorem 2 29

E Omitted Technical Lemmas **30**

F Omitted Details for Section 5 **31**

F.1 Implementation Details 31

F.2 More Experimental Results 32

A More Related Works

In this section, we provide additional related works and outline several future directions.

Dueling Bandits. Dueling bandits extend the multi-armed bandit framework by allowing the learner to sequentially compare pairs of arms and observe binary preference feedback (Yue et al., 2012). Contextual dueling bandits further incorporate contextual information into the comparison process (Dudík et al., 2015; Saha, 2021; Bengs et al., 2022; Di et al., 2024). Beyond linear reward models, neural dueling bandit methods use neural networks to model nonlinear reward functions (Verma et al., 2025a;b; Qu et al., 2026).

Recent works have improved the computational efficiency of preference-based online learning by replacing repeated maximum likelihood estimation with online mirror descent updates (Li et al., 2024a; 2025; Lee et al., 2025). These methods process each new observation online, making the per-round update cost independent of the size of the accumulated interaction history. Our work addresses a complementary computational bottleneck: the remaining dependence on the feature dimension d. By sketching the accumulated preference geometry, GeoSRM further reduces the time and space complexity with respect to d while retaining the one-pass nature of these online estimators.

Active Learning and Deployment-time Learning. Active learning aims to reduce labeling costs by selectively querying the most informative samples for annotation (Settles, 2009; Muldrew et al., 2024). A typical active learning procedure iteratively selects informative samples from a large unlabeled pool, queries their labels, updates the model, and repeats this process (Freund et al., 1997; Huang et al., 2014). This paradigm has also been extended to RLHF, where preference queries are actively selected to improve feedback efficiency (Das et al., 2025; Verma et al., 2025b).

Deployment-time learning allows models to continue adapting during deployment by learning from experience collected through interactions with the environment (Zhang et al., 2026; Suzgun et al., 2026; Guo et al., 2026). One of the central challenges in this setting is balancing immediate deployment performance with long-term model improvement: the learner must provide high-quality responses to current users while selecting sufficiently informative actions to improve future behavior. Dueling reinforcement learning formalizes this trade-off through cumulative regret, where the rewards of both queried trajectories contribute to the online objective (Saha et al., 2023; Wu & Sun, 2024; Qu et al., 2026).

Sketching-based Approaches. Streaming matrix sketching techniques provide compact representations of large-scale matrices while approximately preserving their essential structure. Existing approaches can be broadly classified into three categories. The first category is sampling-based sketching, which approximates a matrix by selecting a subset of representative rows or columns (Frieze et al., 2004; Cohen et al., 2020). The second category uses randomized transformations, where matrix rows are combined through random projections or hashing to construct a lower-dimensional representation (Achlioptas, 2001; Sarlós, 2006). The third category adopts deterministic sketching methods, such as Frequent Directions proposed by Liberty (2013), which extends the Misra–Gries algorithm (Misra & Gries, 1982) to matrix streams by preserving directions with large spectral contributions. For further details, we refer readers to the survey (Woodruff, 2014).

Sketching techniques have been extensively studied for balancing computational efficiency and statistical accuracy in large-scale learning. This line of research spans statistical estimation (Bach, 2013; Pilanci & Wainwright, 2016; Shi & Phillips, 2021), sequential decision-making (Kuzborskij et al., 2019; Chen et al., 2020; Wen et al., 2026b), online optimization (Luo et al., 2016; Feinberg et al., 2023; Yang et al., 2025), and streaming learning (Zhou, 2024; Wang et al., 2024; Wen et al., 2026a). These developments motivate the use of sketching in LLM applications such as RLHF, where high-dimensional representations can impose substantial computational costs.

Future Directions. Due to length constraints in the main text, we outline the limitations of this work and several potential future directions here. First, GeoSRM improves the efficiency of reward estimation and uncertainty evaluation, while end-to-end online RLHF also involves other computationally intensive stages, including response generation, feature extraction, and policy optimization. Integrating GeoSRM with efficient RLHF systems such as HybridFlow (Sheng et al., 2025) would enable a more comprehensive study of end-to-end training efficiency. Second, our current framework applies sketching only to reward modeling. Extending this principle to policy optimization,

Algorithm 2 Update procedure of FD

Input: Current sketch $(\boldsymbol{B}_t, \boldsymbol{M}_t)$ with sketch size l, newly arrived row $\boldsymbol{x}_t$, and $\lambda > 0$

1: Append $\boldsymbol{x}_t$ to the last row of $\boldsymbol{B}_t$
2: Compute $[\boldsymbol{U}_t, \boldsymbol{\Sigma}_t, \boldsymbol{V}_t] \leftarrow \text{SVD}(\boldsymbol{B}_t)$
3: Define $\boldsymbol{\Sigma}_t = \text{diag}(\sigma_1, \cdots, \sigma_l)$
4: Set $\rho \leftarrow \sigma_l^2$
5: Update sketch matrix $\boldsymbol{B}_t$ by

$$\boldsymbol{B}_{t+1} \leftarrow \sqrt{\boldsymbol{\Sigma}_t^2 - \rho \boldsymbol{I}_l} \cdot \boldsymbol{V}_t^\top$$

6: Update sketch matrix $\boldsymbol{M}_t$ by

$$\boldsymbol{M}_{t+1} \leftarrow \text{diag}\left\{\frac{1}{\lambda + \sigma_1^2 - \rho}, \cdots, \frac{1}{\lambda}\right\}$$

Output: Sketch matrices $\boldsymbol{B}_{t+1}$, $\boldsymbol{M}_{t+1}$.

Algorithm 3 Update procedure of Batch-FD

Input: Current sketch $(\boldsymbol{B}_t, \boldsymbol{M}_t)$ with sketch size $2l$, newly arrived row $\boldsymbol{x}_t$, and $\lambda > 0$

1: Inset $\boldsymbol{x}_t$ to a zero valued row of $\boldsymbol{B}_t$
2: **if** $\boldsymbol{B}_t$ has no zero valued rows **then**
3: Compute $[\boldsymbol{U}_t, \boldsymbol{\Sigma}_t, \boldsymbol{V}_t] \leftarrow \text{SVD}(\boldsymbol{B}_t)$
4: Define $\boldsymbol{\Sigma}_t = \text{diag}(\sigma_1, \cdots, \sigma_{2l})$
5: Set $\rho \leftarrow \sigma_l^2$
6: Update sketch matrix $\boldsymbol{B}_t$ by

$$\boldsymbol{B}_{t+1} \leftarrow \sqrt{\max\left(\boldsymbol{\Sigma}_t^2 - \rho \boldsymbol{I}_{2l}, 0\right)} \cdot \boldsymbol{V}_t^\top$$

The last $l+1$ rows are zero valued
7: Set $\boldsymbol{M}_{t+1} \leftarrow \text{diag}\left\{\frac{1}{\lambda + \sigma_1^2 - \rho}, \cdots, \frac{1}{\lambda}\right\}$
8: **else**
9: Update $\boldsymbol{M}_{t+1}$ via Eq. (11)
10: **end if**

Output: Sketch matrices $\boldsymbol{B}_{t+1}$, $\boldsymbol{M}_{t+1}$.

building on recent progress in sketch-based preconditioning (Feinberg et al., 2023; Yang et al., 2025) and low-rank optimization (Zhao et al., 2024), may further reduce the computational and memory costs of online RLHF. Third, GeoSRM currently uses a fixed sketch size, although the complexity of the preference geometry may change over the course of learning. Developing adaptive sketching strategies, potentially inspired by dyadic block sketching (Wen et al., 2026b), could provide a more flexible trade-off between computational efficiency and statistical accuracy.

B Omitted Algorithms

This section provides the omitted algorithms. We first present the standard and batched FD updates in Appendix B.1, and then give the pseudocode of GeoSRM-D and GeoSRM-A for deployment-time adaptation and active preference optimization in Appendix B.2.

B.1 Pseudo-code of Frequent Directions

Frequent Directions (FD) is a deterministic matrix sketching algorithm for streaming data (Liberty, 2013; Ghashami et al., 2016). After each sketch update, FD maintains a zero-valued last row in the sketch matrix $\boldsymbol{B}$, which is used to accommodate the next incoming row. Given a new row $\boldsymbol{x}_t$, FD inserts $\boldsymbol{x}_t$ into this row and computes the singular value decomposition $\boldsymbol{B}_t = \boldsymbol{U}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^\top$. It then shrinks the singular values by setting $\rho = \sigma_l^2$ and updating $\boldsymbol{B}_{t+1} = \sqrt{\boldsymbol{\Sigma}_t^2 - \rho \boldsymbol{I}_l} \boldsymbol{V}_t^\top$. This update sets the last singular value to zero, leaving the last row of $\boldsymbol{B}_{t+1}$ available for the next observation. Moreover, since the rows of $\boldsymbol{B}_{t+1}$ are orthogonal, $\boldsymbol{M}_{t+1} := (\boldsymbol{B}_{t+1}\boldsymbol{B}_{t+1}^\top + \lambda \boldsymbol{I}_l)^{-1}$ is diagonal and can therefore be maintained efficiently.

Algorithm 2 summarizes the standard FD update. Its computational cost is dominated by the SVD of the $l \times d$ sketch matrix. Because standard FD performs an SVD after every incoming row, each update requires $O(dl^2)$ time. This cost can be reduced to $O(dl)$ amortized time by maintaining a sketch of size $2l$ and performing compression only after l new rows have been accumulated (Liberty, 2013; Luo et al., 2016; Kuzborskij et al., 2019). Algorithm 3 summarizes this batched variant.

Maintaining $\boldsymbol{M}_t$ in Batch-FD. Unlike standard FD, the rows of the Batch-FD sketch are generally not orthogonal between two consecutive compression steps, and thus $\boldsymbol{M}_t$ is not necessarily diagonal during this period. Nevertheless, it can be maintained efficiently through a block inverse update. Let $\bar{\boldsymbol{B}}_t \in \mathbb{R}^{m \times d}$ contain the current nonzero rows of the sketch, where $m < 2l$, and let $\boldsymbol{M}_t = (\bar{\boldsymbol{B}}_t \bar{\boldsymbol{B}}_t^\top + \lambda \boldsymbol{I}_m)^{-1}$. When a new row $\boldsymbol{x}_t$ is inserted, define

$$\boldsymbol{c}_t = \bar{\boldsymbol{B}}_t \boldsymbol{x}_t, \qquad s_t = \lambda + \|\boldsymbol{x}_t\|_2^2 - \boldsymbol{c}_t^\top \boldsymbol{M}_t \boldsymbol{c}_t.$$

Algorithm 4 GeoSRM-A

Input: Sketch size l, step size η, regularization parameter λ, maximum assortment size K
Output: Final policy $\widehat{\pi}_{T+1}$

1: Initialize estimator $\widehat{\boldsymbol{\theta}}_1 \in \Theta$ and sketch matrices $\boldsymbol{B}_1 = \mathbf{0}_{l\times d}$ and $\boldsymbol{M}_1 = \frac{1}{\lambda}\boldsymbol{I}_l$.
2: **for** $t = 1$ **to** T **do**
3: For each candidate context $\boldsymbol{x}$, greedily construct $S(\boldsymbol{x})$ according to Eq. (8)
4: Select $(\boldsymbol{x}_t, S_t)$ according to Eq. (9)
5: Observe ranking feedback ω_t and generate q_t pair comparisons by rank breaking
6: Set $\widehat{\boldsymbol{\theta}}_{t,0} = \widehat{\boldsymbol{\theta}}_t$, $\boldsymbol{B}^+_{t,0} = \boldsymbol{B}_t$ and $\boldsymbol{M}^+_{t,0} = \boldsymbol{M}_t$
7: **for** $p = 1$ **to** q_t **do**
8: Construct $\widehat{\boldsymbol{\mathcal{H}}}_{t,p}$ as in Eq. (3) and compute $\widehat{\boldsymbol{\theta}}_{t,p}$ by Eq. (5)
9: Update $\left(\boldsymbol{B}^+_{t,p}, \boldsymbol{M}^+_{t,p}\right) = \mathsf{FD}_l\left(\boldsymbol{B}^+_{t,p-1}, \boldsymbol{M}^+_{t,p-1}, \sqrt{\dot{\mu}(\boldsymbol{z}^\top_{t,p}\widehat{\boldsymbol{\theta}}_{t,p})}\boldsymbol{z}_{t,p}\right)$ by Algorithm 2
10: **end for**
11: Commit $\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_{t,q_t}$, $\boldsymbol{B}_{t+1} = \boldsymbol{B}^+_{t,q_t}$ and $\boldsymbol{M}_{t+1} = \boldsymbol{M}^+_{t,q_t}$
12: **end for**
13: Return final policy $\widehat{\pi}_{T+1}(\boldsymbol{x}) = \arg\max_{a\in\mathcal{A}} \phi(\boldsymbol{x}, \boldsymbol{a})^\top \widehat{\boldsymbol{\theta}}_{T+1}$

The inverse associated with the augmented sketch can then be updated as

$$\boldsymbol{M}_{t+1} = \begin{bmatrix} \boldsymbol{M}_t + s_t^{-1}\boldsymbol{M}_t\boldsymbol{c}_t\boldsymbol{c}_t^\top\boldsymbol{M}_t & -s_t^{-1}\boldsymbol{M}_t\boldsymbol{c}_t \\ -s_t^{-1}\boldsymbol{c}_t^\top\boldsymbol{M}_t & s_t^{-1} \end{bmatrix}. \tag{11}$$

This update follows directly from the block matrix inversion formula and avoids recomputing the inverse from scratch. When the buffer becomes full and Batch-FD performs an SVD compression, the resulting sketch rows become orthogonal again, so $\boldsymbol{M}_{t+1}$ can be reset to the diagonal form shown in Algorithm 3. The block update requires $O(md + m^2)$ time with $m \le 2l$, which is $O(ld + l^2)$ per inserted row and reduces to $O(ld)$ when $l \le d$.

B.2 Pseudocode of GeoSRM-A and GeoSRM-D

We provide the pseudocode for two instantiations of GeoSRM: GeoSRM-D for Deployment-time adaptation with pairwise-comparison feedback, and GeoSRM-A for Active preference optimization with multi-comparison feedback. GeoSRM-D combines the data collection rule in Section 4.2 with the sketched OMD estimator, as summarized in Algorithm 5. GeoSRM-A combines the active query selection rule in Section 4.1 with the rank-broken sketched OMD updates, as summarized in Algorithm 4. For GeoSRM-D, the quantity $\widehat{\beta}_t$ in Eq. (10) can be updated efficiently by maintaining the cumulative sketch shrinkage $\sum_{s<t}\sum_{p=1}^{q_s}\rho_{s,p}$ (details in Appendix C). For clarity, both algorithms are presented using standard FD. The Batch-FD implementation in Algorithm 3 can be used instead to obtain the amortized sketch-update complexity discussed in the main text.

The pairwise active preference optimization setting considered in Section 5.2 of Li et al. (2025) can be handled in the same framework. Its reward-model update uses the same pairwise local geometry, so the sketched OMD estimator in Eq. (5) applies without modification. It suffices to replace the acquisition rule in Line 4 of Algorithm 4 with the corresponding pairwise rule that maximizes the unweighted predictive uncertainty. We therefore omit a separate pseudocode for this special case.

Overall Computational Complexity. Let $N = |\mathcal{X}|$ and $n = |\mathcal{A}|$ denote the numbers of candidate contexts and actions, respectively. We consider $2 \le K \le n$ and the regime $K \ll l < d$ assumed throughout the paper. We count arithmetic operations on available feature vectors and use Batch-FD (Algorithm 3) for amortized sketch maintenance. By Eq. (7), evaluating a pairwise uncertainty costs $O(ld)$. For GeoSRM-A, finding the best initial pair for every context therefore costs $O(Nn^2ld)$. During greedy expansion, we maintain, for each remaining action, the sum of its weighted pairwise uncertainties with the selected actions. Each insertion requires at most n new pairwise evaluations, giving $O(NKnld)$ time over at most $K-2$ insertions. Selecting the context with the largest resulting score costs $O(N)$. For GeoSRM-D, computing predicted rewards and selecting the greedy action costs $O(nd)$, while scanning the optimistic second action costs $O(nld)$. Since its context is observed, no search over $\mathcal{X}$ is required.

Algorithm 5 GeoSRM-D

Input: Sketch size l, step size η, regularization parameter λ
Output: Adapted reward model $\widehat{\boldsymbol{\theta}}_t$ at round t

1: Initialize estimator $\widehat{\boldsymbol{\theta}}_1 \in \Theta$ and sketch matrices $\boldsymbol{B}_1 = \mathbf{0}_{l\times d}$ and $\boldsymbol{M}_1 = \frac{1}{\lambda}\boldsymbol{I}_l$.
2: **for** $t = 1$ **to** T **do**
3: Observe the context $\boldsymbol{x}_t$ from user
4: Select actions $\boldsymbol{a}_t$ and $\boldsymbol{a}'_t$ according to Eq. (10)
5: Observe the preference feedback $y_t = \mathbb{I}(\boldsymbol{a}_t \succ \boldsymbol{a}'_t)$
6: Set $\boldsymbol{z}_t = (2y_t - 1)(\phi(\boldsymbol{x}_t, \boldsymbol{a}_t) - \phi(\boldsymbol{x}_t, \boldsymbol{a}'_t))$
7: Construct $\widehat{\boldsymbol{\mathcal{H}}}_{t,1}$ as in Eq. (3) and compute $\widehat{\boldsymbol{\theta}}_{t+1}$ by Eq. (5) (case $p = 1$)
8: Update $(\boldsymbol{B}_{t+1}, \boldsymbol{M}_{t+1}) = \mathsf{FD}_l\Big(\boldsymbol{B}_t, \boldsymbol{M}_t, \sqrt{\dot{\mu}\big(\boldsymbol{z}_t^\top \widehat{\boldsymbol{\theta}}_{t+1}\big)}\boldsymbol{z}_t\Big)$ by Algorithm 2
9: **end for**

Table 1: Per-round computational complexity of GeoSRM-A and GeoSRM-D in the regime $K \ll l < d$. Sketch-maintenance costs are amortized using Batch-FD. Here $N = |\mathcal{X}|$ and $n = |\mathcal{A}|$.

Component	**GeoSRM-A (Alg. 4)**	**GeoSRM-D (Alg. 5)**
Predicted reward evaluation	$O(Nnd)$	$O(nd)$
Initialize $S(x)$ with the best pair	$O(Nn^2ld)$	—
Expand $S(x)$ / optimistic selection	$O(NKnld)$	$O(nld)$
Context x_t selection	$O(N)$	—
Estimator updates	$O(K^2ld)$	$O(ld)$
Sketch maintenance	$O(K^2ld)$	$O(ld)$

For the estimator updates, GeoSRM-A processes $q_t = \binom{|S_t|}{2} = O(K^2)$ comparisons. All within-round comparison vectors lie in the span of at most $K - 1$ differences between the selected action features. Representing the exact current-round correction in this subspace keeps the rank of the low-rank component of the local geometry at $O(l + K) = O(l)$, so each preconditioned update costs $O(ld)$. The projection in Eq. (6) also admits an $O(ld)$ implementation using a low-rank formulation of the projection procedure of Mhammedi et al. (2019). Thus, the estimator updates require $O(K^2ld)$ time per round. Batch-FD contributes an additional $O(K^2ld)$ amortized time. GeoSRM-D processes one comparison and incurs $O(ld)$ for both estimator updating and sketch maintenance.

Table 1 summarizes these costs. Combining these components and using $K \le n$ gives the amortized per-round computational complexity

$$\mathrm{Time}_{\mathrm{A}} = O\big((Nn^2 + K^2)ld\big) \quad \text{and} \quad \mathrm{Time}_{\mathrm{D}} = O(nld).$$

The Nn^2 and K^2 terms distinguish candidate search from the updates of reward model. In implementations that search candidate minibatches or sampled response pools, N and n are replaced by the corresponding numbers of candidates actually examined. Feature extraction, response generation, preference annotation, and policy optimization are accounted for separately.

C Omitted Proofs for Section 3

In this section, we prove Lemma 1. Appendix C.1 gives the main argument, while Appendix C.2 gives the supporting auxiliary lemmas. We first provide some notations used in our work.

Notations. Let $[n] = \{1, 2, \ldots, n\}$, upper-case bold letters (e.g., $\boldsymbol{A}$) represent matrix and lower-case bold letters (e.g., $\boldsymbol{a}$) represent vectors. We denote by $\|\boldsymbol{A}\|_2$ and $\|\boldsymbol{A}\|_F$ the spectral and Frobenius norms of $\boldsymbol{A}$. We define $|\boldsymbol{A}|$ and $\mathrm{Tr}(\boldsymbol{A})$ as the determinant and trace of matrix $\boldsymbol{A}$. For a positive semi-definite matrix $\boldsymbol{A}$, the matrix norm of vector $\boldsymbol{x}$ is defined by $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$. For two positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we use $\boldsymbol{A} \succeq \boldsymbol{B}$ to represent that $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite. We use $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ to represent the SVD of $\boldsymbol{A}$, where $\boldsymbol{U}, \boldsymbol{V}$ denote the left and right

matrices of singular vectors and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, ..., \sigma_n)$ is the diagonal matrix of singular values in the descending order. We define $\boldsymbol{A}_{[k]} = \boldsymbol{U}_k \mathbf{\Sigma}_k \boldsymbol{V}_k^\top$ for $k \le \mathrm{rank}(\boldsymbol{A})$ as the best rank-k approximation to $\boldsymbol{A}$, where $\boldsymbol{U}_k \in \mathbb{R}^{n\times k}$ and $\boldsymbol{V}_k \in \mathbb{R}^{d\times k}$ are the first k columns of $\boldsymbol{U}$ and $\boldsymbol{V}$.

C.1 Main Proof of Lemma 1

Our proof follows the RB-OMD analysis of Lee et al. (2025). The main technical challenge is to control the propagation of the FD approximation error through sequential rank-broken updates. As in prior analyses, we focus on the small-K regime and impose the following condition to control interactions among pairwise comparisons generated from the same ranking.

Assumption 2. *There exists a known constant $C_{\mathrm{dir}} \ge 0$, fixed before the interaction and independent of d, l, λ, and δ, such that, almost surely, simultaneously for all $t \ge 1$ and $\boldsymbol{v} \in \mathbb{R}^d$, we have $\sum_{1\le p<r\le q_t} |\boldsymbol{z}_{t,p}^\top \boldsymbol{v}|\,|\boldsymbol{z}_{t,r}^\top \boldsymbol{v}| \le C_{\mathrm{dir}} \sum_{p=1}^{q_t} (\boldsymbol{z}_{t,p}^\top \boldsymbol{v})^2$.*

Assumption 2 is used only to control within-ranking interaction terms in the RB-OMD analysis (Lee et al., 2025) and is independent of the sketching mechanism studied in this work. Indeed, the condition always holds with coupling coefficient $C_{\mathrm{dir}} \le (q_t - 1)/2$. When K is treated as a fixed small constant, C_{dir} is therefore also a constant; for example, $K \le 5$ implies $q_t \le 10$ and we may take $C_{\mathrm{dir}} = 9/2$. For pairwise feedback, $q_t = 1$, so the condition holds trivially with $C_{\mathrm{dir}} = 0$.

Consider a single infinite, nonanticipating GeoSRM run on a common probability space. Let $(\mathcal{F}_t)_{t\ge 0}$ be its post-feedback filtration and $\mathcal{G}_t$ the sigma-field after selecting the round-t query but before observing its feedback, so that $\mathcal{F}_{t-1} \subseteq \mathcal{G}_t \subseteq \mathcal{F}_t$. At every $t \ge 1$, the beginning-of-round estimator and sketch are $\mathcal{F}_{t-1}$-measurable. Suppose the query is $\mathcal{G}_t$-measurable, $1 \le q_t < \infty$, and, conditionally on $\mathcal{G}_t$, every queried comparison has its BTL marginal under $\boldsymbol{\theta}^\star$; no conditional independence within a block is required. Fix $1 \le l \le d$, initialize $\widehat{\boldsymbol{\theta}}_1 \in \Theta$ and $\boldsymbol{B}_1 = \mathbf{0}$, and use standard FD throughout the run.

Proof. Recall the ranking-oriented difference $\boldsymbol{z}_{t,p}$ from Eq. (1), and define the curvature row.

$$\boldsymbol{\psi}_{t,p}(\boldsymbol{\theta}) := \sqrt{\dot{\mu}(\boldsymbol{z}_{t,p}^\top \boldsymbol{\theta})}\, \boldsymbol{z}_{t,p}$$

Assumption 1 and Lemma 10 imply $\|\boldsymbol{z}_{t,p}\|_2 \le 2$ and $\|\boldsymbol{\psi}_{t,p}(\boldsymbol{\theta})\|_2 \le 1$. The row inserted into FD is $\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})$, whereas the current-block correction uses $\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})$.

Set $\widehat{\boldsymbol{\mathcal{H}}}_{t,0} := \widehat{\boldsymbol{\mathcal{H}}}_t$ and $\tau = \exp(4\eta/\lambda)$. Our proof proceeds in two stages. We first establish a blockwise OMD recursion that isolates the mismatch between the pre- and post-update curvatures. We then combine this recursion with the sequential-loss bound and the sketched potential bound.

Fix a realized block t and define $\boldsymbol{\delta}_{t,p} := \widehat{\boldsymbol{\theta}}_{t,p} - \widehat{\boldsymbol{\theta}}_{t,p-1}$. We first prove

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_{t+1}} \le \|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t} &+ 2\eta \sum_{p=1}^{q_t} [\ell_{t,p}(\boldsymbol{\theta}^\star) - \ell_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})] \\ &+ 4B\eta e^B(\eta+\tau) C_{\mathrm{int}} \sum_{j=1}^{q_t} \|\boldsymbol{\psi}_{t,j}(\widehat{\boldsymbol{\theta}}_{t,j})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &+ 4B\eta \sum_{p=1}^{q_t} \|\boldsymbol{\delta}_{t,p}\|_2^2 - \sum_{p=1}^{q_t} \|\boldsymbol{\delta}_{t,p}\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_{t,p-1}}.\end{aligned} \tag{12}$$

We derive this recursion in four steps. First, we apply a one-step OMD inequality. Second, we establish an exact blockwise telescoping identity. Third, we bound the resulting curvature boundary. Finally, we invoke the FD metric comparison.

One-step OMD inequality. Fix $p \in [q_t]$. Apply Lemma 6(ii) to the linear objective in the GeoSRM micro-update and expand

$$\widehat{\boldsymbol{\mathcal{H}}}_{t,p} = \widehat{\boldsymbol{\mathcal{H}}}_{t,p-1} + \eta \boldsymbol{H}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1}).$$

To convert the resulting linearized term into a loss difference, apply Lemma 11 with source score $z_{t,p}^\top \widehat{\theta}_{t,p}$ and comparator score $z_{t,p}^\top \theta^\star$. Both scores lie in $[-2B, 2B]$, and $2\eta = \log 2 + 2(B+1)$, so the resulting curvature coefficient is $1/2\eta$.

The remaining discrepancy comes from replacing the post-update gradient by its first-order expansion at the pre-update point. Lemma 10 bounds the absolute third derivative of the scalar logistic loss by $1/4$. Hence Taylor's theorem, $\|z_{t,p}\|_2 \le 2$, and $|z_{t,p}^\top(\widehat{\theta}_{t,p} - \theta^\star)| \le 4B$ give

$$\begin{aligned}&\left|\left\langle g_{t,p}(\widehat{\theta}_{t,p}) - g_{t,p}(\widehat{\theta}_{t,p-1}) - H_{t,p}(\widehat{\theta}_{t,p-1})\delta_{t,p}, \widehat{\theta}_{t,p} - \theta^\star \right\rangle\right| \\ &\quad\le \frac{1}{8}|z_{t,p}^\top \delta_{t,p}|^2 |z_{t,p}^\top(\widehat{\theta}_{t,p} - \theta^\star)| \le 2B\|\delta_{t,p}\|_2^2.\end{aligned}$$

Multiplying this remainder bound by 2η and combining the proximal inequality with the logistic curvature inequality yields

$$\begin{aligned}\|\widehat{\theta}_{t,p} - \theta^\star\|^2_{\widehat{\mathcal{H}}_{t,p-1}} + \|\widehat{\theta}_{t,p} - \theta^\star\|^2_{H_{t,p}(\widehat{\theta}_{t,p})} &\le \|\widehat{\theta}_{t,p-1} - \theta^\star\|^2_{\widehat{\mathcal{H}}_{t,p-1}} + 2\eta[\ell_{t,p}(\theta^\star) - \ell_{t,p}(\widehat{\theta}_{t,p})] \\ &\quad + 4B\eta\|\widehat{\theta}_{t,p} - \widehat{\theta}_{t,p-1}\|_2^2 - \|\widehat{\theta}_{t,p} - \widehat{\theta}_{t,p-1}\|^2_{\widehat{\mathcal{H}}_{t,p-1}}.\end{aligned} \tag{13}$$

Exact block telescoping. The sum of Eq. (13) does not telescope directly. The metric increment uses the pre-update curvature $H_{t,p}(\widehat{\theta}_{t,p-1})$, whereas both the positive curvature term in Eq. (13) and the row retained by FD use the post-update curvature $H_{t,p}(\widehat{\theta}_{t,p})$.

For each $j < q_t$, the contribution of $\eta H_{t,j}(\widehat{\theta}_{t,j-1})$ from the evolving metric telescopes from the intermediate iterate $\widehat{\theta}_{t,j}$ to the terminal iterate $\widehat{\theta}_{t,q_t}$. Combining the resulting endpoint terms with $H_{t,j}(\widehat{\theta}_{t,j})$ leaves the j-th boundary term below. The contribution for $j = q_t$ cancels exactly. Thus,

$$\begin{aligned}&\sum_{p=1}^{q_t}\left[\|\widehat{\theta}_{t,p} - \theta^\star\|^2_{\widehat{\mathcal{H}}_{t,p-1}} - \|\widehat{\theta}_{t,p-1} - \theta^\star\|^2_{\widehat{\mathcal{H}}_{t,p-1}} + \|\widehat{\theta}_{t,p} - \theta^\star\|^2_{H_{t,p}(\widehat{\theta}_{t,p})}\right] \\ &= \|\widehat{\theta}_{t,q_t} - \theta^\star\|^2_{\widehat{\mathcal{H}}_t + \sum_{p=1}^{q_t} H_{t,p}(\widehat{\theta}_{t,p})} - \|\widehat{\theta}_t - \theta^\star\|^2_{\widehat{\mathcal{H}}_t} \\ &\quad + \sum_{p=1}^{q_t-1}\left[\eta\dot{\mu}(z_{t,p}^\top\widehat{\theta}_{t,p-1}) - \dot{\mu}(z_{t,p}^\top\widehat{\theta}_{t,p})\right]\left[(z_{t,p}^\top(\widehat{\theta}_{t,q_t} - \theta^\star))^2 - (z_{t,p}^\top(\widehat{\theta}_{t,p} - \theta^\star))^2\right].\end{aligned} \tag{14}$$

Control boundary. Each summand in the final line of Eq. (14) is the product of a curvature mismatch and a difference of squared errors. Factoring the latter reduces the problem to controlling the curvature mismatch, the sum of the terminal and intermediate errors, and their difference.

For the curvature mismatch, Lemma 6(i), applied to the same micro-update, gives

$$\|\delta_{t,p}\|_{\widehat{\mathcal{H}}_{t,p}} \le \eta\|g_{t,p}(\widehat{\theta}_{t,p-1})\|_{\widehat{\mathcal{H}}_{t,p}^{-1}}.$$

Since $\widehat{\mathcal{H}}_{t,p} \succeq \lambda I_d$, $\|z_{t,p}\|_2 \le 2$, and $\|g_{t,p}(\widehat{\theta}_{t,p-1})\|_2 \le 2$, local-norm Cauchy–Schwarz yields

$$|z_{t,p}^\top\delta_{t,p}| \le \|z_{t,p}\|_{\widehat{\mathcal{H}}_{t,p}^{-1}}\|\delta_{t,p}\|_{\widehat{\mathcal{H}}_{t,p}} \le \eta\|z_{t,p}\|_{\widehat{\mathcal{H}}_{t,p}^{-1}}\|g_{t,p}(\widehat{\theta}_{t,p-1})\|_{\widehat{\mathcal{H}}_{t,p}^{-1}} \le \frac{4\eta}{\lambda}. \tag{15}$$

Combining Eq. (15) with the log-curvature Lipschitz bound in Lemma 10 and the definition of τ gives

$$\tau^{-1}\dot{\mu}(z_{t,p}^\top\widehat{\theta}_{t,p}) \le \dot{\mu}(z_{t,p}^\top\widehat{\theta}_{t,p-1}) \le \tau\dot{\mu}(z_{t,p}^\top\widehat{\theta}_{t,p}).$$

Therefore, we have

$$\left|\eta\dot{\mu}(z_{t,p}^\top\widehat{\theta}_{t,p-1}) - \dot{\mu}(z_{t,p}^\top\widehat{\theta}_{t,p})\right| \le (\eta+\tau)\dot{\mu}(z_{t,p}^\top\widehat{\theta}_{t,p-1}). \tag{16}$$

Because the corresponding curvature rows have the same direction, the same comparison also yields

$$\tau^{-1}\|\psi_{t,p}(\widehat{\theta}_{t,p})\|^2_{\widehat{\mathcal{H}}_t^{-1}} \le \|\psi_{t,p}(\widehat{\theta}_{t,p-1})\|^2_{\widehat{\mathcal{H}}_t^{-1}} \le \tau\|\psi_{t,p}(\widehat{\theta}_{t,p})\|^2_{\widehat{\mathcal{H}}_t^{-1}}. \tag{17}$$

After factoring the difference of squares, boundedness of Θ in Assumption 1 controls the error-sum factor:

$$\left|\boldsymbol{z}_{t,p}^{\top}\big((\widehat{\boldsymbol{\theta}}_{t,q_t}-\boldsymbol{\theta}^{\star})+(\widehat{\boldsymbol{\theta}}_{t,p}-\boldsymbol{\theta}^{\star})\big)\right|\le 8B. \tag{18}$$

It remains to bound the error-difference factor. Direct calculation gives

$$g_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})g_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})^{\top}=\exp(-\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p-1})\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})^{\top}.$$

Together with $|\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p-1}|\le 2B$ and the preceding proximal estimate, we have

$$\begin{aligned}\|\boldsymbol{\delta}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t}\le\|\boldsymbol{\delta}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_{t,p}}&\le\eta\exp\left(-\frac{1}{2}\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p-1}\right)\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_{t,p}^{-1}}\\&\le\eta e^{B}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}},\end{aligned} \tag{19}$$

where the inequalities holds because $\widehat{\boldsymbol{\mathcal{H}}}_{t,p}\succeq\widehat{\boldsymbol{\mathcal{H}}}_t\succ\mathbf{0}$. Applying Eq. (19) along the within-block path $\widehat{\boldsymbol{\theta}}_{t,q_t}-\widehat{\boldsymbol{\theta}}_{t,p}=\sum_{r=p+1}^{q_t}\boldsymbol{\delta}_{t,r}$ yields

$$|\boldsymbol{z}_{t,p}^{\top}(\widehat{\boldsymbol{\theta}}_{t,q_t}-\widehat{\boldsymbol{\theta}}_{t,p})|\le\eta e^{B}\|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\sum_{r=p+1}^{q_t}\|\boldsymbol{\psi}_{t,r}(\widehat{\boldsymbol{\theta}}_{t,r-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}. \tag{20}$$

We now combine the three bounds as in Eq. (16), Eq. (18) and Eq. (20). First,

$$\dot{\mu}(\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p-1})\|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}=\sqrt{\dot{\mu}(\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p-1})}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\le\frac{1}{2}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

Consequently, factoring the difference of squares gives, for every $p<q_t$,

$$\begin{aligned}&\left|\left[\eta\dot{\mu}(\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p-1})-\dot{\mu}(\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p})\right]\left[(\boldsymbol{z}_{t,p}^{\top}(\widehat{\boldsymbol{\theta}}_{t,q_t}-\boldsymbol{\theta}^{\star}))^2-(\boldsymbol{z}_{t,p}^{\top}(\widehat{\boldsymbol{\theta}}_{t,p}-\boldsymbol{\theta}^{\star}))^2\right]\right|\\&\quad\le 4B\eta e^{B}(\eta+\tau)\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\sum_{r=p+1}^{q_t}\|\boldsymbol{\psi}_{t,r}(\widehat{\boldsymbol{\theta}}_{t,r-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.\end{aligned}$$

Summing this inequality over p and applying Lemma 2 yields

$$\begin{aligned}&\left|\sum_{p=1}^{q_t-1}\left[\eta\dot{\mu}(\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p-1})-\dot{\mu}(\boldsymbol{z}_{t,p}^{\top}\widehat{\boldsymbol{\theta}}_{t,p})\right]\left[(\boldsymbol{z}_{t,p}^{\top}(\widehat{\boldsymbol{\theta}}_{t,q_t}-\boldsymbol{\theta}^{\star}))^2-(\boldsymbol{z}_{t,p}^{\top}(\widehat{\boldsymbol{\theta}}_{t,p}-\boldsymbol{\theta}^{\star}))^2\right]\right|\\&\quad\le 4B\eta e^{B}(\eta+\tau)\sum_{1\le p<r\le q_t}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\|\boldsymbol{\psi}_{t,r}(\widehat{\boldsymbol{\theta}}_{t,r-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\\&\quad\le 4B\eta e^{B}(\eta+\tau)C_{\mathrm{int}}\sum_{j=1}^{q_t}\|\boldsymbol{\psi}_{t,j}(\widehat{\boldsymbol{\theta}}_{t,j})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}^2.\end{aligned} \tag{21}$$

Transfer to the sketched term. Sum Eq. (13) over p and substitute the exact identity Eq. (14). Bounding the negative boundary contribution by its absolute value and applying Eq. (21) gives an inequality whose right-hand side is that of Eq. (12) and whose left-hand side is

$$\|\widehat{\boldsymbol{\theta}}_{t,q_t}-\boldsymbol{\theta}^{\star}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t+\sum_{p=1}^{q_t}\boldsymbol{H}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})}^2.$$

Moreover, $\widehat{\boldsymbol{\theta}}_{t+1}=\widehat{\boldsymbol{\theta}}_{t,q_t}$, and the one-sided FD update gives $\widehat{\boldsymbol{\mathcal{H}}}_{t+1}\preceq\widehat{\boldsymbol{\mathcal{H}}}_t+\sum_{p=1}^{q_t}\boldsymbol{H}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})$. Monotonicity of the squared local norm therefore implies

$$\|\widehat{\boldsymbol{\theta}}_{t+1}-\boldsymbol{\theta}^{\star}\|_{\widehat{\boldsymbol{\mathcal{H}}}_{t+1}}^2\le\|\widehat{\boldsymbol{\theta}}_{t,q_t}-\boldsymbol{\theta}^{\star}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t+\sum_{p=1}^{q_t}\boldsymbol{H}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})}^2,$$

which proves the blockwise OMD recursion in Eq. (12).

Fix an arbitrary integer $n \geq 1$, sum Eq. (12) over $t = 1, \ldots, n$, and substitute the sequential-loss bound in Lemma 4. Denote $\kappa_B = 2 + 2\cosh(2B)$ and $C_{\text{int}} = \pi\kappa_B C_{\text{dir}}/8$, this gives

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}_{n+1} - \boldsymbol{\theta}^\star\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_{n+1}} &\leq \|\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^\star\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_1} + \frac{\eta\kappa_B(1+2C_{\text{dir}})}{2}\log\frac{1}{\delta} \\
&\quad + 2\eta e^B[\eta\sqrt{\kappa_B}(\tau + C_{\text{int}}) + 2B(\eta+\tau)C_{\text{int}}]\sum_{t=1}^{n}\sum_{p=1}^{q_t}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\
&\quad + \sum_{t=1}^{n}\sum_{p=1}^{q_t}\left[4B\eta\|\boldsymbol{\delta}_{t,p}\|_2^2 - \|\boldsymbol{\delta}_{t,p}\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_{t,p-1}}\right], \qquad (22) \\
&\leq \|\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^\star\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_1} + \frac{\eta\kappa_B(1+2C_{\text{dir}})}{2}\log\frac{1}{\delta} \\
&\quad + 2\eta e^B[\eta\sqrt{\kappa_B}(\tau + C_{\text{int}}) + 2B(\eta+\tau)C_{\text{int}}]\sum_{t=1}^{n}\sum_{p=1}^{q_t}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}
\end{aligned}$$

where the second inequality holds because $\widehat{\boldsymbol{\mathcal{H}}}_{t,p-1} \succeq \widehat{\boldsymbol{\mathcal{H}}}_t \succeq \lambda \boldsymbol{I}_d$ and $\lambda \geq 4B\eta$.

Finally, $\boldsymbol{B}_1 = \boldsymbol{0}$ implies $\widehat{\boldsymbol{\mathcal{H}}}_1 = \lambda \boldsymbol{I}_d$. Since $\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}^\star \in \Theta$,

$$\|\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^\star\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_1} = \lambda\|\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^\star\|_2^2 \leq 4\lambda B^2.$$

Lemma 3 further gives

$$\sum_{t=1}^{n}\sum_{p=1}^{q_t}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq 2l\log\left(1 + \frac{\sum_{s<t} q_s}{l\lambda}\right) + \frac{2d}{\lambda}\sum_{s<t}\sum_{p=1}^{q_s}\rho_{s,p} := \Gamma_t,$$

where $\rho_{t,p} \geq 0$ is the squared-singular-value shrinkage performed by FD (Algorithm 2) for insertion (t,p). Because the sequential-loss event holds simultaneously for every n, the conclusion holds simultaneously for all integers $t \geq 1$. Substituting these bounds to Eq. (22) yields

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t} &\leq 4\lambda B^2 + \frac{\eta\kappa_B(1+2C_{\text{dir}})}{2}\log\frac{1}{\delta} \\
&\quad + 4\eta e^B[\eta\sqrt{\kappa_B}(\tau + C_{\text{int}}) + 2B(\eta+\tau)C_{\text{int}}]\,\Gamma_t \\
&\lesssim 4\lambda B^2 + \log\frac{1}{\delta} + l\log\left(1 + \frac{tK^2}{l\lambda}\right) + \frac{d}{\lambda}\Delta_l \\
&= O\left(l\log\left(\frac{tK}{\delta}\right) + d\Delta_l\right) := \widehat{\beta}_t(\delta)^2,
\end{aligned}$$

where we simplify the confidence bound by ignoring the constants. □

C.2 Proofs of the Auxiliary Lemmas

This section presents several supporting lemmas used in the proof of Appendix C.1

Lemma 2. *Under Assumption 2 and the condition $\lambda \geq 1 + 2C_{\text{int}}$, every realized block satisfies*

$$\sum_{1\leq p<s\leq q_t}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\|\boldsymbol{\psi}_{t,s}(\widehat{\boldsymbol{\theta}}_{t,s-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq C_{\text{int}}\sum_{j=1}^{q_t}\|\boldsymbol{\psi}_{t,j}(\widehat{\boldsymbol{\theta}}_{t,j})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

Moreover, the total post-update block leverage satisfies $\sum_{p=1}^{q_t}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq 1$.

Proof. Fix a realized block and define

$$u_p := \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}, \qquad v_p := \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

Applying Assumption 2 with $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \widehat{\boldsymbol{\mathcal{H}}}_t^{-1})$ and taking expectations give

$$\sum_{p<s}\|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\|\boldsymbol{z}_{t,s}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq \frac{\pi}{2}C_{\text{dir}}\sum_{p=1}^{q_t}\|\boldsymbol{z}_{t,p}\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}},$$

where we used $\mathbb{E}|\boldsymbol{z}_{t,p}^\top \boldsymbol{g}\, \boldsymbol{z}_{t,s}^\top \boldsymbol{g}| \geq (2/\pi)\|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \|\boldsymbol{z}_{t,s}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}$ in Lemma 9. Since all iterates lie in Θ, $|\boldsymbol{z}_{t,p}^\top \widehat{\boldsymbol{\theta}}_{t,p-1}| \leq 2B$ and $|\boldsymbol{z}_{t,p}^\top \widehat{\boldsymbol{\theta}}_{t,p}| \leq 2B$. Hence,

$$u_p \leq \frac{\|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}}{2}, \qquad v_p \leq \frac{\|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}}{2}, \qquad \|\boldsymbol{z}_{t,p}\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq \kappa_B v_p^2.$$

Therefore, for either $w_p = u_p$ or $w_p = v_p$,

$$\sum_{p<s} w_p w_s \leq \frac{1}{4} \sum_{p<s} \|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \|\boldsymbol{z}_{t,s}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq \frac{\pi \kappa_B}{8} C_{\mathrm{dir}} \sum_{p=1}^{q_t} v_p^2 = C_{\mathrm{int}} \sum_{p=1}^{q_t} v_p^2.$$

Taking $w_p = u_p$ proves the first argument. Taking $w_p = v_p$ gives

$$\left(\sum_{p=1}^{q_t} v_p\right)^2 = \sum_{p=1}^{q_t} v_p^2 + 2 \sum_{p<s} v_p v_s \leq (1 + 2C_{\mathrm{int}}) \sum_{p=1}^{q_t} v_p^2.$$

Let $L_t := \sum_{p=1}^{q_t} v_p^2$, we have

$$L_t^2 \leq \left(\max_{p \in [q_t]} v_p^2\right) \left(\sum_{p=1}^{q_t} v_p\right)^2 \leq \frac{1 + 2C_{\mathrm{int}}}{\lambda} L_t,$$

where the second inequality holds because $\widehat{\boldsymbol{\mathcal{H}}}_t \succeq \lambda \boldsymbol{I}_d$ and $\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|_2 \leq 1$. Therefore, the bound holds trivially for $L_t = 0$. For $L_t > 0$, division by L_t and $\lambda \geq 1 + 2C_{\mathrm{int}}$ finish the proof. □

We next establish a sketched analogue of the standard elliptical potential lemma (Abbasi-Yadkori et al., 2011), which relates the accumulated frozen-block potential to the FD approximation error.

Lemma 3 (Sketched frozen-block potential)**.** *For every integer $n \geq 1$,*

$$\sum_{t=1}^{n} \sum_{p=1}^{q_t} \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq l \log\left(1 + \frac{\sum_{s<t} q_s}{l\lambda}\right) + \frac{2d}{\lambda} \sum_{s<t} \sum_{p=1}^{q_s} \rho_{s,p} := \Gamma_t, \tag{23}$$

where $\rho_{t,p} \geq 0$ is the squared singular-value shrinkage performed by FD for insertion (t,p). Moreover, for every integer $t \geq 1$,

$$\sum_{s<t} \sum_{p=1}^{q_s} \rho_{s,p} \leq \min_{0 \leq k < l} \frac{\|\boldsymbol{A}_t - (\boldsymbol{A}_t)_{[k]}\|_F^2}{l - k}, \tag{24}$$

where the matrix $\boldsymbol{A}_t$ contains the rows $\boldsymbol{\psi}_{i,p}(\widehat{\boldsymbol{\theta}}_{i,p})^\top$ for $i < t$ and $p \in [q_i]$, so that $\boldsymbol{A}_t^\top \boldsymbol{A}_t = \sum_{i=1}^{t-1} \sum_{p=1}^{q_i} \boldsymbol{H}_{i,p}(\widehat{\boldsymbol{\theta}}_{i,p})$.

Proof. For the potential bound, define $\boldsymbol{C}_t := \sum_{p=1}^{q_t} \boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p}) \boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})^\top$. Lemma 2 gives

$$\mathrm{Tr}(\widehat{\boldsymbol{\mathcal{H}}}_t^{-1} \boldsymbol{C}_t) = \sum_{p=1}^{q_t} \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq 1.$$

Let $\zeta_1, \ldots, \zeta_d$ be the eigenvalues of $\widehat{\boldsymbol{\mathcal{H}}}_t^{-1/2} \boldsymbol{C}_t \widehat{\boldsymbol{\mathcal{H}}}_t^{-1/2}$. This matrix is positive semidefinite and has trace at most one, so $0 \leq \zeta_i \leq 1$ for every i. Using $\zeta_i \leq 2\log(1 + \zeta_i)$ therefore yields

$$\sum_{p=1}^{q_t} \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq 2 \log \frac{\det(\widehat{\boldsymbol{\mathcal{H}}}_t + \boldsymbol{C}_t)}{\det(\widehat{\boldsymbol{\mathcal{H}}}_t)}.$$

Let $\boldsymbol{E}_t := \widehat{\boldsymbol{\mathcal{H}}}_t + \boldsymbol{C}_t - \widehat{\boldsymbol{\mathcal{H}}}_{t+1}$, summing the FD residuals generated within block t gives

$$\boldsymbol{0} \preceq \boldsymbol{E}_t \preceq \left(\sum_{p=1}^{q_t} \rho_{t,p}\right) \boldsymbol{I}_d.$$

Since $\widehat{\boldsymbol{\mathcal{H}}}_{t+1} \succeq \lambda \boldsymbol{I}_d$, applying $\log\det(\boldsymbol{I}_d + \boldsymbol{M}) \le \mathrm{Tr}(\boldsymbol{M})$ to $\boldsymbol{M} = \widehat{\boldsymbol{\mathcal{H}}}_{t+1}^{-1/2} \boldsymbol{E}_t \widehat{\boldsymbol{\mathcal{H}}}_{t+1}^{-1/2}$ gives

$$\log \frac{\det(\widehat{\boldsymbol{\mathcal{H}}}_t + \boldsymbol{C}_t)}{\det(\widehat{\boldsymbol{\mathcal{H}}}_t)} \le \log \frac{\det(\widehat{\boldsymbol{\mathcal{H}}}_{t+1})}{\det(\widehat{\boldsymbol{\mathcal{H}}}_t)} + \frac{d}{\lambda} \sum_{p=1}^{q_t} \rho_{t,p}.$$

Summing this inequality over $t = 1, \dots, n$ telescopes the determinant terms. Each inserted row has Euclidean norm at most one, and FD compression can only decrease the squared Frobenius norm. Hence

$$\mathrm{Tr}(\boldsymbol{B}_{n+1}^\top \boldsymbol{B}_{n+1}) \le \sum_{t=1}^{n} q_t.$$

Together with $\mathrm{rank}(\boldsymbol{B}_{n+1}) \le l$, the AM–GM inequality gives

$$\log \frac{\det(\widehat{\boldsymbol{\mathcal{H}}}_{n+1})}{\det(\lambda \boldsymbol{I}_d)} \le l \log\left(1 + \frac{\sum_{t=1}^{n} q_t}{l\lambda}\right).$$

Combining the preceding inequalities proves Eq. (23).

Finally, $\boldsymbol{A}_t$ is precisely the chronological row stream summarized by $\boldsymbol{B}_t$. Applying Lemma 5 to this stream gives

$$\sum_{i=1}^{t-1} \sum_{p=1}^{q_i} \rho_{i,p} \le \min_{0 \le k < l} \frac{\|\boldsymbol{A}_t - (\boldsymbol{A}_t)_{[k]}\|_F^2}{l - k},$$

which proves Eq. (24) and Lemma 1. □

Lemma 4. *Fix $\delta \in (0,1)$. With probability at least $1-\delta$, simultaneously for every integer $n \ge 1$,*

$$2\eta \sum_{t=1}^{n} \sum_{p=1}^{q_t} [\ell_{t,p}(\boldsymbol{\theta}^\star) - \ell_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})] \le \frac{\eta \kappa_B (1 + 2C_{\mathrm{dir}})}{2} \log \frac{1}{\delta} + 2\eta^2 e^B \sqrt{\kappa_B}(\tau + C_{\mathrm{int}}) \sum_{t=1}^{n} \sum_{p=1}^{q_t} \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

No within-block independence is required.

Proof. For concentration, fix an outcome-independent orientation of the q_t comparison positions before observing the round-t feedback. Let $\boldsymbol{d}_{t,p}$ be the corresponding feature difference and $Y_{t,p}$ the indicator that its first action is preferred. Thus $\boldsymbol{d}_{t,p}$ is $\mathcal{G}_t$-measurable and agrees with the ranking-oriented $\boldsymbol{z}_{t,p}$ up to permutation and sign after the ranking is observed.

Set $a_{t,p} := \boldsymbol{d}_{t,p}^\top(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star)$, by the conditional BTL marginals,

$$\mathbb{E}\left[\sum_{p=1}^{q_t} (Y_{t,p} - \mu(\boldsymbol{d}_{t,p}^\top \boldsymbol{\theta}^\star)) a_{t,p} \,\middle|\, \mathcal{G}_t\right] = 0.$$

Logistic curvature gives

$$\sum_{p=1}^{q_t} [\ell_{t,p}(\boldsymbol{\theta}^\star) - \ell_{t,p}(\widehat{\boldsymbol{\theta}}_t)] \le \sum_{p=1}^{q_t} (Y_{t,p} - \mu(\boldsymbol{d}_{t,p}^\top \boldsymbol{\theta}^\star)) a_{t,p} - \frac{1}{2\kappa_B} \sum_{p=1}^{q_t} a_{t,p}^2.$$

Moreover, Assumption 2 implies

$$\left(\sum_{p=1}^{q_t} |a_{t,p}|\right)^2 \le (1 + 2C_{\mathrm{dir}}) \sum_{p=1}^{q_t} a_{t,p}^2.$$

Hence Hoeffding's lemma yields, for every $\gamma > 0$,

$$\mathbb{E}\left[\exp\left\{\gamma \sum_{p=1}^{q_t} [\ell_{t,p}(\boldsymbol{\theta}^\star) - \ell_{t,p}(\widehat{\boldsymbol{\theta}}_t)]\right\} \,\middle|\, \mathcal{G}_t\right] \le \exp\left\{\left(\frac{\gamma^2(1 + 2C_{\mathrm{dir}})}{8} - \frac{\gamma}{2\kappa_B}\right) \sum_{p=1}^{q_t} a_{t,p}^2\right\}.$$

Taking $\gamma = 4/[\kappa_B(1 + 2C_{\mathrm{dir}})]$, the exponential of the cumulative frozen-block loss is therefore a nonnegative supermartingale with respect to $(\mathcal{F}_t)_{t\geq 0}$. Ville's inequality in Lemma 8 gives, with probability at least $1 - \delta$, simultaneously for all $n \geq 1$,

$$\sum_{t=1}^{n}\sum_{p=1}^{q_t}[\ell_{t,p}(\boldsymbol{\theta}^\star) - \ell_{t,p}(\widehat{\boldsymbol{\theta}}_t)] \leq \frac{\kappa_B(1 + 2C_{\mathrm{dir}})}{4}\log\frac{1}{\delta}.$$

It remains to control the within-block updates. Since $(\boldsymbol{d}, Y)$ and $(-\boldsymbol{d}, 1 - Y)$ induce the same logistic loss, after the ranking is observed we may return to the algorithmic order $\boldsymbol{z}_{t,1}, \ldots, \boldsymbol{z}_{t,q_t}$. For each $p \in [q_t]$, the one-Lipschitz property of the logistic loss and $\widehat{\boldsymbol{\theta}}_{t,p} - \widehat{\boldsymbol{\theta}}_t = \sum_{r=1}^{p}\boldsymbol{\delta}_{t,r}$ give

$$\begin{aligned}
\ell_{t,p}(\widehat{\boldsymbol{\theta}}_t) - \ell_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p}) &\leq |\boldsymbol{z}_{t,p}^\top(\widehat{\boldsymbol{\theta}}_t - \widehat{\boldsymbol{\theta}}_{t,p})| \\
&\leq \|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\sum_{r=1}^{p}\|\boldsymbol{\delta}_{t,r}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t} \\
&\leq \eta e^B\|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\sum_{r=1}^{p}\|\boldsymbol{\psi}_{t,r}(\widehat{\boldsymbol{\theta}}_{t,r-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}},
\end{aligned}$$

where the last inequality follows from Eq. (19). Moreover, boundedness of the score and Lemma 10(i) imply

$$\dot{\mu}(\boldsymbol{z}_{t,p}^\top\widehat{\boldsymbol{\theta}}_{t,p-1}) \geq \kappa_B^{-1}, \qquad \|\boldsymbol{z}_{t,p}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq \sqrt{\kappa_B}\,\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

Therefore, summing over p and separating the diagonal and cross terms yields

$$\begin{aligned}
\sum_{p=1}^{q_t}[\ell_{t,p}(\widehat{\boldsymbol{\theta}}_t) - \ell_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})] &\leq \eta e^B\sqrt{\kappa_B}\sum_{p=1}^{q_t}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\
&\quad + \eta e^B\sqrt{\kappa_B}\sum_{r<p}\|\boldsymbol{\psi}_{t,r}(\widehat{\boldsymbol{\theta}}_{t,r-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p-1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\
&\leq \eta e^B\sqrt{\kappa_B}(\tau + C_{\mathrm{int}})\sum_{p=1}^{q_t}\|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}},
\end{aligned}$$

where the second inequality holds by Eq. (17) and Lemma 2. □

D Omitted Proofs for Section 4

We provide the proofs omitted in Section 4. Appendix D.1 establishes the suboptimality-gap bound for GeoSRM-A and Appendix D.2 establishes the regret bound for GeoSRM-D.

D.1 Proof of Theorem 1

After T feedback rounds, Algorithm 4 produces $\widehat{\boldsymbol{\theta}}_{T+1}$. Define the returned policy as $\widehat{\pi}_{T+1}(\boldsymbol{x}) = \arg\max_{\boldsymbol{a}\in\mathcal{A}}\phi(\boldsymbol{x}, \boldsymbol{a})^\top\widehat{\boldsymbol{\theta}}_{T+1}$ With fixed measurable tie-breaking, the round-start estimator and metric make every active query and its size $\mathcal{G}_t$-measurable, before the ranking feedback is observed.

Proof. Work on the anytime confidence event. We first transfer the terminal confidence bound to every historical metric. Let $\varepsilon_t = \sum_{s<t}\sum_{p=1}^{q_s}\rho_{s,p}/\lambda$ and $\boldsymbol{A}_t^\top\boldsymbol{A}_t = \sum_{i=1}^{t-1}\sum_{j=1}^{q_i}\boldsymbol{H}_{i,j}(\widehat{\boldsymbol{\theta}}_{i,j})$. After processing round T, the one-sided FD residual bound gives

$$\mathbf{0} \preceq \boldsymbol{A}_{T+1}^\top\boldsymbol{A}_{T+1} - \boldsymbol{B}_{T+1}^\top\boldsymbol{B}_{T+1} \preceq \lambda\varepsilon_{T+1}\boldsymbol{I}_d \preceq \varepsilon_{T+1}\widehat{\boldsymbol{\mathcal{H}}}_{T+1},$$

where the final inequality uses $\widehat{\boldsymbol{\mathcal{H}}}_{T+1} \succeq \lambda\boldsymbol{I}_d$ and $\varepsilon_{T+1} \geq 0$. For every $t \leq T$, the FD under-approximation, monotonicity of the exact curvature stream, and the terminal residual bound imply

$$\widehat{\boldsymbol{\mathcal{H}}}_t \preceq \lambda\boldsymbol{I}_d + \boldsymbol{A}_t^\top\boldsymbol{A}_t \preceq \lambda\boldsymbol{I}_d + \boldsymbol{A}_{T+1}^\top\boldsymbol{A}_{T+1} \preceq (1 + \varepsilon_{T+1})\widehat{\boldsymbol{\mathcal{H}}}_{T+1}.$$

All matrices in this chain are positive definite because they dominate $\lambda \boldsymbol{I}_d$. Hence, for every $\boldsymbol{h} \in \mathbb{R}^d$ and $t \le T$,

$$\begin{aligned} |\boldsymbol{h}^\top(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1})| &\le \|\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t} \|\boldsymbol{h}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &\le \sqrt{1+\varepsilon_{T+1}}\, \|\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1}\|_{\widehat{\boldsymbol{\mathcal{H}}}_{T+1}} \|\boldsymbol{h}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &\le \sqrt{1+\varepsilon_{T+1}}\, \widehat{\beta}_{T+1} \|\boldsymbol{h}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}. \end{aligned} \tag{25}$$

The first inequality is Cauchy–Schwarz in the local norm, the second follows from the preceding matrix comparison, and the last uses the terminal confidence bound. Recall from Eq. (8) that

$$f_t(\boldsymbol{x}, S) := \frac{1}{2|S|} \sum_{\boldsymbol{a},\boldsymbol{a}' \in S} \dot{\mu}\big([\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')]^\top \widehat{\boldsymbol{\theta}}_t\big) \|\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

Both $\dot{\mu}$ and the squared local norm are invariant under reversing the comparison orientation. Thus, for any two distinct actions $\boldsymbol{a}, \boldsymbol{a}'$,

$$f_t(\boldsymbol{x}, \{\boldsymbol{a}, \boldsymbol{a}'\}) = \frac{1}{2} \dot{\mu}\big([\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')]^\top \widehat{\boldsymbol{\theta}}_t\big) \|\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

The best seed dominates this candidate pair, each accepted expansion does not decrease the score, and the final context maximization dominates every other context. Therefore, for every $\boldsymbol{x} \in \mathcal{X}$ and every pair of distinct actions $\boldsymbol{a}, \boldsymbol{a}' \in \mathcal{A}$,

$$f_t(\boldsymbol{x}_t, S_t) \ge \frac{1}{2} \dot{\mu}\big([\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')]^\top \widehat{\boldsymbol{\theta}}_t\big) \|\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}. \tag{26}$$

For every selected pair, $\widehat{\boldsymbol{\theta}}_t, \widehat{\boldsymbol{\theta}}_{t,p} \in \Theta$ and $\|\boldsymbol{z}_{t,p}\|_2 \le 2$, so $|\boldsymbol{z}_{t,p}^\top(\widehat{\boldsymbol{\theta}}_t - \widehat{\boldsymbol{\theta}}_{t,p})| \le 4B$. The log-curvature Lipschitz inequality in Lemma 10 consequently gives $\dot{\mu}(\boldsymbol{z}_{t,p}^\top \widehat{\boldsymbol{\theta}}_t) \le e^{4B} \dot{\mu}(\boldsymbol{z}_{t,p}^\top \widehat{\boldsymbol{\theta}}_{t,p})$. Because the realized comparisons enumerate all unordered pairs in S_t, orientation invariance yields

$$f_t(\boldsymbol{x}_t, S_t) = \frac{1}{|S_t|} \sum_{p=1}^{q_t} \dot{\mu}(\boldsymbol{z}_{t,p}^\top \widehat{\boldsymbol{\theta}}_t) \|\boldsymbol{z}_{t,p}\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \le \frac{e^{4B}}{|S_t|} \sum_{p=1}^{q_t} \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}. \tag{27}$$

Moreover, $\|\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')\|_2 \le 2$ and $\widehat{\boldsymbol{\theta}}_t \in \Theta$, so $|[\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')]^\top \widehat{\boldsymbol{\theta}}_t| \le 2B$. The curvature term in Eq. (26) is therefore at least $1/\kappa_B$. Combining Eq. (26) and Eq. (27) gives

$$\|\phi(\boldsymbol{x},\boldsymbol{a}) - \phi(\boldsymbol{x},\boldsymbol{a}')\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \le \frac{2\kappa_B e^{4B}}{|S_t|} \sum_{p=1}^{q_t} \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}. \tag{28}$$

Fix $\boldsymbol{x} \in \mathcal{X}$. If $\pi^\star(\boldsymbol{x}) = \widehat{\pi}_{T+1}(\boldsymbol{x})$, the reward gap is zero. Otherwise, the greedy definition of final policy $\widehat{\pi}_{T+1}(\boldsymbol{x}) = \arg\max_{a \in \mathcal{A}} \phi(\boldsymbol{x},\boldsymbol{a})^\top \widehat{\boldsymbol{\theta}}_{T+1}$ gives

$$[\phi(\boldsymbol{x}, \pi^\star(\boldsymbol{x})) - \phi(\boldsymbol{x}, \widehat{\pi}_{T+1}(\boldsymbol{x}))]^\top \widehat{\boldsymbol{\theta}}_{T+1} \le 0.$$

Therefore, for every $t \le T$, Eq. (25) and Eq. (28) imply

$$\begin{aligned} r(\boldsymbol{x}, \pi^\star(\boldsymbol{x})) - r(\boldsymbol{x}, \widehat{\pi}_{T+1}(\boldsymbol{x})) &= [\phi(\boldsymbol{x}, \pi^\star(\boldsymbol{x})) - \phi(\boldsymbol{x}, \widehat{\pi}_{T+1}(\boldsymbol{x}))]^\top \boldsymbol{\theta}^\star \\ &\le \big|[\phi(\boldsymbol{x}, \pi^\star(\boldsymbol{x})) - \phi(\boldsymbol{x}, \widehat{\pi}_{T+1}(\boldsymbol{x}))]^\top (\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1})\big| \\ &\le \sqrt{1+\varepsilon_{T+1}}\, \widehat{\beta}_{T+1} \|\phi(\boldsymbol{x}, \pi^\star(\boldsymbol{x})) - \phi(\boldsymbol{x}, \widehat{\pi}_{T+1}(\boldsymbol{x}))\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &\le e^{2B} \sqrt{\frac{2(1+\varepsilon_{T+1})\kappa_B}{|S_t|}}\, \widehat{\beta}_{T+1} \sqrt{\sum_{p=1}^{q_t} \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}}. \end{aligned}$$

Averaging this bound over $t = 1, \dots, T$ and applying Cauchy–Schwarz across rounds yields

$$\begin{aligned} & r(\boldsymbol{x}, \pi^\star(\boldsymbol{x})) - r(\boldsymbol{x}, \widehat{\pi}_{T+1}(\boldsymbol{x})) \\ &\le \frac{e^{2B}\sqrt{2(1+\varepsilon_{T+1})\kappa_B}\, \widehat{\beta}_{T+1}}{T} \sqrt{\left(\sum_{t=1}^{T} \sum_{p=1}^{q_t} \|\boldsymbol{\psi}_{t,p}(\widehat{\boldsymbol{\theta}}_{t,p})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}\right)\left(\sum_{t=1}^{T} \frac{1}{|S_t|}\right)} \\ &\le \frac{2e^{2B}\sqrt{(1+\varepsilon_{T+1})\kappa_B}\, \widehat{\beta}_{T+1}}{T} \sqrt{\Gamma_{T+1} \sum_{t=1}^{T} \frac{1}{|S_t|}}, \end{aligned}$$

where the second inequality follows from Lemma 3. All preceding inequalities hold deterministically for every $\boldsymbol{x}$ on the same anytime confidence event, so taking the supremum requires no additional union bound. Consequently, apply Lemma 1, we have

$$\begin{aligned}\mathrm{SubOpt}_{T+1} &= \sup_{\boldsymbol{x}\in\mathcal{X}}\left[r(\boldsymbol{x},\pi^\star(\boldsymbol{x})) - r(\boldsymbol{x},\widehat{\pi}_{T+1}(\boldsymbol{x}))\right] \\ &\le \frac{2e^{2B}\sqrt{(1+\Delta_l/\lambda)\kappa_B}\,\widehat{\beta}_{T+1}}{T}\sqrt{\left(l\log\left(1+\frac{\sum_{s<T}q_s}{l\lambda}\right)+\frac{2d\Delta_l}{\lambda}\right)\sum_{t=1}^{T}\frac{1}{|S_t|}} \\ &\lesssim \widetilde{O}\left(\frac{l+d\Delta_l}{T}\sqrt{\sum_{t=1}^{T}\frac{1+\Delta_l}{|S_t|}}\right),\end{aligned}$$

which completes the proof. □

D.2 Proof of Theorem 2

The result below uses an ordered two-slot query $(\boldsymbol{a}_t, \boldsymbol{a}'_t) \in \mathcal{A}^2$, selected before feedback and hence $\mathcal{G}_t$-measurable. Distinct slots use the original BTL feedback. If the slots coincide, the label is conditionally $\mathrm{Bernoulli}(1/2)$; the comparison row is zero and both updates are no-ops. This preserves the conditional-BTL hypothesis of Lemma 1.

Proof. Work on the confidence event. By Cauchy–Schwarz in the local norm, for every $\boldsymbol{h} \in \mathbb{R}^d$,

$$|\boldsymbol{h}^\top(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_t)| \le \|\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_t\|_{\widehat{\boldsymbol{\mathcal{H}}}_t}\|\boldsymbol{h}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \le \widehat{\beta}_t\|\boldsymbol{h}\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

Applying this inequality with $\boldsymbol{h} = \phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)$, we first analyze the instantaneous regret at round t. According to the selection rule of $\boldsymbol{a}_t$ in Eq. (10) gives

$$\begin{aligned}r(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - r(\boldsymbol{x}_t,\boldsymbol{a}_t) &= \left[\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\right]^\top\boldsymbol{\theta}^\star \\ &\le \left[\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\right]^\top\widehat{\boldsymbol{\theta}}_t \\ &\quad + \widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &\le \left[\phi(\boldsymbol{x}_t,\boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\right]^\top\widehat{\boldsymbol{\theta}}_t + \widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &\le \widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.\end{aligned} \tag{29}$$

The second inequality follows from the optimistic choice of $\boldsymbol{a}'_t$ in Eq. (10). The final inequality uses the greedy choice of $\boldsymbol{a}_t$, which implies $[\phi(\boldsymbol{x}_t,\boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)]^\top\widehat{\boldsymbol{\theta}}_t \le 0$. Applying the confidence inequality to $\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}'_t)$ gives

$$\begin{aligned}&r(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - r(\boldsymbol{x}_t,\boldsymbol{a}'_t) \\ &\le \left[\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}'_t)\right]^\top\widehat{\boldsymbol{\theta}}_t + \widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}'_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &\le \widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} - \widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &\quad + \widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}'_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\ &\le 2\widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}},\end{aligned} \tag{30}$$

where the second inequality follows by rearranging the optimism comparison in Eq. (10). The final inequality uses

$$\|\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}'_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \le \|\phi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} + \|\phi(\boldsymbol{x}_t,\boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

Averaging Eq. (29) and Eq. (30) yields

$$r(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - \frac{1}{2}\{r(\boldsymbol{x}_t,\boldsymbol{a}_t) + r(\boldsymbol{x}_t,\boldsymbol{a}'_t)\} \le \frac{3}{2}\widehat{\beta}_t\|\phi(\boldsymbol{x}_t,\boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t,\boldsymbol{a}_t)\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}. \tag{31}$$

The feedback-oriented difference $\boldsymbol{z}_{t,1}$ agrees with $\phi(\boldsymbol{x}_t, \boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t, \boldsymbol{a}_t)$ up to sign, which does not affect the local norm. Since $\boldsymbol{\psi}_{t,1}(\widehat{\boldsymbol{\theta}}_{t,1}) = \sqrt{\dot{\mu}(\boldsymbol{z}_{t,1}^\top \widehat{\boldsymbol{\theta}}_{t,1})}\, \boldsymbol{z}_{t,1}$ and $\dot{\mu}(\boldsymbol{z}_{t,1}^\top \widehat{\boldsymbol{\theta}}_{t,1}) \geq 1/\kappa_B$, we have

$$\|\phi(\boldsymbol{x}_t, \boldsymbol{a}'_t) - \phi(\boldsymbol{x}_t, \boldsymbol{a}_t)\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq \kappa_B \|\boldsymbol{\psi}_{t,1}(\widehat{\boldsymbol{\theta}}_{t,1})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}.$$

Summing Eq. (31), using the monotonicity of $\widehat{\beta}_t$, and applying Cauchy–Schwarz gives

$$\begin{aligned}
\mathrm{Regret}_T &= \sum_{t=1}^{T} \left[r(\boldsymbol{x}_t, \pi^\star(\boldsymbol{x}_t)) - \tfrac{1}{2}\{r(\boldsymbol{x}_t, \boldsymbol{a}_t) + r(\boldsymbol{x}_t, \boldsymbol{a}'_t)\} \right] \\
&\leq \frac{3}{2}\sqrt{\kappa_B} \sum_{t=1}^{T} \widehat{\beta}_t \|\boldsymbol{\psi}_{t,1}(\widehat{\boldsymbol{\theta}}_{t,1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\
&\leq \frac{3}{2}\widehat{\beta}_{T+1}\sqrt{\kappa_B} \sum_{t=1}^{T} \|\boldsymbol{\psi}_{t,1}(\widehat{\boldsymbol{\theta}}_{t,1})\|_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \\
&\leq \frac{3}{2}\widehat{\beta}_{T+1}\sqrt{\kappa_B T \sum_{t=1}^{T} \|\boldsymbol{\psi}_{t,1}(\widehat{\boldsymbol{\theta}}_{t,1})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}}}.
\end{aligned}$$

Since $q_t = 1$ in the deployment setting, Lemma 3 gives $\sum_{t=1}^{T} \|\boldsymbol{\psi}_{t,1}(\widehat{\boldsymbol{\theta}}_{t,1})\|^2_{\widehat{\boldsymbol{\mathcal{H}}}_t^{-1}} \leq 2\Gamma_{T+1}$. Therefore, we have

$$\mathrm{Regret}_T \leq \frac{3}{2}\widehat{\beta}_{T+1}\sqrt{2\kappa_B T \Gamma_{T+1}} \lesssim \widetilde{O}((l + d\Delta_l)\sqrt{T}),$$

which completes the proof. □

E Omitted Technical Lemmas

Lemma 5 (Ghashami et al. (2016, Theorem 1.1)). *Let $\boldsymbol{X}_n \in \mathbb{R}^{n\times d}$ have rows $\boldsymbol{v}_1^\top, \ldots, \boldsymbol{v}_n^\top$. Let $\boldsymbol{B}_i$ be the standard FD sketch after processing the first i rows, with sketch size $1 \leq l < d$. Denote by ρ_i the amount subtracted from the squared singular values at step i. Then, for each $i = 1, \ldots, n$,*

$$\boldsymbol{0} \preceq \boldsymbol{B}_{i-1}^\top \boldsymbol{B}_{i-1} + \boldsymbol{v}_i \boldsymbol{v}_i^\top - \boldsymbol{B}_i^\top \boldsymbol{B}_i \preceq \rho_i \boldsymbol{I}_d.$$

Moreover,

$$\sum_{i=1}^{n} \rho_i \leq \min_{0\leq r<l} \frac{\|\boldsymbol{X}_n - (\boldsymbol{X}_n)_{[r]}\|_F^2}{l - r},$$

where $(\boldsymbol{X}_n)_{[r]}$ is a best approximation to $\boldsymbol{X}_n$ of rank at most r.

Lemma 6 (Part (i): Orabona (2019, Lemma 6.39); Part (ii): Campolongo & Orabona (2020, Proposition 4.1)). *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be closed and convex, $\boldsymbol{H} \succ \boldsymbol{0}$, $\boldsymbol{x}_0 \in \mathcal{C}$, and $\eta > 0$.*

(i) For any $\boldsymbol{g} \in \mathbb{R}^d$, let

$$\boldsymbol{x}_+ \in \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{ \langle \boldsymbol{g}, \boldsymbol{x} \rangle + \frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2_{\boldsymbol{H}} \right\}.$$

Then

$$\|\boldsymbol{x}_+ - \boldsymbol{x}_0\|_{\boldsymbol{H}} \leq \eta \|\boldsymbol{g}\|_{\boldsymbol{H}^{-1}}.$$

(ii) For any differentiable convex function $F : \mathbb{R}^d \to \mathbb{R}$ and $\boldsymbol{u} \in \mathcal{C}$, let

$$\boldsymbol{x}_+ \in \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{ F(\boldsymbol{x}) + \frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2_{\boldsymbol{H}} \right\}.$$

Then

$$\langle \nabla F(\boldsymbol{x}_+), \boldsymbol{x}_+ - \boldsymbol{u} \rangle \leq \frac{1}{2\eta}\left(\|\boldsymbol{x}_0 - \boldsymbol{u}\|^2_{\boldsymbol{H}} - \|\boldsymbol{x}_+ - \boldsymbol{u}\|^2_{\boldsymbol{H}} - \|\boldsymbol{x}_+ - \boldsymbol{x}_0\|^2_{\boldsymbol{H}} \right).$$

Lemma 7 (Abbasi-Yadkori et al. (2011, Elliptical potential lemma)). *Let $\{\boldsymbol{z}_t\}_{t\geq 1} \subseteq \mathbb{R}^d$ satisfy $\|\boldsymbol{z}_t\|_2 \leq L$ for all $t \geq 1$. For $\lambda > 0$ and $t \geq 1$, define $\boldsymbol{\Lambda}_t := \lambda \boldsymbol{I}_d + \sum_{s=1}^{t-1} \boldsymbol{z}_s \boldsymbol{z}_s^\top$. Then, for every $T \geq 1$,*

$$\sum_{t=1}^{T} \min\{1, \|\boldsymbol{z}_t\|^2_{\boldsymbol{\Lambda}_t^{-1}}\} \leq 2d \log\left(1 + \frac{L^2 T}{d\lambda}\right).$$

Lemma 8 (Howard et al. (2020, Lemma 1)). *Let $(M_n)_{n\geq 0}$ be a nonnegative supermartingale with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$. Then, for every $a > 0$,*

$$\mathbb{P}\left(\sup_{n\geq 0} M_n \geq a\right) \leq \frac{\mathbb{E}[M_0]}{a}.$$

Lemma 9 (Nabeya (1951)). *Let $\boldsymbol{H} \succ \boldsymbol{0}$ and $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}^{-1})$. Then, for every $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$,*

$$\mathbb{E}\left[|\boldsymbol{a}^\top \boldsymbol{g}|\, |\boldsymbol{b}^\top \boldsymbol{g}|\right] \geq \frac{2}{\pi} \|\boldsymbol{a}\|_{\boldsymbol{H}^{-1}} \|\boldsymbol{b}\|_{\boldsymbol{H}^{-1}}.$$

Lemma 10 (Abeille et al. (2021, Lemma 9)). *Let $B \geq 0$, $\mu(u) = (1 + e^{-u})^{-1}$, $\dot{\mu}(u) = \mu(u)(1 - \mu(u))$, and $h(u) = \log(1 + e^{-u})$. Then, for all $u, v \in \mathbb{R}$,*

$$0 < \dot{\mu}(u) \leq \tfrac{1}{4}, \qquad |\log \dot{\mu}(u) - \log \dot{\mu}(v)| \leq |u - v|, \qquad |h'''(u)| \leq \tfrac{1}{4}.$$

Moreover, for all $u \in [-2B, 2B]$,

$$\dot{\mu}(u)^{-1} \leq \kappa_B := 2 + 2\cosh(2B).$$

Lemma 11 (Zhang & Sugiyama (2023, Lemma 1)). *Let $B \geq 0$, $\boldsymbol{a}(u) := (u/2, -u/2)$, and*

$$L_1(\boldsymbol{z}) := \log(e^{z_1} + e^{z_2}) - z_1, \qquad c_B := \log 2 + 2(B + 1).$$

Then, for all $s, v \in [-2B, 2B]$,

$$L_1(\boldsymbol{a}(v)) \geq L_1(\boldsymbol{a}(s)) + \langle \nabla L_1(\boldsymbol{a}(s)), \boldsymbol{a}(v) - \boldsymbol{a}(s) \rangle + \frac{1}{c_B} \|\boldsymbol{a}(v) - \boldsymbol{a}(s)\|^2_{\nabla^2 L_1(\boldsymbol{a}(s))}.$$

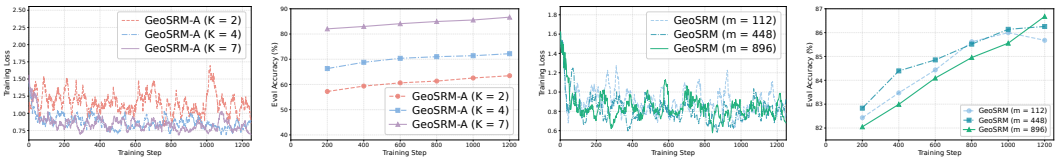
F Omitted Details for Section 5

In this section, we provide the omitted details of experiments in section 5. We provide the detailed experimental setups and configurations in Appendix F.1 and additional experiments in F.2.

F.1 Implementation Details

Experimental Setups in Section 5.1. Following the experimental setups in Lee et al. (2025), we use the first 5,000 examples from Nectar's training split and retain their provided response rankings. Each GPU process receives a minibatch of four candidate prompts. The proposed GeoSRM-A and M-AUPO (Lee et al., 2025) construct a candidate assortment for each prompt using GreedyMax and select the prompt with the largest uncertainty score within that minibatch, as in Section 4.1. Uniform instead samples a prompt uniformly from the minibatch while retaining assorment selection and OMD updates in Eq. (9). To ensure a fair comparison, we adopt the hyperparameter settings from Lee et al. (2025). Unless specified, we set max assortment size $K = 7$, $\eta = 10^{-3}$ and $\lambda = 0.1$ for all methods. For GeoSRM-A, we set sketch size $l = d/4$ for different size of LLMs backbones.

Evaluation uses the same fixed prompt pool as active query selection. For each choice of K, we retain the top-K ranked responses per prompt in both the query and evaluation pools. We report pairwise preference accuracy, defined as the fraction of evaluated prompts for which the highest-ranked retained response receives a higher predicted reward than the lowest-ranked retained response; for $K = 7$, these correspond to ranks 1 and 7. Incomplete evaluation batches are discarded, yielding 4,992 evaluated prompts per checkpoint in the four-GPU runs. Algorithm time includes context and assortment selection, reward-head updates, and curvature maintenance.



(a) Training loss (b) Evaluation accuracy (c) Training loss (d) Evaluation accuracy

Figure 6: Online RLHF with active preference optimization in `Qwen2.5-7B-Instruct`. We report the training loss and evaluation accuracy of GeoSRM-A under (a), (b): varying maximum assortment sizes K; and (c), (d): varying sketch sizes m.

Experimental Setups in Section 5.2. We use the binarized UltraFeedback dataset (Rafailov et al., 2023), a widely used benchmark for RLHF. We partition its `test_gen` split into 20 consecutive chunks of 50 prompts. The Reward model backbones are initialized from supervised fine-tuned `Qwen2.5-Instruct` checkpoints and remain frozen during reward learning. The first reward-head fit uses four preference pairs from `train_prefs`; subsequent fits make one pass over the pairs collected in the preceding iteration. PPO operates on the current prompt chunk.

Following Li et al. (2025), we adopt a rejection sampling-based approximation to avoid quantifying uncertainty over the large response set (Gülçehre et al., 2023; Dong et al., 2024). Specifically, given a prompt, the updated policy then samples 64 responses per prompt, and the learned reward model selects the highest-scoring response and the q-th highest-scoring response to ensure diversity, where q is a hyperparameter and set $q = 4$ by default. The fixed `Llama-3-OffsetBias-RM-8B` oracle (Park et al., 2024) labels the selected pair, providing feedback for the next iteration. All reward-head methods set learning rate $\eta = 10^{-3}$. GeoSRM-D set sketch size $l = d/4$ for the 1.5B, 3B, and 7B LLMs backbones, respectively. OMD-HVP choose the linear function $f(t/T) = t/T$ as the damping function, set $\lambda_0 = 0.8$, and performs at most 3 CG steps in the updates. PPO uses one training epoch, rank-16 LoRA adapters (Hu et al., 2022).

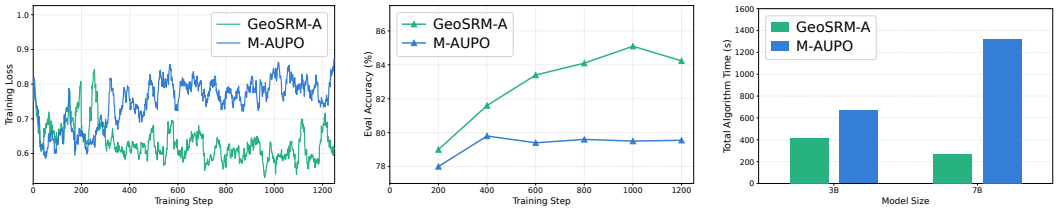
Deployment performance is evaluated using the oracle scores on the online prompt stream. We define the pair score as the mean oracle reward of the two selected responses. Average cumulative reward is the running average of the per-iteration mean pair scores. Prompt-level win rate measures the fraction of shared prompt–iteration instances on which a method obtains the highest pair score among the compared methods, with ties sharing the credit equally.

F.2 More Experimental Results

Learning Dynamics with Varying K. Complementing Figure 4c, we report the training loss and evaluation accuracy curves of GeoSRM-A for maximum assortment sizes $K \in \{2, 4, 7\}$ in Figures 6a and 6b. For $K = 4$ and $K = 7$, the training loss decreases rapidly in the early stage and remains generally lower than that for $K = 2$, which exhibits larger fluctuations throughout training. A similar trend emerges in evaluation accuracy: all three settings improve steadily, with $K = 7$ consistently achieving the highest accuracy, followed by $K = 4$ and $K = 2$. These observations are consistent with Theorem 1, supporting the benefit of larger assortment sizes for learning from multi-comparison feedback.

Learning Dynamics with Varying m. We further examine how the sketch size affects the learning dynamics of GeoSRM-A with sketch size $m \in \{112, 448, 896\}$. As shown in Figures 6c and 6d, while smaller sketches reduce computational and memory costs, stronger compression can result in less stable learning dynamics. In particular, $m = 112$ exhibits more pronounced fluctuations in training loss, and its evaluation accuracy drops at the final checkpoint. In contrast, $m = 896$ shows more stable training dynamics, with evaluation accuracy improving steadily and reaching the highest final value. Together, these results illustrate the efficiency–stability trade-off induced by sketch size: smaller sketches reduce computational overhead, but excessive compression can make training more volatile and the resulting accuracy gains less consistent.

Active Learning Efficacy across Model Sizes. Complementing the 7B results in Figures 4a and 4b, we examine the learning dynamics with `Qwen2.5-3B-Instruct`. As shown in Figures 7a and 7b, M-AUPO fails to sustain its initial progress: the training loss subsequently rebounds,



(a) Training loss (b) Evaluation accuracy (c) Running Time

Figure 7: Comparison of GeoSRM-A and M-AUPO in online RLHF with active preference optimization. We report (a) training loss and (b) evaluation accuracy for `Qwen2.5-3B-Instruct`, and (c) total algorithm time for `Qwen2.5-3B-Instruct` and `Qwen2.5-7B-Instruct`.

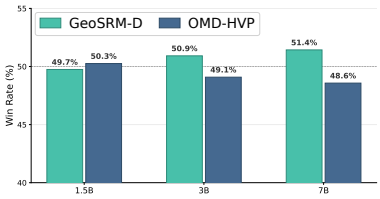
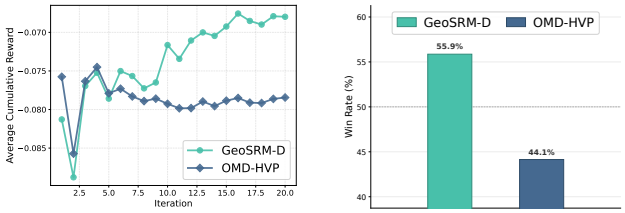


Figure 8: Win rates of different methods in deployment-time adaptation with Qwen2.5-Instruct.



(a) Cumulative reward (b) Win rate

Figure 9: Online RLHF with deployment-time adaptation in `Llama-3.2-3B-Instruct`, we report the comparison about (a) average cumulative reward and (b) win rate.

while the evaluation accuracy stagnates at around 79.5%. GeoSRM-A maintains a substantially lower training loss and achieves a final accuracy of approximately 84.2%. These results indicate that GeoSRM-A exhibits more stable learning dynamics than M-AUPO with the smaller backbone. Our method employs matrix sketching to reshape the curvature spectrum used for parameter updates, yielding an update scale that can be more conducive to stable learning when the capabilities of the base model are limited.

Figure 7c further shows that the computational advantage of GeoSRM-A becomes more pronounced as the model size increases. This scaling behavior reflects two effects. First, the sketched curvature operations in GeoSRM-A have a weaker dependence on the hidden-state dimension than the full-matrix operations in M-AUPO. Second, the 7B runs use smaller assortments, resulting in fewer rank-broken comparisons and hence fewer parameter and curvature updates. Under the greedy assortment construction in Eq. (8), this corresponds to fewer assortment extensions before the stopping criterion is met. Thus, the observed runtime difference reflects both more efficient curvature operations and variation in the number of updates induced by the adaptive assortment construction.

Deployment-time Learning Efficacy across Model Sizes. In addition to Figures 5a, 5b, and 5c, we report the win rates of GeoSRM-D and OMD-HVP across different `Qwen2.5-Instruct` model sizes. As shown in Figure 8, the two methods perform comparably on the smallest backbone, while the advantage of GeoSRM-D becomes more pronounced as the model size increases. This trend suggests that FD-based geometry sketching remains effective as the backbone scales up.

Deployment-time Learning on Other Foundation Models. To examine whether the gains observed on Qwen extend to another model family, we further evaluate GeoSRM-D and OMD-HVP with `Llama-3.2-3B-Instruct`. As shown in Figure 9a, GeoSRM-D exhibits an overall upward trend in average cumulative reward after the initial fluctuations, whereas OMD-HVP largely plateaus. The gap becomes more pronounced in later deployment rounds, suggesting that GeoSRM-D continues to improve under prolonged interaction while OMD-HVP makes limited further progress. GeoSRM-D also achieves a higher win rate in Figure 9b, selecting response pairs with higher mean oracle rewards more frequently than OMD-HVP.