Feel-Good Thompson Sampling for Contextual Dueling Bandits

Xuheng Li [1] **Heyang Zhao** [1] **Quanquan Gu** [1]

Abstract

Contextual dueling bandits, where a learner compares two options based on context and receives feedback indicating which was preferred, extends classic dueling bandits by incorporating contextual information for decision-making and preference learning. Several algorithms based on the upper confidence bound (UCB) have been proposed for linear contextual dueling bandits. However, no algorithm based on posterior sampling has been developed in this setting, despite the empirical success observed in traditional contextual bandits. In this paper, we propose a Thompson sampling algorithm, named FGTS.CDB, for linear contextual dueling bandits. At the core of our algorithm is a new Feel-Good exploration term specifically tailored for dueling bandits. This term leverages the independence of the two selected arms, thereby avoiding a cross term in the analysis. We show that our algorithm achieves nearly minimax-optimal regret, i.e., $\widetilde{\mathcal{O}}(d\sqrt{T})$, where d is the model dimension and T is the time horizon. Finally, we evaluate our algorithm on synthetic data and observe that FGTS.CDB outperforms existing algorithms by a large margin.

1 Introduction

Reinforcement learning from human feedback (RLHF) has become a popular methodology in the alignment of large language models (LLMs, Ouyang et al. 2022; Diao et al. 2023). In RLHF, it is often easier for the human user to compare two responses than providing a numerical reward/score based on a common standard. Therefore, existing works on RLHF (Zhu et al., 2023; Ji et al., 2023) focus on a model where the learning agent has a dataset of users' preferences among several choices. The preferences are often assumed to follow the Plackett-Luce (PL) model (Soufiani et al., 2014; Khetan & Oh, 2016; Ren et al., 2018), where the probability of the user favoring a certain choice is proportional to the exponential of the reward function, and the special case where two choices are presented to the user is called the Bradley-Terry-Luce (BTL) model (Hunter, 2004; Luce, 2005). The online version of the preference-based model, called the *dueling bandits*, has been studied extensively (Yue et al., 2012; Zoghi et al., 2014; Komiyama et al., 2015) when the set of the action space is fixed and finite (i.e., the multi-armed dueling bandits). Recently, a more general model, the *(linear) contextual dueling bandit* (Saha, 2021; Bengs et al., 2022), has been proposed. This model has important features including time-varying and possibly infinite action spaces, along with a context-dependent reward function with a linear structure, which capture important practical situations. A number of algorithms have been proposed for (linear) contextual dueling bandits, including MaxInP (Saha, 2021), CoLSTIM (Bengs et al., 2022) and VACDB (Di et al., 2023), all of which are based on the upper confidence bound (UCB) technique for exploration.

Under the setting of traditional contextual bandits, Thompson sampling (Thompson, 1933) is another technique for exploration apart from UCB-based methods, and superior empirical performance has been observed when applying Thompson sampling to various tasks (Chapelle & Li, 2011; Osband & Van Roy, 2017). Instead of deterministically learning a model, in Thomson sampling, models are sampled from a posterior distribution constructed on historic observations. It has been widely studied in both the multi-armed setting (Agrawal & Goyal, 2012; Kaufmann et al., 2012; Agrawal & Goyal, 2017; Jin et al., 2021) and the linear setting (Agrawal & Goyal, 2013). Later, Zhang (2022) showed that the frequentist regret of linear Thompson sampling is suboptimal in the worst case and proposed a new variant of Thompson sampling called *Feel-Good Thomson sampling* (FGTS) to overcome this issue. The effectiveness of FGTS is theoretically justified: when applied to linear contextual bandits, FGTS can achieve the minimax-optimal regret bound as UCB-based algorithms like LinUCB (Li et al., 2010) or OFUL (Abbasi-Yadkori et al., 2011).

Despite the success of Thompson sampling algorithms in traditional contextual bandits, there have been few works that apply this technique to contextual dueling bandits. The notable exception is a double Thompson sampling approach proposed by Wu & Liu (2016). However, this approach is limited to multi-armed dueling bandits and cannot be modi-

[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

fied for (linear) contextual dueling bandits. In addition, it is also unknown whether algorithms based on Thompson sampling can achieve the same minimax-optimal regret bounds as UCB-based algorithms for contextual dueling bandits. Therefore, we raise the following question:

Is it possible to design an efficient algorithm for contextual dueling bandits based on Feel-Good Thompson sampling?

In this paper, we affirmatively answer this question by solving the problem of linear contextual dueling bandits under the framework of Feel-Good Thompson sampling. We summarize our contributions as follows:

- We propose a new algorithm named FGTS.CDB for the problem of linear contextual dueling bandits, which is based on Feel-Good Thompson sampling. Compared with existing FGTS algorithms for standard contextual dueling bandits (Zhang, 2022), we introduce a new Feel-Good exploration term designed specially for the comparison of two actions. Compared with UCB-based approaches, our algorithm can handle the case of large action spaces more efficiently.
- We prove that our algorithm enjoys a minimax-optimal regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T})$ in expectation, where d is the feature dimensionality and T is the number of rounds. The new Feel-Good exploration term plays a crucial role in the proof by eliminating cross terms that arise from the comparison of actions.
- We extend our analysis to the setting of general reward functions, and manage to recover the regret bound for several cases of interest, including the cases of finite action sets and finite model sets.
- We conduct experiments to compare our algorithms with several strong baselines, including MaxInP, MaxPairUCB (Saha, 2021), CoLSTIM (Bengs et al., 2022) and VACDB (Di et al., 2023). We observe that the performance of FGTS.CDB is significantly better than all baselines.

Notation. We use plain case letters to denote scalars and lowercase boldface letters to denote vectors. We use $\langle \cdot, \cdot \rangle$ to denote the inner product of vectors. For a vector $\mathbf{x}$, $\|\mathbf{x}\|_2$ denotes its ℓ_2-norm. We use $[N]$ as a shorthand for the set $\{1, 2, \ldots, N\}$. We use standard asymptotic notations including $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$, while $\widetilde{\mathcal{O}}(\cdot)$ $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ hide logarithmic factors.

2 Related Work

Dueling bandits. First proposed by Yue et al. (2012), the dueling bandit problem involves a learner sequentially selecting a pair of arms among multiple choices based on the noisy binary observations revealing the relative preference of the chosen arms. Under their multi-armed dueling bandit setting, Zoghi et al. (2014) proposed RUCB, a UCB-based algorithm which achieves an $\mathcal{O}(K \log T/\Delta)$ regret upper bound where K is the number of arms, T is the number of rounds, and Δ stands for the gap between the best arm and the second-best arm. Later, Komiyama et al. (2015) proposed RMED with a more sophisticated arm selection phase whose regret matches the lower bound with optimal constant. Relaxing the typical Condorcet winner setting where it is assumed that there is one arm that beats all the other arms, researchers also investigated other variants of multi-armed dueling bandits which assumed the existence of Copeland Winner (Zoghi et al., 2015; Wu & Liu, 2016; Komiyama et al., 2016), Borda winner (Jamieson et al., 2015; Falahatgar et al., 2017; Heckel et al., 2016; Saha et al., 2021; Brandt et al., 2022; Wu et al., 2023), or von Neumann winner (Dudík et al., 2015; Balsubramani et al., 2016; Ramamohan et al., 2016).

Contextual dueling bandits. There is also a large body of literature on contextual dueling bandits, where dueling bandits with contextual information is considered (Kumagai, 2017; Saha, 2021; Saha & Krishnamurthy, 2022; Bengs et al., 2022; Di et al., 2023). Kumagai (2017) studied dueling bandits with a cost function over a continuous space and achieved a dimension-free regret under the strong convexity and smoothness assumption. Saha (2021) considered contextual dueling bandits with generalized linear classes and proposed an algorithm MaxInP with an $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret and Sta'D with an $\widetilde{\mathcal{O}}(\sqrt{dT \log K})$ regret. Bengs et al. (2022) proposed CoLSTIM and further extended it to the contextual linear stochastic transitivity model. Recently, Di et al. (2023) proposed an action-elimination based algorithm VACDB, with a tighter variance-dependent regret bound. It is worth mentioning that all the existing algorithms for contextual dueling bandits need to either maintain a subset of eligible arms or maximize the randomly perturbed rewards over all the possible arms, which are only applicable to finite action space.

Feel-Good Thompson sampling (FGTS). FGTS was proposed by Zhang (2022) to fill the gap between the practical effectiveness of Thompson sampling and a lack of frequentist-type regret guarantee. When applied to linear contextual bandits, FGTS achieves a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T})$ that matches the lower bound of $\Omega(d\sqrt{T})$. The analysis of this algorithm is based on the decoupling of arm selection with model parameters. Fan & Gu (2023) proposed a unified framework for the analysis of FGTS applied to a number of variants of linear contextual bandits. Another line of works extends the idea of FGTS to reinforcement learning, including Model-based Optimistic Posterior Sampling (MOPS) for Markov decision processes (Agarwal & Zhang, 2022) and conditional Posterior Sampling with Booster for two-player Markov games (Xiong et al., 2022). Our work is the first attempt to apply FGTS to contextual dueling bandits.

Table 1. Comparison of our algorithm, FGTS.CDB, against existing algorithms for linear contextual dueling bandits. FGTS.CDB is the first algorithm for linear contextual dueling bandits using the technique of Thompson sampling. Our algorithm is also the first that can be easily applied to the case of infinite action spaces (modification for MaxInP is more complex). The regret bounds hold for linear contextual dueling bandits of T rounds, with d-dimensional feature vectors and the action space of size K.

Algorithm	Main technique	Infinite action space?	Regret
MaxInP (Saha, 2021)	UCB + Adaptive Selection	✓	$\widetilde{\mathcal{O}}(d\sqrt{T})$
CoLSTIM (Bengs et al., 2022)	Perturbed UCB	✗	$\widetilde{\mathcal{O}}(d\sqrt{T})$
Sta'D (Saha, 2021)	SupLinUCB + Adaptive Selection	✗	$\widetilde{\mathcal{O}}(\sqrt{dT\log K})$
SupCoLSTIM (Bengs et al., 2022)	Perturbed SupLinUCB	✗	$\widetilde{\mathcal{O}}(\sqrt{dT\log K})$
FGTS.CDB (This work)	Feel-Good Thompson Sampling	✓	$\widetilde{\mathcal{O}}(d\sqrt{T})$

Sampling-based algorithms for dueling bandits. Wu & Liu (2016) proposed a double Thompson sampling algorithm for multi-arm dueling bandits which achieves a regret bound of $\mathcal{O}(K^2 \log T)$ where K is the number of arms. Sui et al. (2017) also proposed an algorithm based on Thompson sampling that converted multi-dueling bandits to standard bandits. However, these two algorithms cannot be modified for the setting of (linear) contextual dueling bandits because they depend on the count of comparison outcomes between the arms and T is the number of rounds. In contextual dueling bandits, this is infeasible because the set of arms are different across different rounds. Other algorithms are based on the sampling of policies rather than model parameters. For example, Xiong et al. (2023) proposed a KL-constrained framework which uses Gibbs sampling. Nonetheless, this work focuses on fine-tuning LLMs, and the regret studied in this work has an additional term that measures the difference between the learned policy and the original policy. Novoseller et al. (2020) studied the application of posterior sampling in preference-based reinforcement learning. However, the regret bound of the algorithm relies on the assumption of finite state and action sets and cannot be trivially extended to linear contextual dueling bandits.

3 Problem Setting

In this work, we study the setting where the agent repeatedly interact with the agent to receive prompts and query preferences between the two chosen responses.

Linear contextual dueling bandits. We focus on the setting of contextual dueling bandits with contextual information embodied in both the prompt and the action space, similar to Zhang (2022). Let $\mathcal{X}$ be the set of prompts and $\mathcal{A}$ be the set of all possible responses. During round t in a total of T rounds, the agent receives a prompt $x_t \in \mathcal{X}$, along with a corresponding action space $\mathcal{A}_t \subset \mathcal{A}$ which can both vary across different rounds. The agent then selects two responses (more commonly referred to as arms in the bandit context) $a_t^1, a_t^2 \in \mathcal{A}_t$ and receives a randomized preference y_t whose distribution depends on an underlying reward function $r_* : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. $y_t = 1$ represents the case where a_t^1 is preferred over a_t^2, and $y_t = -1$ otherwise. We assume that the reward function class adopts a linear structure:

Assumption 3.1 (Linear reward)**.** We assume that the reward function is parameterized by $r_{\boldsymbol{\theta}} = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(s,a) \rangle$ for some known feature mapping $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$. Specifically, the real value function is $r_*(x,a) = \langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x,a) \rangle$ for some vector $\boldsymbol{\theta}_* \in \mathbb{R}^d$ hidden from the learning agent. We assume that $\|\boldsymbol{\phi}(s,a)\|_2 \le 1$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, and $\|\boldsymbol{\theta}\|_2 \le B$. Thus, the reward function is bounded by $|r_{\boldsymbol{\theta}}(\cdot,\cdot)| \le B$.

The setting we study is equivalent to those of previous works on contextual dueling bandits. Saha (2021) and Bengs et al. (2022) considered a time-varying action space $\mathcal{S}_t = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ where each action is represented by a d-dimensional vector called the *contextual vector*, and the reward function is defined as $r_*(\mathbf{a}) = \langle \boldsymbol{\theta}_*, \mathbf{a} \rangle$. The contextual vector depends on both the prompt x_t and the response a_t, and can be viewed as a counterpart of the feature mapping $\boldsymbol{\phi}(\cdot,\cdot)$ in Assumption 3.1. Compared with the contextual vector, our formulation is more general when considering other types of function approximations (see Section 6).

Stochastic preference model. In this work, we assume that the preference y_t follows a Bernoulli distribution according to the Bradley-Terry-Luce (BTL) model (Hunter, 2004; Luce, 2005): Given context x_t and responses a_t^1, a_t^2, the probability of a_t^1 is preferred over a_t^2 is

$$\begin{aligned}\mathbb{P}(y_t = 1 | x_t, a_t^1, a_t^2) &= \frac{\exp(r_*(x_t, a_t^1))}{\exp(r_*(x_t, a_t^1)) + \exp(r_*(x_t, a_t^2))} \\ &= \exp(-\sigma(r_*(x_t, a_t^1) - r_*(x_t, a_t^2))),\end{aligned}$$

where $\sigma(z) = \log(1 + \exp(-z))$.

Some other works study a more general setting called the Plackett-Luce (PL) model (Soufiani et al., 2014; Khetan & Oh, 2016; Ren et al., 2018), where the learning agent selects $q \ge 2$ arms $a_t^1, \ldots, a_t^q \in \mathcal{A}_t$ in round t and receives the

preference $o_t \in [q]$. The probability of a_t^j being preferred is

$$\mathbb{P}(o_t = j) = \frac{\exp(r_*(x_t, a_t^j))}{\sum_{j'=1}^{q} \exp(r_*(x_t, a_t^{j'}))}.$$

The BTL model can be seen as a special case of the PL model by fixing $q = 2$. Saha (2021) showed that under the PL model, the worst-case regret of any algorithm for dueling bandits is $\Omega(d\sqrt{T})$, regardless of the choice of q. Therefore, provided that the learner is permitted to select any number of arms, it suffices to design a minimax-optimal algorithm where two arms are selected in each round, which is shown to be true for our algorithm in Section 5.

Learning Objective. Our goal is to minimize the cumulative average regret defined as

$$\mathrm{Regret}(T) := \sum_{t=1}^{T} \left[r_*(x_t, a_t^*) - \frac{r_*(x_t, a_t^1) + r_*(x_t, a_t^2)}{2} \right],$$

where $a_t^* = \mathrm{argmax}_{a \in \mathcal{A}} r_*(x_t, a)$ is the optimal response at time t. The regret we study is exactly the same as the regret studied in Saha (2021) and Bengs et al. (2022). The regret is also equivalent to the dueling bandit regret studied in Yue et al. (2012), defined as

$$\sum_{t=1}^{T} \frac{1}{2} \Big[\Big(\exp(-\sigma(r_*(x_t, a_t^*) - r_*(x_t, a_t^1))) - \frac{1}{2} \Big) + \Big(\exp(-\sigma(r_*(x_t, a_t^*) - r_*(x_t, a_t^2))) - \frac{1}{2} \Big) \Big],$$

because $\exp(-\sigma(z)) - 1/2 = \Theta(z)$ for $z \in [-2B, 2B]$.

4 Algorithm Description

We now present our algorithm, named FGTS.CDB, for linear contextual dueling bandits. The pseudocode is shown in Algorithm 1.

Algorithm 1 FGTS.CDB

Given hyperparameters η, μ. Initialize $S_0 = \varnothing$.
for $t = 1, \ldots, T$ **do**
  Receive prompt x_t and action space $\mathcal{A}_t$.
  for $j = 1, 2$ **do**
    Sample model parameter $\boldsymbol{\theta}_t^j$ from the posterior distribution $p^j(\cdot|S_{t-1})$, defined in (4.1).
    Select response $a_t^j = \mathrm{argmax}_{a \in \mathcal{A}_t} \langle \boldsymbol{\theta}_t^j, \boldsymbol{\phi}(x_t, a) \rangle$.
  end for
  Receive preference y_t.
  Update dataset $S_t \leftarrow S_{t-1} \cup \{(x_t, a_t^1, a_t^2, y_t)\}$.
end for

In our algorithm, the agent first samples model parameters $\boldsymbol{\theta}_t^1$ and $\boldsymbol{\theta}_t^2$ independently, following posterior distributions $p^1(\cdot|S_{t-1})$ and $p^2(\cdot|S_{t-1})$, respectively. The posterior distributions are defined as

$$p^j(\boldsymbol{\theta}|S_{t-1}) \propto \exp\Big(-\sum_{i=1}^{t-1} L^j(\boldsymbol{\theta}, x_i, a_i^1, a_i^2, y_i) \Big) p_0(\boldsymbol{\theta}), \tag{4.1}$$

where L^j is the likelihood function, and $p_0(\cdot)$ is the prior distribution. Sampling from such a posterior distribution can be implemented via Langevin Monte Carlo (LMC), which has been studied extensively in the literature (Roberts & Tweedie, 1996; Bakry et al., 2014). Afterwards, actions a_t^j are selected to maximize the inner product of the parameter $\boldsymbol{\theta}_t^j$ and the feature mapping $\boldsymbol{\phi}(x_t, a_t^j)$ for $j = 1, 2$. Finally, the agent receives the binary preference $y_t \in \{\pm 1\}$ and augments the dataset with (x_t, a_t^1, a_t^2, y_t).

Feel-Good Thompson sampling. In our algorithm, the likelihood function is defined as

$$L^j(\boldsymbol{\theta}, x, a^1, a^2, y) = \eta \sigma(y \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a^1) - \boldsymbol{\phi}(x, a^2) \rangle) - \mu \max_{a' \in \mathcal{A}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a') - \boldsymbol{\phi}(x, a^{3-j}) \rangle,$$

where η and μ are hyperparameters. In the definition above, the first term can be treated as the log-likelihood function on the observation (x, a^1, a^2, y); the second term encourages exploration of $\boldsymbol{\theta}$ with large reward in previous rounds, which is referred to as *Feel-Good exploration* in the literature (Zhang, 2022). Without the Feel-Good exploration term, i.e., when $\mu = 0$, L^j reduces to the likelihood function used in standard Thompson sampling algorithms.

Comparison with FGTS for traditional contextual bandits. The differences between FGTS.CDB and existing FGTS algorithms for traditional contextual bandits (Zhang, 2022) are twofold. Firstly, due to the preferential feedback, the least-squares term in previous algorithms is naturally replaced with a term in the form of logistic regression. The more important difference lies in the Feel-Good exploration terms. In our Feel-Good exploration term, there is an additional inner product of the current model parameter $\boldsymbol{\theta}$ and the feature vector of the adversarial arm $\boldsymbol{\phi}(x, a^{3-j})$. This additional term is a better design for the setting of contextual dueling bandits because the affecting factor of the observation y_t is the difference between the rewards of two arms rather than the reward of a single arm. Additionally, this term plays a crucial role in the proof, as we will show in Section 7.

Comparison with UCB-based algorithms. We highlight that besides the model learning technique, there is also a stark difference in the arm selection scheme between UCB-based algorithms (including MaxInP (Saha, 2021), CoLSTIM (Bengs et al., 2022), VACDB (Di et al., 2023)) and FGTS.CDB. In UCB-based algorithms, arms are often selected based on a bonus term in the form of $\|\boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\|_{\boldsymbol{\Sigma}^{-1}}$, for some positive definite matrix $\boldsymbol{\Sigma}$, which encourages the selection of more separated arms. Thus, the bonus term is essential for exploration in UCB-based algorithms, and results in the dependence of the selected arms. In FGTS.UCB, however, the arms a_t^1 and a_t^2 are just maximizers of the learned reward function, and are independent conditioned on the history S_{t-1}. This is possible because exploration is accomplished by Thompson

sampling in our algorithm, and the bonus term becomes unnecessary. In addition, the independence of arms a_t^1 and a_t^2 (conditioned on S_{t-1}) is a crucial property in our proof as we will show in Section 7.

From the viewpoint of computational complexity, the arm selection scheme of FGTS.CDB is also superior to those of existing algorithms. When the action space $\mathcal{A}_t$ is infinite and continuous, the arm selection phase of FGTS.CDB can still be implemented by solving an optimization problem. In contrast, MaxInP calculates a set of promising arms in each round, which causes additional computational overhead in the case of infinite action spaces. CoLSTIM needs to take the maximum of randomly perturbed rewards corresponding to each contextual vector, which is infeasible when the number of arms is infinite. Therefore, FGTS.CDB is the first algorithm that can be easily applied to the setting of infinite action spaces.

5 Main Results

In this section, we present the regret bounds of Algorithm 1, which is minimax-optimal. We first introduce the following assumption about the prior distribution p_0:

Assumption 5.1. The logarithm of the prior distribution is L-Lipschitz, i.e., for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \le B\}$, we have

$$|\log p_0(\boldsymbol{\theta}_1) - \log p_0(\boldsymbol{\theta}_2)| \le L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.$$

Assumption 5.1 is satisfied for most commonly-used prior distributions, including the uniform distribution ($L = 0$) and the Gaussian distribution $\mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$ restricted to the ball $\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \le B\}$ ($L = B/\sigma_0^2$).

We now present the main theorem:

Theorem 5.2. *Under Assumptions 3.1 and 5.1, assume that the hyperparameters are selected as $\eta = 0.25$ and $\mu = 1/(10e^B\sqrt{T})$, then the expectation of the regret of Algorithm 1 satisfies*

$$\mathbb{E}[\mathrm{Regret}(T)] = \widetilde{\mathcal{O}}(d\sqrt{T}).$$

The following theorem provides a regret lower bound for contextual dueling bandits:

Theorem 5.3. *Under Assumption 3.1, assume that for all $t \in [T]$, we have $\{\boldsymbol{\phi}(x_t, a) : a \in \mathcal{A}_t\} = \{\mathbf{u} : \|\mathbf{u}\|_2 \le 1\}$. We also assume that $T \ge \max\{72B^{-2}d^2, d\}$. Then for any algorithm for linear contextual dueling bandits, there exists $\boldsymbol{\theta}_*$ such that the expectation of the regret satisfies*

$$\mathbb{E}[\mathrm{Regret}(T)] = \Omega(d\sqrt{T}).$$

Remark 5.4. Theorem 5.3 shows that the regret lower bound of any algorithm for linear contextual dueling bandits is $\Omega(d\sqrt{T})$. Note that Theorem 3.1 in Bengs et al. (2022) also provides a regret lower bound of algorithms for linear contextual dueling bandits, but their lower bound is looser than ours by a factor of $\sqrt{d}$. They applied the analysis for the case of bounded ℓ_∞-norm to the case of bounded ℓ_2-norm, which yields loose inequalities (Lattimore & Szepesvári, 2020). Combining Theorem 5.2 and Theorem 5.3, we conclude that the regret bound of FGTS.CDB matches the worst-case regret. Our regret bound also matches that of UCB-based algorithms including MaxInP (Saha, 2021) and CoLSTIM (Bengs et al., 2022). More recently, Di et al. (2023) proposed an algorithm that has a variance-dependent regret bound, and it is an interesting future direction to design a variance-aware sampling-based algorithms for contextual dueling bandits.

Remark 5.5. Some other works assume that the size of the action space K is finite and derive algorithms with the regret bound $\widetilde{\mathcal{O}}(\sqrt{dT\log K})$, including Sta'D (Saha, 2021) and SupCoLSTIM (Bengs et al., 2022). We first note that the regret bound of $\widetilde{\mathcal{O}}(\sqrt{dT\log K})$ is not a contradiction against Theorem 5.2 due to the assumption of a total of K arms. More specifically, the proof of Theorem 5.3 involves contructing $\{\boldsymbol{\phi}(x_t, a) : a \in \mathcal{A}_t\} = \{\mathbf{u} : \|\mathbf{u}\|_2 \le 1\}$, so $K = 2^d$, and $\widetilde{\mathcal{O}}(\sqrt{dT\log K}) = \widetilde{\mathcal{O}}(d\sqrt{T})$. When K is exponential in the model dimensionality d, which is more often the case in the setting of contextual dueling bandits, the regret bound of FGTS.CDB is at least as good as these algorithms. In addition, our algorithm is more computationally efficient because it does not need to perform arm elimination or to apply random pertubations to each arm.

6 Extension to Nonlinear Reward

In this section, we relax the assumption of linear reward functions. Instead, we make the following assumption about the reward function class:

Assumption 6.1. The parameter space Θ is a measurable space with measure $\bar{\mu}$ and metric d. The model is well-specified, i.e., $\theta_* \in \Theta$. The reward function is uniformly bounded by B and is L_0-Lipschitz in θ.

We define a shorthand notation

$$\Delta r_\theta(x, a^1, a^2) := r_\theta(x, a^1) - r_\theta(x, a^2).$$

In order to characterize the complexity of the reward function class, similar to Zhang (2022), we define the decoupling coefficient dc to be such that for any $\lambda > 0$ and any joint distribution P over $\Theta \times \mathcal{A} \times \mathcal{A}$, we have

$$\begin{aligned}&\mathbb{E}_{(\theta,a^1,a^2)\sim P}[\Delta r_\theta(x, a^1, a^2) - \Delta r_{\theta_*}(x, a^1, a^2)] \\ &\le \lambda\mathrm{dc} + \frac{1}{4\lambda}\mathbb{E}_{\tilde{\theta}\sim P|_\theta}\mathbb{E}_{(a^1,a^2)\sim P|_{(a^1,a^2)}} \\ &\quad [\Delta r_{\tilde{\theta}}(x, a^1, a^2) - \Delta r_{\theta_*}(x, a^1, a^2)]^2,\end{aligned}$$

where $P|_\theta$ and $P|_{(a^1,a^2)}$ are the marginal distributions of P. We have shown in Lemma D.1 that in the setting of linear reward functions, the decoupling coefficient is d. Furthermore, if the size of the action space is K, then the decoupling coefficient is bounded by $K(K-1)/2 = \mathcal{O}(K^2)$, shown by Lemma 2 in Zhang (2022).

We now introduce modifications to FGTS.CDB for the nonlinear reward. In accordance with the change of the reward function's structure, the arm selection scheme and the likelihood function are changed to

$$a_j^t = \max_{a\in\mathcal{A}_t} r_{\theta_t^j}(x_t, a),$$
$$L^j(\theta, x, a^1, a^2, y) = \eta\sigma(y\Delta r_\theta(x, a^1, a^2)) - \mu \max_{a'\in\mathcal{A}} \Delta r_\theta(x, a', a^{3-j}).$$

The following theorem characterizes the regret bound of FGTS.CDB with the aforementioned modifications:

Theorem 6.2. *Suppose that Assumptions 5.1 and 6.1 hold, and the hyperparameters are chosen as* $\eta = 0.25$ *and*

$$\mu = \frac{e^{-B}}{5}\sqrt{\frac{-\log(p_0(\theta_*)\bar{\mu}(\{\theta : d(\theta,\theta_*) \le 2/(L_0T)\})}{T\mathrm{dc}}}.$$

Then the regret of FGTS.CDB satisfies

$$\mathbb{E}[\mathrm{Regret}(T)] = \mathcal{O}(\sqrt{-\log(p_0(\theta_*)\bar{\mu}(\{\theta : d(\theta,\theta_*) \le 2/(L_0T)\}) \cdot T\mathrm{dc}}).$$

Remark 6.3. Theorem 6.2 can be reduced to Theorem 5.2 by noting that $-\log(p_0(\theta_*)\bar{\mu}(\{\theta : d(\theta,\theta_*) \le 2/(L_0T)\})) = \widetilde{\mathcal{O}}(d)$, and $\mathrm{dc} = \mathcal{O}(d)$. Furthermore, if we assume that the size of the action space is bounded by K, then the regret bound is $\widetilde{\mathcal{O}}(K\sqrt{dT})$. This regret bound is minimax-optimal if K is treated as a constant, although the regret bounds of existing algorithms, e.g., Sta'D (Saha, 2021) and SupCoLSTIM (Bengs et al., 2022), scale with $\sqrt{\log K}$.

Remark 6.4. Another case of interest is that the reward function has a linear structure, but Θ is a finite set of size N. In this case, if we choose the prior p_0 to be the uniform distribution on Θ, then $-\log p_0(\theta_*) = \log N$, and $-\log(\bar{\mu}(\{\theta : d(\theta,\theta_*) \le 2/L_0T\})) = 0$ when T is large enough. Therefore, the regret bound is $\mathcal{O}(\sqrt{dT\log N})$.

7 Overview of Proof

In this section, we present the key techniques in the proof of Theorem 5.2, and details are given in Appendix A.1. The proof of Theorem 6.2 is similar and is given in Appendix A.3. In Subsection 7.1, we first introduce a special regret decomposition scheme corresponding to our algorithm and discuss the advantage of our Feel-Good exploration term. Then, in Subsection 7.2, we get into details of the analysis based on the difference of potentials between steps.

7.1 Regret Decomposition

The proof for standard Thompson sampling (Zhang, 2022) performs the following regret decomposition:

$$r_*(x_t, a_t^*) - r_*(x_t, a_t) = \underbrace{[r_\theta(x_t, a_t) - r_*(x_t, a_t)]}_{\text{Bellman Error}} - \underbrace{[\max_a r_\theta(x_t, a) - r_*(x_t, a_t^*)]}_{\text{Feel-Good Exploration}},$$

where the Bellman Error term refers to the estimation error of θ evaluated on historic data, and the Feel-Good exploration term refers to the difference between the maximum reward corresponding to θ and θ_*. The Bellman Error term can be bounded using the decoupling technique which converts the joint expectation of the model sampling and trajectory into independent expectations. The Feel-Good Exploration term adopts a crucial structure that does not explicitly contain a_t, so the decoupling trivially applies to the Feel-Good Exploration term. However, the following two challenges arise when studying contextual dueling bandits due to the different definition of the regret: 1. How we can perform the regret decomposition, and 2. whether the Feel-Good Exploration term can also be decoupled.

Challenge 1: Regret decomposition. The starting point of deriving a new regret decomposition is the Bellman Error term, which should correspond to the estimation error on historic data. As the likelihood in the posterior distribution in (4.1) is the inner product between $\boldsymbol{\theta}$ and the difference between the arms, the Bellman Error term is

$$\mathrm{BE}_t^j = \langle \boldsymbol{\theta}_t^j - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^j) - \boldsymbol{\phi}(x_t, a_t^{3-j})\rangle.$$

The remaining part of the regret is the Feel-Good Exploration term, which is

$$\begin{aligned}
&\frac{\mathrm{BE}_t^1 + \mathrm{BE}_t^2}{2} - \left[r_*(x_t, a_t^*) - \frac{r_*(x_t, a_t^1) + r_*(x_t, a_t^2)}{2}\right] \\
&= \frac{1}{2}\Big[\max_{a\in\mathcal{A}_t}\langle \boldsymbol{\theta}_t^1, \boldsymbol{\phi}(x_t, a) - \boldsymbol{\phi}(x_t, a_t^2)\rangle \\
&\qquad - \langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t^2)\rangle\Big] \\
&+ \frac{1}{2}\Big[\max_{a\in\mathcal{A}_t}\langle \boldsymbol{\theta}_t^2, \boldsymbol{\phi}(x_t, a) - \boldsymbol{\phi}(x_t, a_t^1)\rangle \\
&\qquad - \langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t^1)\rangle\Big].
\end{aligned}$$

Therefore, we set the Feel-Good Exploration term to be

$$\mathrm{FG}_t^j(\boldsymbol{\theta}) = \max_{a\in\mathcal{A}_t}\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a) - \boldsymbol{\phi}(x_t, a_t^{3-j})\rangle - \langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t^{3-j})\rangle,$$

and the regret decomposition is

$$\begin{aligned}
&r_*(x_t, a_t^*) - [r_*(x_t, a_t^1) + r_*(x_t, a_t^2)]/2 \\
&= [\mathrm{BE}_t^1 + \mathrm{BE}_t^2 - \mathrm{FG}_t^1(\boldsymbol{\theta}_t^1) - \mathrm{FG}_t^2(\boldsymbol{\theta}_t^2)]/2,
\end{aligned}$$

The regret decomposition inspires the design of the Feel-Good exploration term in the posterior distribution. Without the additional term $\langle \boldsymbol{\theta}_t^j, \boldsymbol{\phi}(x_t, a_t^{3-j})\rangle$ in the likelihood function, the decomposition would contain additional cross terms $\langle \boldsymbol{\theta}_t^j - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^{3-j})\rangle$ that are hard to analyze.

Challenge 2: Decoupling for both the Bellman Error and the Feel-Good Exploration term. We bound BE_t^j using a

decoupling argument that is standard in the literature (Zhang, 2022; Fan & Gu, 2023):

$$\mathbb{E}[\mathrm{BE}_t^j] \le d\lambda + \frac{1}{4\lambda}\mathbb{E}_{S_{t-1},x_t,a_t^1,a_t^2}\Big[\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^j(\cdot|S_{t-1})}\mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})\Big], \tag{7.1}$$

where the inequality holds for any constant $\lambda > 0$, and the least square term is defined as

$$\mathrm{LS}_t(\boldsymbol{\theta}) = \langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle^2.$$

Details of this inequality are given in Lemma D.1.
However, additional challenges arise in the analysis of the Feel-Good exploration term $\mathrm{FG}_t^j(\boldsymbol{\theta}_t^j)$ because it is different from its counterpart in standard contextual bandits, and the property that the Feel-Good Exploration term does not explicitly contain the trajectory no longer holds.
To counter this challange, we make the following two crucial observations:

1. The Feel-Good exploration term $\mathrm{FG}_t^j(\boldsymbol{\theta}_t^j)$ only contains parameter $\boldsymbol{\theta}_t^j$ and the adversarial arm a_t^{3-j}, while a_t^j does not appear in its definition.

2. Conditioned on the history S_{t-1}, $\boldsymbol{\theta}_t^j$ and a_t^{3-j} are independent because a_t^{3-j} is determined by $\boldsymbol{\theta}_t^{3-j}$, which is independent with $\boldsymbol{\theta}_t^j$ according to Algorithm 1.

Therefore, $\boldsymbol{\theta}_t^j$ can also be decoupled with a_t^j:

$$\mathbb{E}[\mathrm{FG}_t^j(\boldsymbol{\theta}_t^j)] = \mathbb{E}_{S_{t-1},x_t,a_t^1,a_t^2}[\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^j(\cdot|S_{t-1})}\mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}})]. \tag{7.2}$$

Combining (7.1) and (7.2), we have

$$\mathbb{E}[\mathrm{BE}_t^j - \mathrm{FG}_t^j(\boldsymbol{\theta})] \le d\lambda + \mathbb{E}_{S_{t-1},x_t,a_t^1,a_t^2}\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^j(\cdot|S_{t-1})} \Big[\mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})/(4\lambda) - \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}})\Big]. \tag{7.3}$$

To further bound (7.3), we use a technique based on the analysis of the potential which is shown in Subsection 7.2.

7.2 Potential-Based Analysis

Following Zhang (2022), we consider a potential Z_t^j for each round, defined as

$$Z_t^j := \mathbb{E}_{S_t}\log\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p_0} W_t^j(\widetilde{\boldsymbol{\theta}}|S_t), \tag{7.4}$$

where

$$W_t^j(\boldsymbol{\theta}|S_t) := \exp\Big(-\sum_{i=1}^t \Delta L^j(\boldsymbol{\theta}, x_i, a_i^1, a_i^2, y_i)\Big), \tag{7.5}$$

$$\Delta L^j(\boldsymbol{\theta}, x, a^1, a^2, y) := L^j(\boldsymbol{\theta}, x, a^1, a^2, y) - L^j(\boldsymbol{\theta}_*, x, a^1, a^2, y). \tag{7.6}$$

Then the posterior distribution satisfies

$$p^j(\widetilde{\boldsymbol{\theta}}|S_{t-1}) = \frac{p_0(\widetilde{\boldsymbol{\theta}})W_{t-1}^j(\widetilde{\boldsymbol{\theta}}|S_{t-1})}{\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p_0}W_{t-1}^j(\widetilde{\boldsymbol{\theta}}|S_{t-1})}. \tag{7.7}$$

Using this expression, we can then obtain

$$Z_t^j - Z_{t-1}^j = \mathbb{E}_{S_t}\log\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^j(\cdot|S_{t-1})}\exp(-\Delta L^j(\widetilde{\boldsymbol{\theta}}, x_t, a_t^1, a_t^2, y_t)).$$

Based on the above property and the design of L^j, we can establish the connection between (7.3) and the potential difference as follows:

Lemma 7.1. *Let Z_t be defined in (7.4), and suppose $\eta \le 1/2$. Then we have*

$$\mathbb{E}_{S_{t-1},x_t,a_t^1,a_t^2}\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^j(\cdot|S_{t-1})}\Big[\frac{e^{-2B}\eta}{18\mu}\mathrm{LS}_t(\widetilde{\boldsymbol{\theta}}) - \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}})\Big] \le \mu^{-1}(Z_{t-1}^j - Z_t^j) + 32\mu B^2,$$

The proof of Lemma 7.1 is detailed in Appendix C. Inspired by Lemma 7.1, we can choose $\lambda = (9\mu e^{2B})/(2\eta)$ in (7.3). Taking the sum over t, noting that $Z_0^j = 0$, through telescope sum we obtain that

$$\sum_{t=1}^T \mathbb{E}[\mathrm{BE}_t^j - \mathrm{FG}_t^j(\boldsymbol{\theta}_t^j)] \le \frac{-Z_T^j}{\mu} + \mu T\left(\frac{9de^{2B}}{2\eta} + 32B^2\right).$$

It then suffices to derive an upper bound for $-Z_T^j$, which is characterized by the following lemma:

Lemma 7.2. *Let Z_T^j be defined by (7.4), then we have*

$$-Z_T^j = \widetilde{\mathcal{O}}(d), \quad j = 1, 2.$$

Lemma 7.2 shows that the upper bound of $-Z_t^j$ only contains logarithmic factors of T despite the sum of T terms in its definition. The proof of Lemma 7.2 is shown in Appendix B.2 and uses a technique similar to that of Section 5.2 in Zhang (2022). Finally, by selecting $\eta = \Theta(1)$ and $\mu = \Theta(1/\sqrt{T})$, we can obtain the regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T})$.

8 Experiments

In this section, we investigate the performance of FGTS.CDB through simulation, comparing it with other efficient algorithms proposed for contextual dueling bandits. For each experiment, we run $T = 2500$ rounds. The underlying unknown parameter $\boldsymbol{\theta}^*$ is randomly generated and normalized to a unit vector. The dimension of feature vectors is set to $d = 5, 10, 15$. We generate a total of $|\mathcal{A}_t| = 32$ distinct arms with feature vectors randomly chosen from $\{-1, 1\}^d$ following the uniform distribution. In every round, given the arm pair selected by the algorithm, a response is generated following the random process described in Section 3. Each experiment comprises 10 independent runs.

We report and plot the average cumulative regret in Figure 1, along with the standard deviation shown in the shaded region. For simplicity, we choose the logistic function $\sigma(\cdot)$ as the link function.

8.1 Implementation Details of Different Algorithms

MaxInP. The maximum informative pair method introduced by Saha (2021) maintains an active set of potential optimal arms in each round. Pairs are selected based on maximizing the uncertainty in the difference between the two arms. Instead of incorporating a warm-up period τ_0 as part of their definitions, we initialize the covariance matrix $\mathbf{\Sigma}_0 = \lambda \mathbf{I}$ for regularization. Empirically, we found that when λ is set to 0.001, this approach shows no substantial impact on the performance compared to the warm-up strategy.

MaxPairUCB. In this algorithm, we use the same MLE estimator as that of MaxInP. However, we use a different arm selection scheme defined as follows:

$$(\mathbf{x}_t, \mathbf{y}_t) = \underset{(\mathbf{x},\mathbf{y})\in\mathcal{A}_t\times\mathcal{A}_t}{\operatorname{argmax}} \left[\langle \widehat{\boldsymbol{\theta}}_t, \mathbf{x}_t + \mathbf{y}_t\rangle + \beta\|\mathbf{x}_t - \mathbf{y}_t\|_{\widehat{\mathbf{\Sigma}}_t^{-1}}\right],$$

which is a variant of MaxInP without the need for an arm elimination phase proposed by Di et al. (2023), where $\mathcal{A}_t$ is the decision set at round t.

CoLSTIM. This technique is proposed in Bengs et al. (2022). Initially, they augment each arm with randomly perturbed utilities and select the arm with the best estimation. They argue that this procedure yields improved empirical performance. The second arm is then selected by maximizing the sum of the estimated reward and the uncertainty term proposed by Saha (2021).

VACDB. This approach, proposed by Di et al. (2023), adopts a layered arm elimination process. Arms in higher layers have lower estimation uncertainty, and the elimination process starts when all arms within a layer have lower uncertainty. The arm selection scheme is similar to that of MaxPairUCB, where arms are selected to maximize the same weighted sum within a certain layer.

FGTS.CDB. The proposed sampling-based algorithm for contextual dueling bandits in this paper as presented in Algorithm 1. η and μ is set to 1 and $T^{-1/2}\cdot\alpha$, respectively. Empirically, we generate the model parameter $\{\boldsymbol{\theta}_t^j\}_{j=1,2}$ by repeatedly taking the following stochastic gradient Langevin dynamics (SGLD) step:

$$\boldsymbol{\theta}_t^j \leftarrow \boldsymbol{\theta}_t^j + \sqrt{2\delta}\boldsymbol{\xi}_i - \delta\nabla_{\boldsymbol{\theta}}\underbrace{\left[\sum_{\tau=1}^{t-1} L^j(\boldsymbol{\theta}_t^j, x_\tau, a_\tau^1, a_\tau^2, y_\tau) - \ln p_0(\boldsymbol{\theta}_t^j)\right]}_{I(\boldsymbol{\theta}_t^j)}, \tag{8.1}$$

where $\boldsymbol{\xi}_i \sim \mathcal{N}(0, \mathbf{I}_d)$. If we set the step size δ to a sufficiently small number, the dynamic of $\boldsymbol{\theta}_t^j$ can be regarded as the solution of the following stochastic differential equation:

$$\mathrm{d}\boldsymbol{\theta}_s = -\nabla I(\boldsymbol{\theta}_s)\mathrm{d}s + \sqrt{2}\mathrm{d}\mathbf{W}_s,$$

where W_s is the Brownian motion in $\mathbb{R}^d$ with $W_0 = 0$. We denote by $q_s(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ the distribution of $\boldsymbol{\theta}_s$ at time s. Then it is known that q_s satisfies the following Fokker-Planck equation:

$$\frac{\partial q_s}{\partial s} = \nabla\cdot(q_s\nabla I(\boldsymbol{\theta}_s)) + \lambda\Delta q_s.$$

The stationary distribution of $\boldsymbol{\theta}_t^j$ after SGLD iterations can then be derived by setting $\partial q_s/\partial s = 0$, from which we have $\boldsymbol{\theta}_t^j$ converges to a sample from $p^j(\boldsymbol{\theta}) \propto \exp(-I(\boldsymbol{\theta}))$.
In our experiment, we implement (8.1) with initial step size $\delta = 0.005$. After each round, we schedule the step size δ with decaying rate 0.99 to stabilize the optimization process.

8.2 Regret Comparison

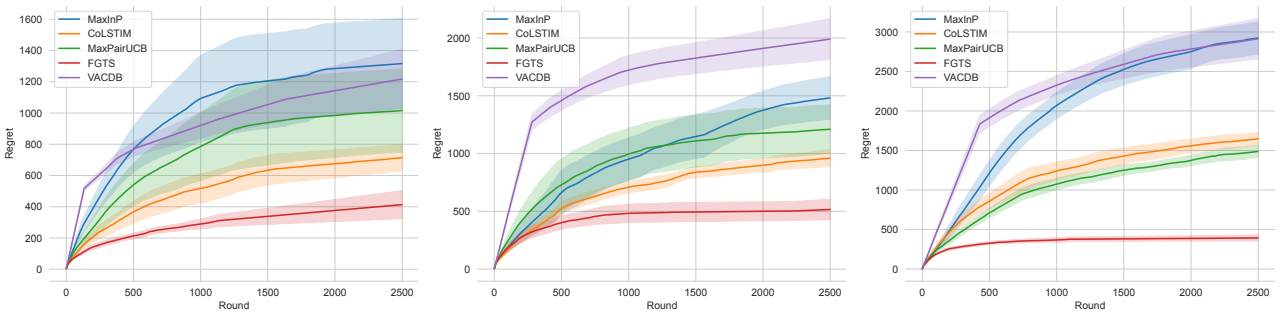
We plot the regret with respect to the number of rounds in Figure 1. The results are averaged over 30 trials. In Figure 1, we run FGTS.CDB with $\alpha = 0.1$. For the benchmarks, we select the hyperparameters, including the conficence radius in MaxInP and MaxPairUCB and the magnitude of perturbations in CoLSTIM, to be the best-performing hyperparameter within $\{10^{-2}, 10^{-1}, 10^0, 10^1\}$. It is shown that, FGTS.CDB outperforms the previous algorithms by a large margin with dimension d set to 5, 10, 15. Additionally, the standard deviation of FGTS.CDB among different trials is also the smallest according to our experiments, which indicates that FGTS.CDB is more stable under different random seeds.

8.3 Ablation Study

One benefit of FGTS.CDB in empirical tasks is that the hyperparameters $\mu = T^{-1/2}\cdot\alpha$ and η are independent of the feature dimension d, simplifying the tuning of parameters. In contrast, the confidence radius in UCB-based algorithms usually need to be carefully tuned to achieve the desirable performance (Lattimore & Szepesvári, 2020). In Figure 2, we study the performance of FGTS.CDB when $d = 5, 10, 15$ under different values of α. It is observed that the regret of FGTS.CDB is robust to different values of α. Surprisingly, it turns out that FGTS.CDB performs well even without Feel-Good exploration ($\alpha = 0$). We conjecture that a Thompson-sampling-based algorithm may also work for contextual dueling bandit setting due to its stochastic nature, which also encourages the agent to explore different pairs of arms. We leave the study of Thompson-sampling without Feel-Good exploration under contextual dueling bandit setting for future work.

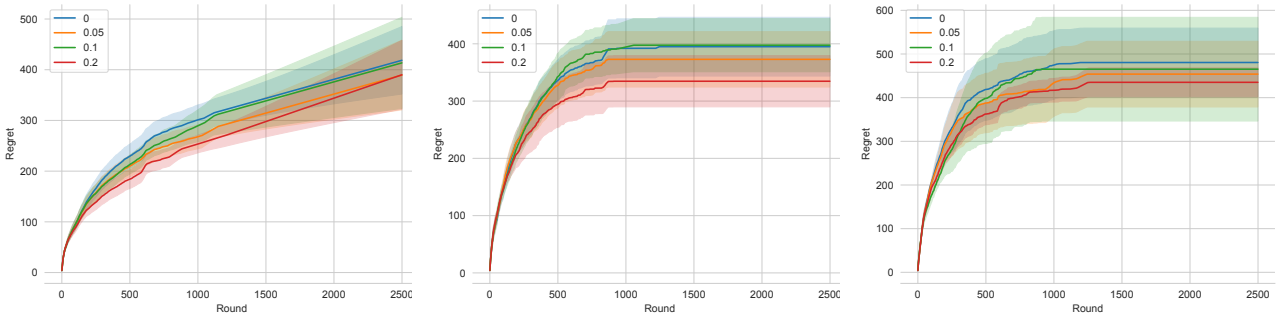
9 Conclusion

In this work, we apply the technique of Feel-Good Thompson sampling to the setting of contextual dueling bandits.



(a) $d = 5$ (b) $d = 10$ (c) $d = 15$

Figure 1. Regret comparison with MaxInP, MaxPairUCB and CoLSTIM.



(a) $d = 5$ (b) $d = 10$ (c) $d = 15$

Figure 2. Performance of FGTS.CDB over different α.

We propose an algorithm, FGTS.CDB, whose pivotal design is a Feel-Good exploration term that contains an additional term representing the reward of the adversarial arm, compared with standard Thompson sampling. We show that our algorithm achieves a nearly minimax-optimal regret bound. Furthermore, experiments on synthetic data show that the performance of our algorithm based on FGTS is comparable with UCB-based algorithms. As a future direction, it is interesting to explore the possibility of variance-aware algorithms based on the FGTS technique. The extension of our algorithm to the setting of preference-based reinforcement learning is also an interesting topic to study.

Acknowledgements

We thank the anonymous reviewers and area chair for their helpful comments. XL, HZ and QG are supported in part by the NSF grant DMS-2323113, Cisco Research Award, and the Sloan Research Fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

Impact Statement

This paper presents work whose goal is to advance the field of bandit algorithms. Our algorithm can be applied in a lot of RLHF scenarios including finetuning LLMs, but we don't think the social impacts must be specifically highlighted here.

References

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

Agarwal, A. and Zhang, T. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *Advances in Neural Information Processing Systems*, 35:35284–35297, 2022.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Agrawal, S. and Goyal, N. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5): 1–24, 2017.

Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Balsubramani, A., Karnin, Z., Schapire, R. E., and Zoghi, M. Instance-dependent regret bounds for dueling bandits. In *Conference on Learning Theory*, pp. 336–360. PMLR, 2016.

Bengs, V., Saha, A., and Hüllermeier, E. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pp. 1764–1786. PMLR, 2022.

Brandt, J., Bengs, V., Haddenhorst, B., and Hüllermeier, E. Finding optimal arms in non-stochastic combinatorial bandits with semi-bandit feedback and finite budget. *Advances in Neural Information Processing Systems*, 35: 20621–20634, 2022.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Di, Q., Jin, T., Wu, Y., Zhao, H., Farnoud, F., and Gu, Q. Variance-aware regret bounds for stochastic contextual dueling bandits. *arXiv preprint arXiv:2310.00968*, 2023.

Diao, S., Pan, R., Dong, H., Shum, K. S., Zhang, J., Xiong, W., and Zhang, T. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. *arXiv preprint arXiv:2306.12420*, 2023.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. *ArXiv*, abs/1502.06362, 2015.

Falahatgar, M., Orlitsky, A., Pichapati, V., and Suresh, A. T. Maximum selection and ranking under noisy comparisons. In *International Conference on Machine Learning*, pp. 1088–1096. PMLR, 2017.

Fan, Z. and Gu, Q. The power of feel-good thompson sampling: A unified framework for linear bandits, 2023.

Heckel, R., Shah, N. B., Ramchandran, K., and Wainwright, M. J. Active ranking from pairwise comparisons and when parametric assumptions do not help. *arXiv: Learning*, 2016.

Hunter, D. R. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

Jamieson, K., Katariya, S., Deshpande, A., and Nowak, R. Sparse dueling bandits. In *Artificial Intelligence and Statistics*, pp. 416–424. PMLR, 2015.

Ji, X., Wang, H., Chen, M., Zhao, T., and Wang, M. Provable benefits of policy learning from human preferences in contextual bandit problems. *arXiv preprint arXiv:2307.12975*, 2023.

Jin, T., Xu, P., Shi, J., Xiao, X., and Gu, Q. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Khetan, A. and Oh, S. Data-driven rank breaking for efficient rank aggregation. In *International Conference on Machine Learning*, pp. 89–98. PMLR, 2016.

Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pp. 1141–1154. PMLR, 2015.

Komiyama, J., Honda, J., and Nakagawa, H. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *International Conference on Machine Learning*, pp. 1235–1244. PMLR, 2016.

Kumagai, W. Regret analysis for continuous dueling bandit. *Advances in Neural Information Processing Systems*, 30, 2017.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

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

Luce, R. D. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

Novoseller, E., Wei, Y., Sui, Y., Yue, Y., and Burdick, J. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020.

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike,

J., and Lowe, R. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022.

Ramamohan, S., Rajkumar, A., and Agarwal, S. Dueling bandits: Beyond condorcet winners to general tournament solutions. In *NIPS*, 2016.

Ren, W., Liu, J., and Shroff, N. B. Pac ranking from pairwise and listwise queries: Lower bounds and upper bounds. *arXiv preprint arXiv:1806.02970*, 2018.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

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

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

Saha, A., Koren, T., and Mansour, Y. Adversarial dueling bandits. *ArXiv*, abs/2010.14563, 2021.

Soufiani, H. A., Parkes, D., and Xia, L. Computing parametric ranking models via rank-breaking. In *International Conference on Machine Learning*, pp. 360–368. PMLR, 2014.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253*, 2017.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Wu, H. and Liu, X. Double thompson sampling for dueling bandits. *Advances in neural information processing systems*, 29, 2016.

Wu, Y., Jin, T., Lou, H., Farnoud, F., and Gu, Q. Borda regret minimization for generalized linear dueling bandits. *arXiv preprint arXiv:2303.08816*, 2023.

Xiong, W., Zhong, H., Shi, C., Shen, C., and Zhang, T. A self-play posterior sampling algorithm for zero-sum markov games. In *International Conference on Machine Learning*, pp. 24496–24523. PMLR, 2022.

Xiong, W., Dong, H., Ye, C., Zhong, H., Jiang, N., and Zhang, T. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*, 2023.

Yue, Y., Broder, J., Kleinberg, R. D., and Joachims, T. The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78:1538–1556, 2012.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Zhu, B., Jiao, J., and Jordan, M. I. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.

Zoghi, M., Whiteson, S., Munos, R., and de Rijke, M. Relative upper confidence bound for the k-armed dueling bandit problem. *ArXiv*, abs/1312.3393, 2014.

Zoghi, M., Karnin, Z. S., Whiteson, S., and de Rijke, M. Copeland dueling bandits. In *NIPS*, 2015.

A Proof of Main Results

A.1 Proof of Theorem 5.2

We first restate Theorem 5.2 more precisely:

Theorem A.1 (Restatement of Theorem 5.2)**.** *Assume that the hyperparameters are selected as $\mu = 0.25$ and $\mu = 1/(10e^B\sqrt{T})$, and the logarithm of the prior distribution p_0 is L-Lipschitz. Then the expected regret is bounded by*

$$\sum_{t=1}^{T}\mathbb{E}\left[r_*(x_t,a_t^*)-\frac{r_*(x_t,a_t^1)+r_*(x_t,a_t^2)}{2}\right]\le 10e^B d\sqrt{T}\left[2-\frac{\log p_0(\boldsymbol{\theta}_*)}{d}+\log\left(\frac{L}{\sqrt{d}}+\frac{1}{5e^B}\sqrt{\frac{T}{d}}+\frac{T}{2\sqrt{d}}\right)\right].$$

The following Lemma bounds the regret with terms related with Z_T^j:

Lemma A.2. *Under Assumption 3.1, if $\eta \le 1/2$, then we have*

$$\sum_{t=1}^{T}\mathbb{E}\left[r_*(x_t,a_t^*)-\frac{r_*(x_t,a_t^1)+r_*(x_t,a_t^2)}{2}\right]\le\left(\frac{9de^{2B}}{2\eta}+32B^2\right)\mu T-\frac{Z_T^1+Z_T^2}{2\mu}.$$

We need the following lemma to characterize Z_T^j in Lemma A.2:

Lemma A.3 (Restatement of Lemma 7.2)**.** *Assume that the log of the prior distribution p_0 is L-Lipschitz. Then for $j = 1, 2$, we have*

$$Z_T^j\ge\log p_0(\boldsymbol{\theta}_*)-d-d\log\frac{L+2(\mu+\eta)T}{\sqrt{d}}.$$

The proofs of Lemma A.2 and Lemma A.3 are presented in Section B.1 and Section B.2, respectively. We now present the proof of Theorem 5.2:

Proof of Theorem 5.2. Combining Lemma A.2 and Lemma A.3, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[r_*(x_t,a_t^*)-\frac{r_*(x_t,a_t^1)+r_*(x_t,a_t^2)}{2}\right]\\
&\le\frac{9de^{2B}}{2\eta}\left(1+\frac{64B^2\eta}{de^{2B}}\right)\mu T+\frac{1}{\mu}\left[d-\log p_0(\boldsymbol{\theta}_*)+d\log\frac{L+2(\mu+\eta)T}{\sqrt{d}}\right]\\
&\le\frac{9de^{2B}}{2\eta}\left(1+\frac{32}{e^2}\right)\mu T+\frac{1}{\mu}\left[d-\log p_0(\boldsymbol{\theta}_*)+d\log\frac{L+2(\mu+\eta)T}{\sqrt{d}}\right]\\
&\le\frac{25e^{2B}dT\mu}{\eta}+\frac{1}{\mu}\left[d-\log p_0(\boldsymbol{\theta}_*)+d\log\frac{L+2(\mu+\eta)T}{\sqrt{d}}\right]\\
&=10e^Bd\sqrt{T}\left[2-\frac{\log p_0(\boldsymbol{\theta}_*)}{d}+\log\left(\frac{L}{\sqrt{d}}+\frac{1}{5e^B}\sqrt{\frac{T}{d}}+\frac{T}{2\sqrt{d}}\right)\right],
\end{aligned}$$

where the first inequality holds due to Lemma A.2 and Lemma A.3, the second inequality holds because $B/e^B \le 1/e$, the third inequality holds because $9/2\cdot(1+32/e^2)\le 25$, and the equality holds by substituting $\eta = 0.25$ and $\mu = 1/(10e^B\sqrt{T})$. □

A.2 Proof of Theorem 5.3

In this section, we provide the proof for Theorem 5.3. Instead of applying using the proof that is similar to the case of bounded ℓ_∞ norm, which is the approach used in the proof of Theorem 3.1 of (Bengs et al., 2022), we follow the proof for the case of bounded ℓ_2-norm in the standard contextual bandit setting (Lattimore & Szepesvári, 2020).

Notations. For $\boldsymbol{\theta}\in\Theta$, let $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{E}_{\boldsymbol{\theta}}$ be the distribution and expectation over the trajectory generated by the algorithm, respectively. Let ϕ_{ti}^j be the shorthand notation for $(\boldsymbol{\phi}(x_t,a_t^j))_i$. For two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, let $D_{\mathrm{KL}}(\mathbb{P}_1||\mathbb{P}_2)$ be their KL-divergence.

Proof of Theorem 5.3. We fix $i\in[d]$, and define

$$\tau_i\coloneqq T\wedge\min\left\{\tau:\sum_{t=1}^{\tau}[(\phi_{ti}^1)^2+(\phi_{ti}^2)^2]\ge\frac{2T}{d}\right\}$$

For $x \in \{\pm 1\}$, we define

$$U_{\boldsymbol{\theta},i}(x) = \mathbb{E}_{\boldsymbol{\theta}}\bigg[\sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^1 x\bigg)^2 + \sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^2 x\bigg)^2\bigg],$$

We fix $\boldsymbol{\theta} \in \Theta = \{\pm\Delta\}^d$ for some Δ to be determined. For the fixed i, define $\boldsymbol{\theta}'$ to be the vector such that $\theta_i' = -\theta_i$ and $\theta_j' = \theta_j$ for all $j \neq i$. We then have

$$U_{\boldsymbol{\theta},i}(\mathrm{sign}(\theta_i)) + U_{\boldsymbol{\theta}',i}(\mathrm{sign}(\theta_i')) = \underbrace{U_{\boldsymbol{\theta},i}(\mathrm{sign}(\theta_i)) - U_{\boldsymbol{\theta}',i}(\mathrm{sign}(\theta_i))}_{I_1} + \underbrace{U_{\boldsymbol{\theta}',i}(\mathrm{sign}(\theta_i)) + U_{\boldsymbol{\theta}',i}(\mathrm{sign}(\theta_i'))}_{I_2}.$$

For I_1, note that

$$\begin{aligned}
&\sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^1\,\mathrm{sign}(\theta_i)\bigg)^2 + \sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^2\,\mathrm{sign}(\theta_i)\bigg)^2 \\
&= \frac{2\tau_i}{d} - \frac{2\,\mathrm{sign}(\theta_i)}{\sqrt{d}}\sum_{t=1}^{\tau_i}(\phi_{ti}^1 + \phi_{ti}^2) + \sum_{t=1}^{\tau_i}[(\phi_{ti}^1)^2 + (\phi_{ti}^2)^2] \\
&\le \frac{4\tau_i}{d} + 2\sum_{t=1}^{\tau_i}[(\phi_{ti}^1)^2 + (\phi_{ti}^2)^2] \\
&\le \frac{4T}{d} + 2\cdot\frac{2T}{d} + 2\cdot(1+1) = \frac{8T}{d} + 4,
\end{aligned}$$

where first inequality holds due to Cauchy-Schwarz inequality, and the second inequality holds due to the definition of τ_i and $|\phi_{ti}^j| \le 1$. We thus bound I_1 as follows:

$$\begin{aligned}
I_1 &= \mathbb{E}_{\boldsymbol{\theta}}\bigg[\sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^1\,\mathrm{sign}(\theta_i)\bigg)^2 + \sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^2\,\mathrm{sign}(\theta_i)\bigg)^2\bigg] \\
&\quad - \mathbb{E}_{\boldsymbol{\theta}'}\bigg[\sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^1\,\mathrm{sign}(\theta_i)\bigg)^2 + \sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^2\,\mathrm{sign}(\theta_i)\bigg)^2\bigg] \\
&\ge -4(2T/d+1)\sqrt{D_{\mathrm{KL}}(\mathrm{Ber}(\mathbb{P}_\theta||\mathbb{P}_{\theta'}))/2} \\
&\ge -(2T/d+1)\sqrt{\mathbb{E}_{\boldsymbol{\theta}}\bigg[\sum_{t=1}^{\tau_i}\langle\boldsymbol{\theta}-\boldsymbol{\theta}', \boldsymbol{\phi}(x_t,a_t^1) - \boldsymbol{\phi}(x_t,a_t^2)\rangle^2\bigg]} \\
&\ge -2\Delta(2T/d+1)\sqrt{2\mathbb{E}_{\boldsymbol{\theta}}\bigg[\sum_{t=1}^{\tau_i}((\phi_{ti}^1)^2 + (\phi_{ti}^2)^2)\bigg]} \\
&\ge -4\Delta(2T/d+1)\sqrt{T/d+1} \\
&\ge -12\sqrt{2}\Delta(T/d)^{1.5},
\end{aligned} \tag{A.1}$$

where the first ineequality holds due to Pinker's inequality, the second inequality holds due to Lemma D.4, the third inequality holds due to the definition of $\boldsymbol{\theta}'$ and because $|\phi_{ti}^1 - \phi_{ti}^2|^2 \le 2(\phi_{ti}^1)^2 + 2(\phi_{ti}^2)^2$, the fourth inequality holds due to the definition of τ_i and $|\phi_{ti}^j| \le 1$, and the last inequality holds because $T \ge d$. For I_2, we have

$$\begin{aligned}
I_2 &= \mathbb{E}_{\boldsymbol{\theta}'}\bigg[\sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^1\bigg)^2 + \sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} - \phi_{ti}^2\bigg)^2\bigg] + \mathbb{E}_{\theta'}\bigg[\sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} + \phi_{ti}^1\bigg)^2 + \sum_{t=1}^{\tau_i}\bigg(\frac{1}{\sqrt{d}} + \phi_{ti}^2\bigg)^2\bigg] \\
&= \frac{4\tau_i}{d} + 2\mathbb{E}_{\boldsymbol{\theta}'}\bigg[\sum_{t=1}^{\tau_i}[(\phi_{ti}^1)^2 + (\phi_{ti}^2)^2]\bigg] \ge \frac{4T}{d},
\end{aligned} \tag{A.2}$$

where the inequality holds due to the definition of τ_i. Combining (A.1) and (A.2), we have

$$U_{\boldsymbol{\theta},i}(\mathrm{sign}(\theta_i)) + U_{\boldsymbol{\theta}',i}(\mathrm{sign}(\theta_i')) \ge \frac{4T}{d} - 12\sqrt{2}\Delta(T/d)^{1.5}.$$

Thus, by pairing each $\boldsymbol{\theta}$ with the corresponding $\boldsymbol{\theta}'$ for each i, we have

$$\frac{1}{|\Theta|}\sum_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{d} U_{\boldsymbol{\theta},i}(\mathrm{sign}(\theta_i)) \geq 2T - 6\sqrt{2}\Delta T^{1.5} d^{-0.5}.$$

Therefore, there exists $\boldsymbol{\theta} \in \Theta$ such that

$$\sum_{i=1}^{d} U_{\boldsymbol{\theta},i}(\mathrm{sign}(\theta_i)) \geq 2T - 6\sqrt{2}\Delta T^{1.5} d^{-0.5}. \tag{A.3}$$

For this $\boldsymbol{\theta}$, the regret is

$$\mathrm{Regret}(T) = \frac{\Delta}{2}\cdot\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}\left(\frac{1}{\sqrt{d}} - \phi_{ti}^{1}\,\mathrm{sign}(\theta_i)\right)\right] + \frac{\Delta}{2}\cdot\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}\left(\frac{1}{\sqrt{d}} - \phi_{ti}^{2}\,\mathrm{sign}(\theta_i)\right)\right]. \tag{A.4}$$

Note that

$$\begin{aligned}
\sum_{i=1}^{d}\left(\frac{1}{\sqrt{d}} - \phi_{ti}^{j}\,\mathrm{sign}(\theta_i)\right) &= \sum_{i=1}^{d}\left(\frac{1}{2\sqrt{d}} - \phi_{ti}^{j}\,\mathrm{sign}(\theta_i)\right) + \frac{\sqrt{d}}{2} \\
&\geq \sum_{i=1}^{d}\left(\frac{1}{2\sqrt{d}} - \phi_{ti}^{j}\,\mathrm{sign}(\theta_i)\right) + \frac{\sqrt{d}}{2}\sum_{i=1}^{d}(\phi_{ti}^{j})^2 \\
&= \frac{\sqrt{d}}{2}\sum_{i=1}^{d}\left(\frac{1}{\sqrt{d}} - \phi_{ti}^{j}\,\mathrm{sign}(\theta_i)\right)^2,
\end{aligned} \tag{A.5}$$

where the inequality holds because $\sum_{i=1}^{d}(\phi_{ti}^{j})^2 \leq 1$. Plugging (A.5) into (A.4), we have

$$\begin{aligned}
\mathrm{Regret}(T) &\geq \frac{\Delta\sqrt{d}}{4}\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T}\left(\frac{1}{\sqrt{d}} - \phi_{ti}^{1}\,\mathrm{sign}(\theta_i)\right)^2 + \sum_{t=1}^{T}\left(\frac{1}{\sqrt{d}} - \phi_{ti}^{1}\,\mathrm{sign}(\theta_i)\right)^2\right] \\
&\geq \frac{\Delta\sqrt{d}}{4}\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{\tau_i}\left(\frac{1}{\sqrt{d}} - \phi_{ti}^{1}\,\mathrm{sign}(\theta_i)\right)^2 + \sum_{t=1}^{\tau_i}\left(\frac{1}{\sqrt{d}} - \phi_{ti}^{1}\,\mathrm{sign}(\theta_i)\right)^2\right] \\
&= \frac{\Delta\sqrt{d}}{4}\sum_{i=1}^{d} U_{\boldsymbol{\theta},i}(\mathrm{sign}(\theta_i)) \\
&\geq \frac{\Delta\sqrt{d}}{2}(T - 3\sqrt{2}\Delta T^{1.5} d^{-0.5}),
\end{aligned}$$

where the second inequality holds because $\tau_i \leq T$, and the last inequality holds due to (A.3). Let $\Delta = \frac{1}{6}\sqrt{\frac{d}{2T}}$, then

$$\mathrm{Regret}(T) \geq \frac{d\sqrt{T}}{24\sqrt{2}}.$$

□

A.3 Proof of Theorem 6.2

Similar to the proof of Theorem 5.2, we make the following notations:

$$\begin{aligned}
\mathrm{BE}_t^j &:= \Delta r_{\theta_t^j}(x_t, a_t^j, a_t^{3-j}) - \Delta r_{\theta_*}(x_t, a_t^j, a_t^{3-j}), \\
\mathrm{FG}_t^j(\theta) &= \max_{a\in\mathcal{A}} \Delta r_{\theta}(x_t, a, a_t^{3-j}) - \Delta r_{\theta_*}(x_t, a_t^*, a_t^{3-j}), \\
\mathrm{LS}_t(\theta) &= (\Delta r_{\theta}(x_t, a_t^j, a_t^{3-j}) - \Delta r_{\theta_*}(x_t, a_t^j, a_t^{3-j}))^2.
\end{aligned}$$

We first present the following restatement of Theorem 6.2:

Theorem A.4 (Restatement of Theorem 6.2)**.** *Assume that the hyperparameters are selected as* $\eta = 0.25$*, and*

$$\mu = \frac{e^{-B}}{5}\sqrt{\frac{\log 1/[p_0(\theta_*)\bar{\mu}(\{\theta \in \Theta : d(\theta,\theta_*) \le 2/(L_0T)\})]}{T\mathrm{dc}}},$$

We also assume that the $\log p_0$ *is L-Lipschitz and that* r_θ *is* L_0*-Lipschitz in* θ*. Then the regret is bound by*

$$\mathbb{E}[\mathrm{Regret}(T)] \le 4 + 5e^B\sqrt{T\mathrm{dc}\log 1/[p_0(\theta_*)\bar{\mu}(\{\theta \in \Theta : d(\theta,\theta_*) \le 2/(L_0T)\})]} \cdot \left[2 + \frac{1 + 2L/(L_0T)}{\log 1/[p_0(\theta_*)\bar{\mu}(\{\theta \in \Theta : d(\theta,\theta_*) \le 2/(L_0T)\})]}\right].$$

Note that Lemma B.1 does not require linearity of the reward function and can be directly applied in the proof. We only require the following lemma as a counterpart of Lemma A.3:

Lemma A.5. *Assume that* $\log p_0$ *is L-Lipschitz and that* r_θ *is* L_0*-Lipschitz in* θ*. Then for* $j = 1, 2$*, we have*

$$Z_T^j \ge \log p_0(\theta_*) + \log \bar{\mu}(\{\theta \in \Theta : d(\theta,\theta_*) \le 2/(L_0T)\}) - \frac{2L}{L_0T} - 4(\eta + \mu).$$

The proof of Lemma A.5 is given in Appendix B.3. We now present the proof of Theorem A.4:

Proof of Theorem A.4. BE_t^j can be bounded as

$$\begin{aligned}\mathbb{E}[\mathrm{BE}_t^j - \mathrm{FG}_t^j(\theta_t^j)] &\le \frac{9e^{2B}\mu\mathrm{dc}}{2\eta} + \mathbb{E}_{S_{t-1},x_t,a_t^1,a_t^2}\mathbb{E}_{\widetilde{\theta}\sim p^j(\cdot|S_{t-1})}\left[\frac{e^{-2B}\eta}{18\mu}\mathrm{LS}_t(\widetilde{\theta}) - \mathrm{FG}_t^j(\widetilde{\theta})\right] \\ &\le \frac{9e^{2B}\mu\mathrm{dc}}{2\eta} + \frac{Z_{t-1}^j - Z_t^j}{\mu} + 32\mu B^2,\end{aligned}$$

where the first inequality holds due to the definition of dc, and the second inequality holds due to Lemma B.1. Taking the sum over t and substituting $\eta = 0.25$, we have

$$\begin{aligned}&\sum_{t=1}^T \mathbb{E}[\mathrm{BE}_t^j - \mathrm{FG}_t^j(\theta_t^j)] \le (18e^{2B}\mathrm{dc} + 32B^2)\mu T + \frac{e^{-2B}\epsilon}{72\mu} - \frac{Z_T^j}{\mu} \le (18 + 32/e^2)e^{2B}\mathrm{dc}\mu T - \frac{Z_T^j}{\mu} \\ &\le 25e^{2B}\mathrm{dc}\mu T + \frac{1}{\mu} + \frac{2L}{L_0T\mu} + 4 - \frac{\log p_0(\theta_*)}{\mu} + \frac{\log 1/\bar{\mu}(\{\theta \in \Theta : d(\theta,\theta_*) \le 2/(L_0T)\})}{\mu},\end{aligned}$$

where the second inequality holds because $e^{-B}B \le 1/e$, the second inequality holds because $18 + 32/e^2 \le 25$, and the last inequality holds due to Lemma A.5. Taking

$$\mu = \frac{e^{-B}}{5}\sqrt{\frac{\log 1/[p_0(\theta_*)\bar{\mu}(\{\theta \in \Theta : d(\theta,\theta_*) \le 2/(L_0T)\})]}{T\mathrm{dc}}},$$

then we have

$$\sum_{t=1}^T \mathbb{E}[\mathrm{BE}_t^j - \mathrm{FG}_t^j(\theta_t^j)] \le 4 + 5e^B\sqrt{T\mathrm{dc}\log 1/[p_0(\theta_*)\bar{\mu}(\{\theta \in \Theta : d(\theta,\theta_*) \le 2/(L_0T)\})]} \cdot \left[2 + \frac{1 + 2L/(L_0T)}{\log 1/[p_0(\theta_*)\bar{\mu}(\{\theta \in \Theta : d(\theta,\theta_*) \le 2/(L_0T)\})]}\right].$$

□

B Proof of Lemmas in Appendix A

B.1 Proof of Lemma A.2

In order to prove Lemma A.2, we first present the following lemma, which connects LS_t and FG_t^j with the difference of the potential Z_t^j:

Lemma B.1 (Restatement of Lemma 7.1)**.** *Under the same assumptions as Lemma A.2, we have*

$$\mathbb{E}_{S_{t-1},x_t,a_t^1,a_t^2}\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^j(\cdot|S_{t-1})}\left[\frac{e^{-2B}\eta}{18\mu}\mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})-\mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}})\right]\le\mu^{-1}(Z_{t-1}^j-Z_t^j)+32\mu B^2.$$

The proof of Lemma B.1 is given in Appendix C. We now present the proof of Lemma A.2:

Proof of Lemma A.2. The regret at step t can be decomposed as

$$\begin{aligned}
&r_*(x_t,a_t^*)-\frac{r_*(x_t,a_t^1)+r_*(x_t,a_t^2)}{2}=\mathbb{E}\Big\langle\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^*)-\frac{\boldsymbol{\phi}(x_t,a_t^1)+\boldsymbol{\phi}(x_t,a_t^2)}{2}\Big\rangle\\
&=\frac{1}{2}\underbrace{\left[\langle\boldsymbol{\theta}_t^1-\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^1)-\boldsymbol{\phi}(x_t,a_t^2)\rangle-\langle\boldsymbol{\theta}_t^1,\boldsymbol{\phi}(x_t,a_t^1)-\boldsymbol{\phi}(x_t,a_t^2)\rangle+\langle\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^*)-\boldsymbol{\phi}(x_t,a_t^2)\rangle\right]}_{I_1}\\
&\quad+\frac{1}{2}\underbrace{\left[\langle\boldsymbol{\theta}_t^2-\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^2)-\boldsymbol{\phi}(x_t,a_t^1)\rangle-\langle\boldsymbol{\theta}_t^2,\boldsymbol{\phi}(x_t,a_t^2)-\boldsymbol{\phi}(x_t,a_t^1)\rangle+\langle\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^*)-\boldsymbol{\phi}(x_t,a_t^1)\rangle\right]}_{I_2}.
\end{aligned}\tag{B.1}$$

The expectation of I_1 can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\langle\boldsymbol{\theta}_t^1-\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^1)-\boldsymbol{\phi}(x_t,a_t^2)\rangle-\langle\boldsymbol{\theta}_t^1,\boldsymbol{\phi}(x_t,a_t^1)-\boldsymbol{\phi}(x_t,a_t^2)\rangle+\langle\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^*)-\boldsymbol{\phi}(x_t,a_t^2)\rangle\right]\\
&=\mathbb{E}_{S_{t-1},x_t}\mathbb{E}_{\boldsymbol{\theta}_t^1,a_t^1,a_t^2|S_{t-1}}\langle\boldsymbol{\theta}_t^1-\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^1)-\boldsymbol{\phi}(x_t,a_t^2)\rangle-\mathbb{E}_{S_{t-1},x_t}\mathbb{E}_{\boldsymbol{\theta}_t^1,a_t^2|S_{t-1}}\mathrm{FG}_t^1(\boldsymbol{\theta}_t^1)\\
&\le\frac{9e^{2B}\mu d}{2\eta}+\frac{e^{-2B}\eta}{18\mu}\mathbb{E}_{S_{t-1},x_t}\mathbb{E}_{a_t^1,a_t^2|S_{t-1}}\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^1(\cdot|S_{t-1})}\mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})-\mathbb{E}_{S_{t-1},x_t}\mathbb{E}_{\boldsymbol{\theta}_t^1,a_t^2|S_{t-1}}\mathrm{FG}_t^1(\boldsymbol{\theta}_t^1)\\
&=\frac{9e^{2B}\mu d}{2\eta}+\mathbb{E}_{S_{t-1},x_t,a_t^1,a_t^2}\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^1(\cdot|S_{t-1})}\left[\frac{e^{-2B}\eta}{18\mu}\mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})-\mathrm{FG}_t^1(\widetilde{\boldsymbol{\theta}})\right]\\
&\le\left(\frac{9de^{2B}}{2\eta}+32B^2\right)\mu+\frac{Z_{t-1}^1-Z_t^1}{\mu},
\end{aligned}\tag{B.2}$$

where the first equality holds due to the law of total expectation, the first inequality holds due to the decoupling lemma (Lemma D.1), the second equality holds because $\boldsymbol{\theta}_t^0$ and a_t^1 are independent conditioned on S_{t-1}, x_t, and the last inequality holds due to Lemma B.1. For I_2, we can similarly prove that

$$\begin{aligned}
&\mathbb{E}\left[\langle\boldsymbol{\theta}_t^2-\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^2)-\boldsymbol{\phi}(x_t,a_t^1)\rangle-\langle\boldsymbol{\theta}_t^2,\boldsymbol{\phi}(x_t,a_t^2)-\boldsymbol{\phi}(x_t,a_t^1)\rangle+\langle\boldsymbol{\theta}_*,\boldsymbol{\phi}(x_t,a_t^*)-\boldsymbol{\phi}(x_t,a_t^1)\rangle\right]\\
&\le\left(\frac{9de^{2B}}{2\eta}+32B^2\right)\mu+\frac{Z_{t-1}^2-Z_t^2}{\mu}
\end{aligned}\tag{B.3}$$

Plugging (B.2) and (B.3) into (B.1), taking the sum over t, we have

$$\begin{aligned}
\sum_{t=1}^T\mathbb{E}\left[r_*(x_t,a_t^*)-\frac{r_*(x_t,a_t^1)+r_*(x_t,a_t^2)}{2}\right]&\le\left(\frac{9de^{2B}}{2\eta}+32B^2\right)\mu T+\frac{(Z_0^1+Z_0^2)-(Z_T^1+Z_T^2)}{2\mu}\\
&=\left(\frac{9de^{2B}}{2\eta}+32B^2\right)\mu T-\frac{Z_T^1+Z_T^2}{2\mu},
\end{aligned}$$

where the equality holds because $Z_0^j=0$. □

B.2 Proof of Lemma A.3

Proof of Lemma A.3. We first can decompose ΔL^j into its expectation I_1 and a deviation term I_2:

$$\begin{aligned}
\Delta L^j(\boldsymbol{\theta},x_t,a_t^1,a_t^2,y_t)&=\underbrace{\mathbb{E}_{y_t|x_t,a_t^1,a_t^2}\Delta L^j(\boldsymbol{\theta},x_t,a_t^1,a_t^2,y_t)}_{I_1}\\
&\quad+\underbrace{\Delta L^j(\boldsymbol{\theta},x_t,a_t^1,a_t^2,y_t)-\mathbb{E}_{y_t|x_t,a_t^1,a_t^2}\Delta L^j(\boldsymbol{\theta},x_t,a_t^1,a_t^2,y_t)}_{I_2}.
\end{aligned}\tag{B.4}$$

For I_1, note that

$$\frac{1}{1+e^z} = -\sigma'(z), \quad \frac{1}{1+e^{-z}} = 1 - \frac{1}{1+e^z} = 1 + \sigma'(z),$$

so we have

$$\begin{aligned}
&I_1 - \mu \mathrm{FG}_t^j(\boldsymbol{\theta}) \\
&= [1 + \sigma'(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)] \\
&\quad \cdot \eta[\sigma(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) - \sigma(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)] \\
&\quad - \sigma'(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) \\
&\quad \cdot \eta[\sigma(-\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) - \sigma(-\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)] \\
&= \eta[\sigma(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a^1) - \boldsymbol{\phi}(x, a^2)\rangle) - \sigma(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x, a^1) - \boldsymbol{\phi}(x, a^2)\rangle)] \\
&\quad - \eta\sigma'(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) \cdot \langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle, \qquad \text{(B.5)}
\end{aligned}$$

where the second equality holds because $\sigma(z) - \sigma(-z) = -z$. By Taylor expansion, there exists ξ between $\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle$ and $\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle$ such that

$$\begin{aligned}
&\sigma(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a^1) - \boldsymbol{\phi}(x, a^2)\rangle) - \sigma(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x, a^1) - \boldsymbol{\phi}(x, a^2)\rangle) \\
&\qquad - \sigma'(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) \cdot \langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle \\
&= \frac{\sigma''(\xi)}{2} \langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle^2 \\
&\leq 2\sigma''(\xi) \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 \\
&\leq \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|^2, \qquad \text{(B.6)}
\end{aligned}$$

where the first inequality holds because $\langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle^2 \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 \cdot \|\boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\|_2^2 \leq 4\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|^2$, and the second inequality holds because $\sigma''(\xi) = 1/(2 + e^{\xi} + e^{-\xi}) \leq 1/4$. Furthermore, for $\mathrm{FG}_t^j(\boldsymbol{\theta})$, let $\widehat{a} = \mathrm{argmax}_{a \in \mathcal{A}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a)\rangle$, then

$$\begin{aligned}
\mathrm{FG}_t^j(\boldsymbol{\theta}) &= \max_{a \in \mathcal{A}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a)\rangle - \langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^*)\rangle + \langle \boldsymbol{\theta}_* - \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a_t^{3-j})\rangle \\
&\leq \langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, \widehat{a}) - \boldsymbol{\phi}(x_t, a_t^{3-j})\rangle \\
&\leq 2\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2, \qquad \text{(B.7)}
\end{aligned}$$

where the first inequality holds because $\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^*)\rangle \geq \langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, \widehat{a})\rangle$, and the second inequality holds because $\langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, \widehat{a}) - \boldsymbol{\phi}(x_t, a_t^{3-j})\rangle \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \cdot \|\boldsymbol{\phi}(x_t, \widehat{a}) - \boldsymbol{\phi}(x_t, a_t^{3-j})\|_2 \leq 2\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2$. Plugging (B.6) and (B.7) into (B.5), we have

$$I_1 \leq \frac{\eta}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 + 2\mu \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2. \qquad \text{(B.8)}$$

Finally, for I_2, note that for $y \in \{\pm 1\}$, we have

$$\sigma(yp) - \sigma(yq) = \sigma(p) - \sigma(q) + \frac{y-1}{2}(p - q),$$

so we have

$$\begin{aligned}
I_2 &= \eta \left[\frac{y_t - 1}{2} - \sigma'(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) \right] \langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle \\
&\leq \eta \left| \frac{y_t - 1}{2} - \sigma'(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) \right| \cdot 2\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2, \qquad \text{(B.9)}
\end{aligned}$$

where the inequality holds because $\langle \boldsymbol{\theta} - \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \cdot \|\boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\|_2 \leq 2\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2$. Denote

$$\epsilon_t \coloneqq \left| \frac{y_t - 1}{2} - \sigma'(\langle \boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) \right|,$$

then we have

$$\mathbb{E}_{y_t|x_t,a_t^1,a_t^2}\epsilon_t = \frac{2\exp(\langle\boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t,a_t^1)-\boldsymbol{\phi}(x_t,a_t^2)\rangle)}{[1+\exp(\langle\boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t,a_t^1)-\boldsymbol{\phi}(x_t,a_t^2)\rangle)]^2} \le \frac{1}{2}, \tag{B.10}$$

where the inequality holds due to AM-GM inequality. Denote

$$\widehat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\le\delta} p_0(\boldsymbol{\theta}),$$

then we have

$$\begin{aligned}
\log\int_{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2\le\delta} p_0(\widetilde{\boldsymbol{\theta}})\,\mathrm{d}\widetilde{\boldsymbol{\theta}} &\ge \log\int_{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_\infty\le\delta/\sqrt{d}} p_0(\widetilde{\boldsymbol{\theta}})\,\mathrm{d}\boldsymbol{\theta} \\
&\ge \log\Big(p_0(\widehat{\boldsymbol{\theta}})(2\delta/\sqrt{d})^d\Big) \\
&\ge \log p_0(\boldsymbol{\theta}_*) - L\delta + d\log(2\delta/\sqrt{d}),
\end{aligned} \tag{B.11}$$

where the first inequality holds because $\{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_\infty\le\delta/\sqrt{d}\}\subset\{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\le\delta\}$, the second inequality holds due to the definition of $\widehat{\boldsymbol{\theta}}$, and the last inequality holds because $\log p_0$ is L-Lipschitz. Therefore, we have

$$\begin{aligned}
Z_T^j &= \mathbb{E}_{S_T}\log\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p_0}\exp\bigg(-\sum_{t=1}^T \Delta L^j(\widetilde{\boldsymbol{\theta}},x_t,a_t^1,a_t^2,y_t)\bigg) \\
&\ge \mathbb{E}_{S_T}\log\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p_0}\exp\bigg(-\frac{\eta T}{2}\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2^2 - 2\mu T\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2 - 2\eta\sum_{t=1}^T\epsilon_t\cdot\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2\bigg) \\
&\ge \mathbb{E}_{S_T}\log\int_{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2\le\delta} p_0(\widetilde{\boldsymbol{\theta}})\exp\bigg(-\frac{\eta T\delta^2}{2} - 2\mu T\delta - 2\eta\delta\sum_{t=1}^T\epsilon_t\bigg)\,\mathrm{d}\widetilde{\boldsymbol{\theta}} \\
&= -\frac{\eta T\delta^2}{2} - 2\mu T\delta - 2\eta\delta\sum_{t=1}^T\mathbb{E}\epsilon_t + \log\int_{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2\le\delta} p_0(\widetilde{\boldsymbol{\theta}})\,\mathrm{d}\widetilde{\boldsymbol{\theta}} \\
&\ge -2(\mu+\eta)T\delta + \log p_0(\boldsymbol{\theta}_*) - L\delta + d\log(2\delta/\sqrt{d}),
\end{aligned}$$

where the first inequality holds by plugging (B.8) and (B.9) into (B.4), the second inequality holds by restricting the domain to $\{\widetilde{\boldsymbol{\theta}}:\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2\le\delta\}$, and the last inequality holds due to (B.10) and (B.11) and because $\delta/2\le 1$. Take $\delta=\min\{d/(L+2(\mu+\eta)T), B, 2\}$, then

$$Z_T^j \ge \log p_0(\boldsymbol{\theta}_*) - d + d\log\frac{\sqrt{d}}{L+2(\mu+\eta)T}.$$

□

B.3 Proof of Lemma A.5

Proof of Lemma A.5. We first can decompose ΔL^j into its expectation I_1 and a deviation term I_2:

$$\Delta L^j(\theta,x_t,a_t^1,a_t^2,y_t) = \underbrace{\mathbb{E}_{y_t|x_t,a_t^1,a_t^2}\Delta L^j(\theta,x_t,a_t^1,a_t^2,y_t)}_{I_1} \tag{B.12}$$

$$+\underbrace{\Delta L^j(\theta,x_t,a_t^1,a_t^2,y_t) - \mathbb{E}_{y_t|x_t,a_t^1,a_t^2}\Delta L^j(\theta,x_t,a_t^1,a_t^2,y_t)}_{I_2}. \tag{B.13}$$

For I_1, similar to the proof of Lemma A.3, we have

$$I_1 - \mu\mathrm{FG}_t^j(\theta) \le \frac{\eta}{8}(\Delta r_\theta(x_t,a_t^1,a_t^2) - \Delta r_{\theta_*}(x_t,a_t^1,a_t^2))^2 \le \frac{L_0^2\eta}{2}d(\theta,\theta_*)^2, \tag{B.14}$$

where the second inequality holds because Δr_θ is $2L_0$-Lipschitz. For $\mathrm{FG}_j^t(\boldsymbol{\theta})$, let $\widehat{a} = \mathrm{argmax}_{a\in\mathcal{A}} r_\theta(x_t, a)$, then

$$\mathrm{FG}_t^j(\theta) = \Delta r_\theta(x_t, \widehat{a}, a_t^{3-j}) - \Delta r_{\theta_*}(x_t, a_t^*, a_t^{3-j}) \le \Delta r_\theta(x_t, \widehat{a}, a_t^{3-j}) - \Delta r_{\theta_*}(x_t, \widehat{a}, a_t^{3-j}) \le 2L_0 d(\theta, \theta_*), \tag{B.15}$$

where the first inequality holds due to optimality of a_t^*, and the second inequality holds because Δr_θ is $2L_0$-Lipschitz. For I_2, similar to the proof of Lemma A.3, we have

$$|I_2| \le \frac{\eta}{2}|\Delta r_\theta(x_t, a_t^1, a_t^2) - \Delta r_{\theta_*}(x_t, a_t^1, a_t^2)| \le L_0\eta d(\theta, \theta_*), \tag{B.16}$$

where the second inequality holds because Δr_θ is $2L_0$-Lipschitz. Denote

$$\widehat{\theta} = \min_{\theta: d(\theta,\theta_*)\le\delta} p_0(\theta),$$

then we have

$$\log \int_{d(\widetilde{\theta},\theta_*)\le\delta} p_0(\widetilde{\theta})\,\mathrm{d}\theta \ge \log\left(p_0(\widehat{\theta})\int_{d(\widetilde{\theta},\theta_*)\le\delta}\mathrm{d}\widetilde{\theta}\right) \ge \log p_0(\theta_*) - L\delta + \log\bar{\mu}(\{\theta\in\Theta : d(\theta,\theta_*)\le\delta\}), \tag{B.17}$$

where the first inequality holds due to the definition of $\widehat{\mu}$, and the second inequality holds because $\log p_0$ is L-Lipschitz. We thus have

$$\begin{aligned}
Z_T^j &= \mathbb{E}_{S_T}\log\mathbb{E}_{\widetilde{\theta}\sim p_0}\exp\left(-\sum_{t=1}^T \Delta L^j(\widetilde{\theta}, x_t, a_t^1, a_t^2, y_t)\right)\\
&\ge \mathbb{E}_{S_T}\log\mathbb{E}_{\widetilde{\theta}\sim p_0}\exp\left(-\frac{L_0^2T\eta}{2}d(\widetilde{\theta},\theta_*)^2 - L_0T(\eta+2\mu)d(\widetilde{\theta},\theta_*)\right)\\
&\ge \log\int_{\widetilde{\theta}:d(\widetilde{\theta},\theta_*)\le\delta} p_0(\widetilde{\theta}) - \frac{L_0^2T\eta\delta^2}{2} - L_0T(\eta+2\mu)\delta\\
&\ge \log p_0(\theta_*) - L\delta + \log\bar{\mu}(\{\theta\in\Theta : d(\theta,\theta_*)\le\delta\}) - \frac{L_0^2T\eta\delta^2}{2} - L_0T(\eta+2\mu)\delta,
\end{aligned}$$

where the first inequality holds due to (B.14), (B.15) and (B.16), the second inequality holds because $\{\theta\in\Theta : d(\theta,\theta_*)\le\delta\}\subset\Theta$, and the last inequality holds due to (B.17). Taking $\delta = 2/(L_0T)$, we have

$$\begin{aligned}
Z_T^j &\ge \log p_0(\theta_*) + \log\bar{\mu}(\{\theta\in\Theta : d(\theta,\theta_*)\le 2/(L_0T)\}) - \frac{2L}{L_0T} - \frac{2\eta}{T} - 2(\eta+2\mu)\\
&\ge \log p_0(\theta_*) + \log\bar{\mu}(\{\theta\in\Theta : d(\theta,\theta_*)\le 2/(L_0T)\}) - \frac{2L}{L_0T} - 4(\eta+\mu),
\end{aligned}$$

where the second inequality holds because $T\ge 1$. □

C Proof of Lemma B.1

Proof of Lemma B.1. The difference of the potential between steps can be bounded as

$$\begin{aligned}
Z_t^j - Z_{t-1}^j &= \mathbb{E}_{S_t}\log\frac{\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p_0} W_t^j(\widetilde{\boldsymbol{\theta}}|S_t)}{\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p_0} W_{t-1}^j(\widetilde{\boldsymbol{\theta}}|S_{t-1})}\\
&= \mathbb{E}_{S_t}\log\frac{\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p_0}[W_{t-1}^j(\boldsymbol{\theta}|S_{t-1})\exp(-\Delta L^j(\widetilde{\boldsymbol{\theta}}, x_t, a_t^1, a_t^2, y_t))]}{\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p_0} W_{t-1}^j(\widetilde{\boldsymbol{\theta}}|S_{t-1})}\\
&= \mathbb{E}_{S_t}\log\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^j(\cdot|S_{t-1})}\exp(-\Delta L^j(\widetilde{\boldsymbol{\theta}}, x_t, a_t^1, a_t^2, y_t))\\
&\le \mathbb{E}_{S_{t-1},x_t,a_t^1,a_t^2}\log\mathbb{E}_{\widetilde{\boldsymbol{\theta}}\sim p^j(\cdot|S_{t-1})}\mathbb{E}_{y_t|x_t,a_t^1,a_t^2}\exp(-\Delta L^j(\widetilde{\boldsymbol{\theta}}, x_t, a_t^1, a_t^2)),
\end{aligned} \tag{C.1}$$

where the first equality holds due to the definition of Z_t in (7.4), the second equality holds due to the definition of definition of W_t in (7.5), the third equality holds due to (7.7), and the inequality holds due to Jensen's inequality. Note that

$$\mathbb{E}_{y_t|x_t,a_t^1,a_t^2}\exp(-\Delta L(\widetilde{\boldsymbol{\theta}}, x_t, a_t^1, a_t^2)) = \exp(\mu\mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}))$$

$$
\begin{aligned}
&\cdot \bigg[\bigg(\frac{1+\exp(-\langle\boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)}{1+\exp(-\langle\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)}\bigg)^{\eta} \cdot \frac{1}{1+\exp(-\langle\boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)} \\
&+ \bigg(\frac{1+\exp(\langle\boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)}{1+\exp(\langle\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)}\bigg)^{\eta} \cdot \frac{1}{1+\exp(\langle\boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)}\bigg] \\
&= \exp(\mu \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}) + \sigma(\langle(1-\eta)\boldsymbol{\theta}_* + \eta\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) \\
&\quad - (1-\eta)\sigma(\langle\boldsymbol{\theta}_*, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle) - \eta\sigma(\langle\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle)) \\
&\leq \exp(\mu \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}) - e^{-2B}/8 \cdot \eta(1-\eta)\mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})) \\
&\leq \exp(\mu \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}) - \eta e^{-2B}/16 \cdot \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})),
\end{aligned} \tag{C.2}
$$

where the second equality holds due to the definition of ΔL^j in (7.6), the first inequality holds due to Lemma D.3, and the second inequality holds because $\eta \leq 1/2$. Plugging (C.2) into (C.1), we have

$$
\begin{aligned}
Z_t^j - Z_{t-1}^j &\leq \mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \log \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} \exp(\mu \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}) - \eta e^{-2B}/16 \cdot \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})) \\
&\leq \frac{1}{2} \underbrace{\mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \log \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} \exp(2\mu \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}))}_{I_1} \\
&\quad + \frac{1}{2} \underbrace{\mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \log \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} \exp(-\eta e^{-2B}/8 \cdot \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}}))}_{I_2},
\end{aligned} \tag{C.3}
$$

where the second inequality holds due to Cauchy-Schwarz inequality. For I_1, note that $\mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}) \in [-4B, 4B]$, so by Lemma D.2, we have

$$
I_1 \leq 2\mu \mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}) + 64\mu^2 B^2. \tag{C.4}
$$

For I_2, we have

$$
\begin{aligned}
I_2 &\leq \mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \log \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} [1 - \eta e^{-2B}/8 \cdot \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}}) + \eta^2 e^{-4B}/128 \cdot (\mathrm{LS}_t(\widetilde{\boldsymbol{\theta}}))^2] \\
&\leq \mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \log \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} [1 - \eta e^{-2B}/8 \cdot \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})(1 - \eta e^{-2B}/16 \cdot (4B)^2)] \\
&\leq \mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \log \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} [1 - \eta e^{-2B}/8 \cdot \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})(1 - 1/2 \cdot 1/e^2)] \\
&\leq \mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \log \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} [1 - \eta e^{-2B}/9 \cdot \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})] \\
&\leq -\eta e^{-2B}/9 \cdot \mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}}),
\end{aligned} \tag{C.5}
$$

where the first inequality holds because $e^z \leq 1 + z + z^2/2$ for all $z \leq 0$, the second inequality holds because $\mathrm{LS}_t(\boldsymbol{\theta}) \leq (4B)^2$, the second inequality holds because $Be^{-B} \leq 1/e$ for all $B > 0$, the fourth inequality holds because $1/8 \cdot (1 - 1/2 \cdot 1/e^2) \geq 1/9$, and the last inequality holds because $\log(1+z) \leq z$. Plugging (C.4) and (C.5) into (C.3), we have

$$
Z_t^j - Z_{t-1}^j \leq 32\mu^2 B^2 + \mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} [\mu \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}}) - \eta e^{-2B}/18 \cdot \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}})].
$$

Rearranging terms, we obtain

$$
\mathbb{E}_{S_{t-1}, x_t, a_t^1, a_t^2} \mathbb{E}_{\widetilde{\boldsymbol{\theta}} \sim p^j(\cdot|S_{t-1})} \bigg[\frac{e^{-2B}\eta}{18\mu} \mathrm{LS}_t(\widetilde{\boldsymbol{\theta}}) - \mathrm{FG}_t^j(\widetilde{\boldsymbol{\theta}})\bigg] \leq \mu^{-1}(Z_{t-1}^j - Z_t^j) + 32\mu B^2.
$$

□

D Auxiliary Lemmas

Lemma D.1 (Decoupling lemma, Lemma D.1 in Fan & Gu (2023)). *Let P be a joint distribution over two $\mathbb{R}^d$ spaces. For any constant $\lambda > 0$, we have*

$$
\mathbb{E}_{(\boldsymbol{\theta}, \boldsymbol{\phi}) \sim P} \langle \boldsymbol{\theta}, \boldsymbol{\phi} \rangle \leq d\lambda + \frac{1}{4\lambda} \mathbb{E}_{(\boldsymbol{\theta}, \boldsymbol{\phi}) \sim P} \mathbb{E}_{(\boldsymbol{\theta}', \boldsymbol{\phi}') \sim P} \langle \boldsymbol{\theta}', \boldsymbol{\phi} \rangle^2.
$$

Lemma D.2 (Hoeffding's lemma)**.** *Let X be a random variable that is bounded by $a \le X \le b$. Then for any constant λ,*

$$\mathbb{E}\exp(\lambda X) \le \exp(\lambda \mathbb{E}X + \lambda^2(b-a)^2/8).$$

Lemma D.3. *Let $\sigma(z) = \log(1+\exp(-z))$. Then for any $\eta \in (0,1)$ and $p, q \in [-2B, 2B]$, we have*

$$\sigma((1-\eta)p + \eta q) - (1-\eta)\sigma(p) - \eta\sigma(q) \le -\frac{e^{-2B}}{8}\eta(1-\eta)(q-p)^2.$$

Proof. Without loss of generality, we assume that $p \le q$. Otherwise, we substitute $(p, q, \eta) \leftarrow (q, p, 1-\eta)$. By Lagrange's mean value theorem, there exists $\xi(\eta, p, q) \in [p, q]$ such that

$$\eta\sigma'((1-\eta)p + \eta q) - \eta\sigma'(q) = -\eta(1-\eta)(q-p)\sigma''(\xi(\eta, p, q)).$$

Note that for any $\xi \in [-2B, 2B]$, the second derivative $\sigma''(\xi)$ is bounded by

$$\sigma''(\xi) = \frac{1}{e^{\xi} + 2 + e^{-\xi}} \ge \frac{e^{-2B}}{4},$$

so

$$\eta\sigma'((1-\eta)p + \eta q) - \eta\sigma'(q) \le -\frac{e^{-2B}\eta(1-\eta)}{4}(q-p).$$

Taking integral w.r.t. q on both sides, we have

$$\sigma((1-\eta)p + \eta q) - (1-\eta)\sigma(p) - \eta\sigma(q) \le -\frac{e^{-2B}}{8}\eta(1-\eta)(q-p)^2.$$

□

Lemma D.4. *For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, we have*

$$D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\theta}} || \mathbb{P}_{\boldsymbol{\theta}'}) \le \frac{1}{8}\mathbb{E}_{\boldsymbol{\theta}}\bigg[\sum_{t=1}^{T}\langle \boldsymbol{\theta} - \boldsymbol{\theta}', \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle^2\bigg].$$

Proof. We define the shorthand notation

$$p_{\boldsymbol{\theta},t} \coloneqq \mathbb{P}_{\boldsymbol{\theta}}[y_t = 1 | x_t, a_t^1, a_t^2] = \frac{1}{1 + \exp(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a_t^2) - \boldsymbol{\phi}(x_t, a_t^1)\rangle)},$$

then by decomposition properties of the relative entropy, we have

$$D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\theta}} || \mathbb{P}_{\boldsymbol{\theta}'}) = \mathbb{E}_{\boldsymbol{\theta}}\bigg[\sum_{t=1}^{T} D_{\mathrm{KL}}(\mathrm{Ber}(p_{\boldsymbol{\theta},t}) || \,\mathrm{Ber}(p_{\boldsymbol{\theta}',t}))\bigg]. \tag{D.1}$$

Denote

$$u = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle,$$
$$v = \langle \boldsymbol{\theta}', \boldsymbol{\phi}(x_t, a_t^1) - \boldsymbol{\phi}(x_t, a_t^2)\rangle,$$

then

$$\begin{aligned} D_{\mathrm{KL}}(\mathrm{Ber}(p_{\boldsymbol{\theta},t}) || \,\mathrm{Ber}(p_{\boldsymbol{\theta}',t})) &= \frac{1}{1+e^{-u}} \cdot \log\frac{1+e^{-v}}{1+e^{-u}} + \frac{1}{1+e^{u}} \cdot \log\frac{1+e^{v}}{1+e^{u}} \\ &= \log(1+e^{-v}) - \log(1+e^{-u}) - \frac{-1}{1+e^{u}}(v-u) \\ &= \frac{e^{\xi}}{(1+e^{\xi})^2} \cdot \frac{(v-u)^2}{2} \\ &\le \frac{(v-u)^2}{8}, \end{aligned} \tag{D.2}$$

where the first equality holds due to definition of $p_{\boldsymbol{\theta},t}$, the third equality holds due to Taylor expansion with Langragian remainder, and the inequality holds due to AM-GM inequality. Plugging (D.2) into (D.1), we derive the desired upper bound for $D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\theta}} || \mathbb{P}_{\boldsymbol{\theta}'})$. □