Stochastic Contextual Dueling Bandits under Linear Stochastic Transitivity Models

Viktor Bengs [1] Aadirupa Saha [2] Eyke Hüllermeier [1]

Abstract

We consider the regret minimization task in a dueling bandits problem with context information. In every round of the sequential decision problem, the learner makes a context-dependent selection of two choice alternatives (arms) to be compared with each other and receives feedback in the form of noisy preference information. We assume that the feedback process is determined by a linear stochastic transitivity model with contextualized utilities (CoLST), and the learner's task is to include the best arm (with highest latent context-dependent utility) in the duel. We propose a computationally efficient algorithm, CoLSTIM, which makes its choice based on imitating the feedback process using perturbed context-dependent utility estimates of the underlying CoLST model. If each arm is associated with a d-dimensional feature vector, we show that CoLSTIM achieves a regret of order $\tilde{O}(\sqrt{dT})$ after T learning rounds. Additionally, we also establish the optimality of CoLSTIM by showing a lower bound for the weak regret that refines the existing average regret analysis. Our experiments demonstrate its superiority over state-of-art algorithms for special cases of CoLST models.

1. Introduction

The *multi-armed bandit* (MAB) problem represents a class of online machine learning problems in which a learner can perform different actions (metaphorically referred to as "pulling arms") resulting in numerical rewards within a sequential decision process (Lattimore & Szepesvári, 2020). From a learning perspective, the main challenge of this problem is the *exploration-exploitation dilemma*. The learner's lack of knowledge about the underlying reward mechanism associated with the actions forces it to try different actions often enough, so as to gradually understand these mechanisms better in the course of time (exploration). On the other side, the learner is inclined to choose actions deemed rewarding, as as to maximize the rewards cumulated over time (exploitation).

Although this learning scenario has been used in many application areas, such as web advertising or medical treatments, it has been generalized in various ways to take specific aspects arising in practical problems into account. In many applications, for example, additional information is available about the context in which an action is performed, for instance, the user profile in web advertising or a patient's medical record. In the *contextual MAB setting* (Chu et al., 2011; Valko et al., 2013), an action's reward mechanism might depend on the contextual information, so that one or the other action is optimal depending on the context.

Another practically relevant extension of the classical MAB problem is the *dueling* (Yue & Joachims, 2009) or *preference-based* (Bengs et al., 2021) bandit problem, where the learner's action is to select a pair of arms resulting in a noisy qualitative comparison between these arms (which arm generated the highest reward?) rather than selecting a single arm resulting in a numerical reward (what is the generated reward?). An example is A/B testing, where users are presented with two options from which the more preferred one should be selected, but no (meaningful) numerical reward can be measured for that choice. Although the dueling bandits problem itself — just like the standard MAB problem — has already been the starting point for various generalizations, little attention has been paid to a contextual learning scenario in the same spirit as the contextual bandits. One reason might be that, in contrast to numerical reward learning scenarios, in preference-based learning scenarios it is often far from obvious to specify basic characteristics such as the optimal (pair of) arm(s) or a practically meaningful model assumption for the feedback mechanism.

One way to get around these issues, which we propose in this paper, is to use the so-called *linear stochastic transi-*

[1] Institute of Informatics, LMU Munich, Germany [2] Microsoft Research, New York City, US. Correspondence to: Viktor Bengs <Viktor.bengs@lmu.de>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

tivity (LST) models[1], which enjoy great popularity in the fields of economics, behavioral science (Cattelan, 2012) with the Bradley-Terry-Luce and the Thurstone-Mosteller model being their most well-known instantiations. They are especially famous for their use in rating or ranking systems like the Elo system (Elo, 1978) or TrueSkill (Herbrich et al., 2006). These LST models describe the mechanism leading to the observation of the outcome of a comparison between two objects (choice alternatives, players, etc.) in a probabilistic way by assuming latent utility values of the objects and a noisy perception of these utility values. The probability of the outcome of the pairwise comparison, say A over B, is then equal to the probability that A's noisy utility value is greater than B's.

Our Contributions

(1). Contextual dueling bandits under linear stochastic transitivity models: We tackle the contextualized dueling bandits problem and extend the class of possible contextualized feedback mechanisms beyond the contextual Plackett-Luce model by allowing it to be determined by an LST model in which the utility values of the arms (choice alternatives) depend linearly on the context.

(2). A computationally efficient and near-optimal algorithm: We construct a learning algorithm, CoLSTIM, which for the choice of its "first arm" in the duel imitates the selection mechanism of an LST model, while the choice of the second arm is tailored towards the task of regret minimization. CoLSTIM enjoys under mild assumptions a bound on its average regret of order $O(\sqrt{dT}\log(T))$ matching the minimax lower bound up to logarithmic terms. From a practical point of view, CoLSTIM is computationally more efficient than current algorithms for more restrictive assumptions on the feedback mechanism and performs quite well in numerical simulations in comparison.

(3). Stronger lower bound: We show a lower bound for weak regret — a weaker form of the usual average regret — which has the same order as the stronger average regret. This result shows that it is not possible to derive stronger regret guarantees for minimizing weak regret in the contextual dueling bandits problem.

Outline. The paper is organized as follows. In Section 2, we recall the linear stochastic transitivity model along with all important definitions and theoretical notions. We give a formal description of the contextualized dueling problem under LST in Section 3 and propose with CoLSTIM a learning algorithm for tackling this problem, which we also analyze theoretically in the form of an upper bound on its expected average regret. In an experimental study we show its empirical superiority over state-of-art algorithms for special cases of LST models in Section 4. Finally, we discuss the works closest related to ours in Section 5, prior to concluding the paper with an outlook on future work in Section 6. All proofs are given in the supplementary materials.

Notation. For $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, \ldots, n\}$ and by $1_{[\cdot]}$ the indicator function. We write $\|\boldsymbol{x}\|_A = \sqrt{\boldsymbol{x}^\top A \boldsymbol{x}}$ for any $\boldsymbol{x} \in \mathbb{R}^d$ and any positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, while $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_{\boldsymbol{I}_d}$ for $\boldsymbol{I}_d$ being the $d \times d$ identity matrix. For symmetric matrices A, B, we write $A \leq B$ if $B - A$ is positive semi-definite. Let $\mathcal{B}_r(p)$ be the ℓ_p ball of radius $r \geq 0$, i.e., $\mathcal{B}_r(p) = \{\mathbf{x} \mid \|\mathbf{x}\|_p \leq r\}$, where $\|\cdot\|_p$ denotes the standard ℓ_p-norm.

2. Linear Stochastic Transitivity

The linear stochastic transitivity (LST) model is a class of parameterized probability models describing the outcome of a comparison between two choice alternatives from a set of n choice alternatives. An LST model has two parameters: An n-dimensional parameter $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$, where each component represents the (latent) *utility* of a choice alternative, and another functional parameter in the form of a symmetric cumulative distribution function[2] $F : \mathbb{R} \to [0, 1]$ called *comparison function.* According to an LST model, the probability that alternative i is preferred over alternative j, denoted by $i \succ j$, is

$$\mathbb{P}(1_{[i\succ j]} = 1) = F(u_i - u_j). \tag{1}$$

Sorting Perturbed Utilities. One equivalent way to obtain the probability in (1) is as follows. Suppose ϵ_i, ϵ_j are two iid random variables with distribution G and such that $\epsilon_j - \epsilon_i \sim F$. Now let v_i be the *perturbed utility* of i by adding the noise term ϵ_i to u_i. If F is continuous, then the probability that i is the choice alternative with the higher perturbed utility is exactly as in (1):

$$\mathbb{P}\Big(i = \operatorname*{argmax}_{k=i,j} v_k\Big) = \mathbb{P}(\epsilon_j - \epsilon_i \leq u_i - u_j) = F(u_i - u_j).$$

This process of sorting the perturbed utilities is a fairly intuitive way to model how a decision about preferences between two options has come about by a decision maker such as a human user or the environment. It will also be exactly this process based on the *perturbation distribution* G that will lead to the key algorithmic idea of our approach.

Examples. Specific choices of the perturbation distribution G have gained much popularity, as the corresponding

[1] Some authors refer to this model class also as Thurstone preference models. e.g., Jin et al. (2020).

[2] A cumulative distribution function is symmetric if the corresponding probability distribution is symmetric around a specific value, usually the mean.

comparison function F has a known form:

- *Bradley-Terry-Luce (BTL) model* — Setting G to be the standard Gumbel distribution, it is well known that the difference of two iid standard Gumbel distributed random variables is standard logistic distributed. In this case, we have $F(u_i - u_j) = \exp(u_i)/(\exp(u_i) + \exp(u_j))$.
- *Thurstone-Mosteller model* — The distribution of the difference of two iid standard Gaussian distributed random variables is Gaussian with zero mean and variance 2. Hence, if G is the standard Gaussian distribution, it holds that $F(u_i - u_j) = \Phi(\,{}^{(u_i - u_j)}/_{\sqrt{2}})$, where Φ is the cumulative distribution function of a standard Gaussian.
- *Exponential Noise* — If G is the exponential distribution with rate $\lambda > 0$, then F is the cumulative distribution function of a Laplace distribution with location 0 and scale λ^{-1}. Thus, if sgn denotes the sign function, then $F(u_i - u_j) = \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(u_i - u_j)(1 - \exp(-{}^{|u_i - u_j|}/_{\lambda}))$.

Context Information. In order to take context information $\boldsymbol{x}_i \in \mathbb{R}^d$ about the i^{th} choice alternative into account, we follow the approach by Cheng et al. (2010); Schäfer & Hüllermeier (2018) for the BTL model[3] and replace the constant latent utility u_i by a linear function of the features, leading to a contextualized LST (CoLST) model. Formally, given n choice alternatives, which define a contextual decision problem, we summarize all context vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in a matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ and define the context-dependent utilities of the choice alternatives via

$$u_i(\mathbf{X}) = \boldsymbol{\theta}^\top \boldsymbol{x}_i = \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle, \quad \forall i = 1, \ldots, n, \qquad (2)$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is some weight vector. With this, the probability that alternative i is preferred over j given the context $\mathbf{X}$, according to a CoLST model, is

$$\mathbb{P}(1_{[i \succ j]} = 1 \,|\, \mathbf{X}) = F(u_i(\mathbf{X}) - u_j(\mathbf{X})). \qquad (3)$$

The log-likelihood function $\boldsymbol{l}\left(\cdot \,|\, 1_{[i \succ j]}, \{i,j\}, \mathbf{X}\right)$ (in short just $\boldsymbol{l}(\cdot)$) for an observation $(1_{[i \succ j]}, \{i,j\}, \mathbf{X})$ is

$$\begin{aligned} \boldsymbol{l}(\boldsymbol{\theta}) = {} & 1_{[i \succ j]} \log\left(F(\boldsymbol{\theta}^\top(\boldsymbol{x}_i - \boldsymbol{x}_j))\right) \\ & + (1 - 1_{[i \succ j]}) \log\left(F(\boldsymbol{\theta}^\top(\boldsymbol{x}_j - \boldsymbol{x}_i))\right). \end{aligned} \qquad (4)$$

The log-likelihood function can be expressed in a more convenient way, if F is of an exponential family type, i.e., if for $Y = 1_{[i \succ j]}$ we can write $\mathbb{P}(Y \,|\, \mathbf{X})$ as

$$\exp\left(\frac{Y\langle \boldsymbol{\theta}, \boldsymbol{z}_{i,j}\rangle - \tilde{F}(\langle \boldsymbol{\theta}, \boldsymbol{z}_{i,j}\rangle)}{v(\eta)} + c(Y, \eta)\right), \qquad (5)$$

where $\boldsymbol{z}_{i,j} = \boldsymbol{x}_i - \boldsymbol{x}_j$, $\tilde{F}$ is the anti-derivative of F, η is some scale parameter and v as well as c are normalization functions. The comparison functions F in the examples above all admit such a representation.

3. Contextual Dueling Bandits

In this section, we first introduce the problem setting prior to the suggested learning algorithm for this setting.

3.1. Problem Setting

We consider a set of $n \in \mathbb{N}_{\geq 2}$ available choice alternatives that we refer to as *arms*, and simply denote them by their index: $\mathcal{A} = \{1, \ldots, n\}$. The learning problem proceeds in a time horizon T, where in each time step $t \in \{1, \ldots, T\}$, the learner observes a context $\mathbf{X}_t = (\boldsymbol{x}_{t,1} \ldots \boldsymbol{x}_{t,n})$ with $\boldsymbol{x}_{t,i} \in \mathcal{X}$ for any arm i, where $\mathcal{X} \subset \mathbb{R}^d$ is the context space, e.g., $\mathcal{X} = \mathcal{B}_1(2) = \{\boldsymbol{x} \mid \|\boldsymbol{x}\| \leq 1\}$ could be the ℓ_2-unit ball of radius 1. Each vector $\boldsymbol{x}_{t,i}$ encodes features of the context in which an arm must be chosen, but possibly also of the arm i itself. In other words, $\boldsymbol{x}_{t,i}$ contains properties of both the context (determined by the environment) and the arm, for instance obtained by a joint feature map. In what follows it will turn out to be convenient to consider the contrast vectors $\boldsymbol{z}_{t,i,j} = \boldsymbol{x}_{t,i} - \boldsymbol{x}_{t,j}$ and assume that we equivalently obtain in each time step the $d \times \binom{n}{2}$ dimensional contrast matrix

$$\mathbf{Z}_t = (\boldsymbol{z}_{t,1,2}\, \boldsymbol{z}_{t,1,3} \ldots \boldsymbol{z}_{t,1,n}\, \boldsymbol{z}_{t,2,3} \ldots \boldsymbol{z}_{t,n-1,n}).$$

After observing the context (or contrast) information $\mathbf{X}_t$ (or $\mathbf{Z}_t$), the learner selects a pair of arms $S_t = (i_t, j_t) \in [n]^2$, whereupon the learner obtains preference feedback, which is either $i_t \succ j_t$ or $j_t \prec i_t$, i.e., i_t is preferred over j_t or the opposite. We assume that the feedback is generated by means of a CoLST model with a *known* perturbation distribution G^* (and corresponding comparison function F^*) and an *unknown* weight parameter $\boldsymbol{\theta}^*$. Formally, let $Y_t = 1_{[i_t \succ j_t]}$ be the binary feedback, then for any timestep $t \in [T]$,

$$\begin{aligned} \mathbb{P}(Y_t = 1 \,|\, \mathbf{Z}_t) &= F^*\left(\langle \boldsymbol{z}_{t,i_t,j_t}, \boldsymbol{\theta}^*\rangle\right) \\ &= F^*(u^*_{i_t}(\mathbf{X}_t) - u^*_{j_t}(\mathbf{X}_t)), \end{aligned} \qquad (6)$$

where $u^*_i(\mathbf{X}_t) = \boldsymbol{x}_{t,i}^\top \boldsymbol{\theta}^*$.

The goal of the learner is to select, in each time step t, a pair of arms $S_t = (i_t, j_t)$ involving the arm which is best for the current context $\mathbf{X}_t$. In the realm of preference-based multi-armed bandits, the notion of a best arm can be defined in various ways (Bengs et al., 2021). In our setting, where we assume choices to be guided by a linear stochastic transitivity model (with fixed but unknown weight parameter $\boldsymbol{\theta}^*$), it is natural to define the best arm for a time step t by the arm with the highest (latent) utility (see (2)). More specifically, the best arm for the current time step t is

$$i^*(t) = \arg\max\nolimits_{i \in \mathcal{A}} u^*_i(\mathbf{X}_t). \qquad (7)$$

With this, one can leverage the two most prevalent notions of instantaneous regret a learner suffers for selecting a pair of

[3] Strictly speaking, these works consider a Plackett-Luce model, which is a generalization of the BTL model, see (Hunter, 2004).

arms $S_t = (i_t, j_t)$ at time t from the non-contextual dueling bandits, namely the average regret and the weak regret:

$$r_t^a(S_t, \mathbf{X}_t) = \frac{2u_{i^*(t)}^*(\mathbf{X}_t) - u_{i_t}^*(\mathbf{X}_t) - u_{j_t}^*(\mathbf{X}_t)}{2},$$
$$r_t^w(S_t, \mathbf{X}_t) = u_{i^*(t)}^*(\mathbf{X}_t) - \max_{k \in S_t} u_k^*(\mathbf{X}_t)\,.$$

Thus, the average and weak cumulative regret for selecting pairs $(S_t)_{t\in[T]}$ for context information $(\mathbf{X}_t)_{t\in[T]}$ during the time horizon T are, respectively,

$$\begin{aligned} R_T^a((S_t)_{t\in[T]}) &= \sum\nolimits_{t=1}^T r_t^a(S_t, \mathbf{X}_t), \\ R_T^w((S_t)_{t\in[T]}) &= \sum\nolimits_{t=1}^T r_t^w(S_t, \mathbf{X}_t). \end{aligned} \tag{8}$$

It holds that $R_T^w((S_t)_{t\in[T]}) \leq R_T^a((S_t)_{t\in[T]})$, so that a learner with theoretical guarantees for the average regret satisfies theoretical guarantees for the weak regret as well. However, being a weaker notion of regret, the weak regret may permit stronger theoretical guarantees than average regret. Saha (2021) has shown a lower bound of order $\Omega(\sqrt{dT})$ for the average regret, and the following result shows that this bound also holds for the weak regret, indicating that stronger theoretical guarantees cannot be obtained for weak regret learners.

Theorem 3.1. *For any learning algorithm $\mathcal{A}$ for the contextual dueling bandits problem under linear stochastic transitivity models in dimension d, there exists an instance of the problem characterized by a weight vector $\boldsymbol{\theta}^* \in \mathcal{B}_1(2)$, context space $\mathcal{X} \subseteq \mathcal{B}_1(\infty)$, and a comparison function $F : \mathbb{R} \to [0,1]$, such that the expected weak regret incurred by $\mathcal{A}$ in any $T > \max(d, 16)$ rounds is*

$$\mathbb{E}\big[R_T^w((S_t^{\mathcal{A}})_{t\in[T]})\big] = \Omega\big(d\sqrt{T}\big).$$

Further, if we restrict $\mathcal{X} \subseteq \mathcal{B}_1(2)$, then

$$\mathbb{E}\big[R_T^w((S_t^{\mathcal{A}})_{t\in[T]})\big] = \Omega\big(\sqrt{dT}\big).$$

It is worth noting that our proof of Theorem 3.1 does not use a reduction to contextual linear bandits as the proof by Saha (2021) for the average regret (see Theorem 10), but is rather based on a direct proof technique (see Section B in the supplementary material).

3.2. CoLST Imitator

At the core of the learning task is the estimation of the unknown weight parameter $\boldsymbol{\theta}^*$, which basically determines the underlying CoLST model of the feedback mechanism. Assuming that we have a compact parameter space Θ such that $\boldsymbol{\theta}^* \in \Theta$, the arguably most natural way to derive an estimate is to use the maximum likelihood estimate (MLE):

$$\hat{\boldsymbol{\theta}}_t \in \mathrm{argmax}_{\boldsymbol{\theta}\in\Theta} \sum\nolimits_{s=1}^{t-1} \boldsymbol{l}\left(\boldsymbol{\theta} \,|\, Y_s, \{i_s, j_s\}, \mathbf{X}_s\right). \tag{9}$$

If the log-likelihood function is of the form (5), then $\hat{\boldsymbol{\theta}}_t$ can be computed by solving

$$0 = \sum\nolimits_{s=1}^{t-1} \big(Y_s - F(\langle \boldsymbol{z}_{s,i_s,j_s}, \boldsymbol{\theta}\rangle)\big)\boldsymbol{z}_{s,i_s,j_s}. \tag{10}$$

Equipped with an estimate of the underlying weight parameter, the question now becomes how to choose an appropriate pair of arms to deal with the exploration-exploitation trade-off. Our approach is essentially to imitate the choice process of the underlying CoLST model using the current estimator and their confidence widths to generate perturbed, context-dependent utilities giving rise to our CoLST Imitator (CoLSTIM) algorithm (see Algorithm 1).

Algorithm 1 CoLSTIM

```
Input: Exploration length $\tau > 0$, perturbation distribution $G$ with induced comparison (or cumulative distribution) function $F$, threshold $C_{\mathrm{thresh}} > 0$, coupling probabilities $(p_t)_t$, confidence width constant $c_1 > 0$
Initialization: Randomly choose $\{i_t, j_t\} \subset [n]$ for $\tau$ many time steps
Set $\mathbf{M}_{\tau+1} = \sum_{s=1}^{\tau} \boldsymbol{z}_{s,i_s,j_s}\boldsymbol{z}_{s,i_s,j_s}^\top$
for $t = \tau+1, 2, \ldots, T$ do
    Observe context vectors $\mathbf{X}_t = (\boldsymbol{x}_{t,1} \ldots \boldsymbol{x}_{t,n})$
    Compute MLE $\hat{\boldsymbol{\theta}}_t$ via (9) (or (10))
    Sample $B_t \sim \mathrm{Ber}(p_t)$
    if $B_t = 1$ then
        Sample $\tilde{\epsilon}_{t,i} \sim G$ for each $i \in [n]$
    else
        Sample $\tilde{\epsilon} \sim G$ and set $\tilde{\epsilon}_{t,i} = \tilde{\epsilon}$ for each $i \in [n]$
    end if
    $\epsilon_{t,i} = \min(C_{\mathrm{thresh}}, \max(-C_{\mathrm{thresh}}, \tilde{\epsilon}_{t,i}))\ \forall i \in [n]$
    $i_t = \arg\max_{i\in[n]} \boldsymbol{x}_{t,i}^\top \hat{\boldsymbol{\theta}}_t + \epsilon_{t,i} \|\boldsymbol{x}_{t,i}\|_{\mathbf{M}_t^{-1}}$
    $j_t = \arg\max_{i\in[n]} \langle \boldsymbol{z}_{t,i,i_t}, \hat{\boldsymbol{\theta}}_t\rangle + c_1 \|\boldsymbol{z}_{t,i_t,i}\|_{\mathbf{M}_t^{-1}}$
    Choose $(i_t, j_t)$ and observe $Y_t = 1_{[i_t \succ j_t]}$
    Update $\mathbf{M}_{t+1} \leftarrow \mathbf{M}_t + \boldsymbol{z}_{t,i_t,j_t}\boldsymbol{z}_{t,i_t,j_t}^\top$
end for
```

More specifically, we first generate additive noise terms by sampling for each arm a random observation of the underlying perturbation distribution G and multiply it with the corresponding confidence width. Due to technical reasons we need to (a) threshold the generated perturbation variables from above (and below) by means of some constant C_{thresh} (and $-C_{\mathrm{thresh}}$), and (b) use the same perturbation variable with a sufficient high probability (*coupling*). These additive noise terms are added to the current utility estimates resulting in perturbed, context-dependent utilities, which in turn specify our imitated CoLST model. The arm having the largest perturbed context-dependent utilities is used as the "first arm" i_t for the action pair, as this one is most likely to win any duel according to our imitated CoLST model. Roughly speaking, the multiplication of the generated perturbation variables with the confidence widths (of the current estimate) takes the degree of accuracy of the imitated CoLST model into account. The “second arm” j_t is chosen as the first arm’s toughest competitor, i.e., the arm that has the highest (optimistic) chance to win the duel against the latter according to the current upper confidence width (represented by $c_1\|\boldsymbol{z}_{t,i_t,i}\|_{\mathbf{M}_t^{-1}}$). Such a choice mechanism for the second arm is common in the realm of non-contextual dueling bandits and is tailored towards the task of average regret minimization (Bengs et al., 2021).

The underlying idea of choosing the first arm is inspired by randomized exploration strategies used in the follow-the-perturbed-leader approach (Kim & Tewari, 2019) or the RandUCB algorithm (Vaswani et al., 2020) for numerical bandit problems. However, in our preference-based setting, we have to deal with a more complex action space leading to a different theoretical analysis. In addition, unlike the numerical setting, we do not necessarily need to adopt the optimism in the face of uncertainty principle in the sense that only positive perturbation values need to be sampled. Finally, thanks to the underlying CoLST model, we have a natural candidate for the perturbation distribution.

3.2.1. Near-optimality

For our theoretical analysis of CoLSTIM, we make the following assumptions:

(A1) $\mathcal{X} = \mathcal{B}_1(2)$ and there exist basis vectors $\{b_j\}_{j=1}^d \subset \{\boldsymbol{x}_{t,i}\}_{i=1}^n$ such that $\rho \boldsymbol{I}_d \leq \sum_{j=1}^d b_j b_j^\top$ for some $\rho > 0$, i.e., the context vectors span the d-dimensional Euclidean space.

(A2) There exists some $L > 0$ such that the derivative of the comparison function F^* is L-Lipschitz continuous, i.e., $|(F^*)'(x) - (F^*)'y)| \leq L|x - y|$ for any $x, y \in \mathbb{R}$. Moreover,

$$\mu := \inf\left\{(F^*)'(\boldsymbol{x}^\top\boldsymbol{\theta}) \,\middle|\, \|\boldsymbol{x}\| \leq 2,\ \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 1\right\} > 0.$$

These assumptions are quite common in the realm of contextual bandits (Li et al., 2017; Vaswani et al., 2020). Note that assumption **(A2)** holds for the comparison functions of the prominent LST models such as the Bradley-Terry-Luce and the Thurstone-Mosteller model (see Section 2). We obtain the following theoretical guarantees for CoLSTIM (proven in Section C in the supplementary material).

Theorem 3.2. *Let* $c_1 = \frac{1}{2\mu}\sqrt{d\log(T/d) + 2\log(T)}$ *and* $\tau = d + \max\{^{d^2\log(T)}/_{\mu^2\rho}, d/\rho\}$. *Further, let* $c_2 \in (0, c_1]$ *be some constant, and for any time step* t, *set* $p_t = \min\left(1, \frac{\sqrt{2d}}{2\sqrt{t-\tau}}\left(3\,c_1 + c_2\right)\sqrt{\log\left(\frac{2T}{d}\right)}\right)$. *For any* $C_{\text{thresh}} \in (0, c_2)$, *it holds that the expected cumulative average regret of* CoLSTIM *is in* $O(d\sqrt{T}\log(T))$ *for any CoLST model with weight vector* $\boldsymbol{\theta}^*$ *such that* $\|\boldsymbol{\theta}^*\| \in \mathcal{B}_1(2)$ *and comparison function* F^* *satisfying assumption* ***(A2)***, *and any* $(\mathbf{X}_t)_{t\in[T]}$ *satisfying assumption* ***(A1)***.

If we make an additional assumption on the smallest eigenvalues of the Gram matrix $\mathbf{M}_t$, we can show a bound $\tilde{O}(\sqrt{dT})$ almost matching (up to logarithmic terms) the lower bound $\Omega(\sqrt{dT})$ shown by Saha (2021). The proof is deferred to Section C in the supplementary material.

Corollary 3.3. *Under the assumptions of Theorem 3.2 and if* $\sum_{t=\tau+1}^T \lambda_{\min}^{-1/2}(\mathbf{M}_t) \leq c\sqrt{T}$, *where* c *is some positive constant and* $\lambda_{\min}(A)$ *denotes the smallest eigenvalue of a square matrix* A, *the expected cumulative average regret of* CoLSTIM *is in* $O(\sqrt{dT}\log(T))$.

The condition on the Gram matrix in the Corollary 3.3 is satisfied if the context vectors are dense (Li et al., 2017).

Algorithm 2 Sup-CoLSTIM

```
Input: $\tau > 0$, $G$, $F$, $C_{\text{thresh}} > 0$, $p_t \in [0,1]$, $c_1 > 0$
Initialization: Same as in CoLSTIM
Set $S = \lfloor \log_2 T \rfloor$, $\Psi^{(0)} = \emptyset$, $\Psi^{(1)} = \ldots = \Psi^{(S)} = [\tau]$
for $t = \tau + 1, 2, \ldots, T$ do
    Observe context vectors $\mathbf{X}_t = (\boldsymbol{x}_{t,1} \ldots \boldsymbol{x}_{t,n})$
    Set $s = 1$ and $A_t^{(s)} = [n]$
    while $i_t, j_t$ not found do
        Compute MLE $\hat{\boldsymbol{\theta}}_t^{(s)}$ via (9) (or (10)) using only data from the time steps in $\Psi^{(s)}$
        Set $\mathbf{M}_t^{(s)} = \sum_{l\in\Psi^{(s)}} \boldsymbol{z}_{l,i_l,j_l}\boldsymbol{z}_{l,i_l,j_l}^\top$
        Set $w_{i,j}^{(s)}(\mathbf{X}_t) = c_1\|\boldsymbol{z}_{t,i,j}\|_{(\mathbf{M}_t^{(s)})^{-1}} \forall i,j \in [n]$
        if $w_{i,j}^{(s)}(\mathbf{X}_t) \leq 1/\sqrt{T}, \quad \forall i,j \in A_t^{(s)}$ then
            Sample $\epsilon_{t,i}$ $\forall i \in [n]$ as in lines 7–13 of Alg.1
            $\Psi^{(0)} = \Psi^{(0)} \cup \{t\}$
            $i_t = \arg\max_{i\in A_t^{(s)}} \boldsymbol{x}_{t,i}^\top\hat{\boldsymbol{\theta}}_t^{(s)} + \epsilon_{t,i}\|\boldsymbol{x}_{t,i}\|_{(\mathbf{M}_t^{(s)})^{-1}}$
            $j_t = \arg\max_{i\in A_t^{(s)}} \langle \boldsymbol{z}_{t,i,i_t}, \hat{\boldsymbol{\theta}}_t^{(s)}\rangle + w_{t,i_t,j}^{(s)}(\mathbf{X}_t)$
        else if $w_{i,j}^{(s)}(\mathbf{X}_t) \leq 1/2^s, \quad \forall i,j \in A_t^{(s)}$ then
            $A_t^{(s+1)} = \{i \in A_t^{(s)} \,|\, \boldsymbol{x}_{t,i}^\top\hat{\boldsymbol{\theta}}_t^{(s)} + 2^{-s} \geq \max_{j\in A_t^{(s)}} \boldsymbol{x}_{t,j}^\top\hat{\boldsymbol{\theta}}_t^{(s)}\}$
            $s \leftarrow s + 1$
        else
            $\Psi^{(s)} = \Psi^{(s)} \cup \{t\}$
            Choose $(i_t, j_t)$ uniformly at random from $\{i,j \in A_t^{(s)} \,|\, w_{i,j}^{(s)}(\mathbf{X}_t) > 1/2^s\}$
        end if
    end while
    Choose $(i_t, j_t)$ and observe $Y_t = 1_{[i_t \succ j_t]}$
end for
```

Note that the theoretical results in fact do not depend on the knowledge of the true perturbation distribution G^*, but rather it is sufficient to know the true comparison function F^*. Finally, it is worth noting that the threshold C_{thresh} can be made arbitrarily small, while the theoretical results will still hold (as long as c_2 is not even smaller). In this case, the choice of the first arm i_t would essentially correspond to a greedy choice, which prevents exploration from the outset. However, thanks to the (optimistic) choice mechanism of the second arm j_t, the algorithm would still explore (i.e., learn about the structure of the bandit problem) and not "get stuck".

3.2.2. SUP-COLST IMITATOR

By adapting the technique introduced by Auer (2002) for the underlying learning scenario (cf. (Chu et al., 2011; Li et al., 2017; Xue et al., 2020)), we can extend the COLSTIM algorithm to SUP-COLSTIM (Algorithm 2) in order to obtain a regret bound of order $\tilde{O}(\sqrt{dT\log(n)})$ without making an additional assumptions on the Gram matrix as in Corollary 3.3. The idea is to embed the choice mechanism of COLSTIM into a stage-wise approach which keeps track of "sufficiently accurately estimated promising arms" (cf. Section 3.2 in (Saha, 2021)).

Theorem 3.4. *Let $c_1 = \frac{3}{2\mu}\sqrt{2\log(3nT^2)}$ and τ as in Theorem 3.2. Further, let $c_2 \in (0, c_1]$ be some constant and p_t be as in Theorem 3.2. Under the assumptions of Theorem 3.2, it holds for any $C_{\text{thresh}} < c_2$ that the expected cumulative average regret of* SUP-COLSTIM *is in $O(\sqrt{dT\log(n)}\log^{3/2}(T))$.*

To the best of our knowledge, this is the first time that the technique introduced by Auer (2002) is combined with a randomized learning strategy and analyzed theoretically (see Section D in the supplementary material for the proof).

3.2.3. COMPUTATIONAL ASPECTS

It is evident that the computation of the MLE involves a computationally expensive operation, as the entire history is used for this estimation step — an issue shared by most of the algorithms for logistic bandits (Filippi et al., 2010; Li et al., 2017; Vaswani et al., 2020) or stochastic contextual dueling bandits (Saha, 2021). However, instead of optimizing the (log-)likelihood function based on the entire history in each time step, we could also optimize the (log-)likelihood function more efficiently in an online manner using stochastic gradient descent. More precisely, we could replace line 6 in Algorithm 1 (and line 9 in Algorithm 2 accordingly) by

$$\hat{\boldsymbol{\theta}}_t \leftarrow \hat{\boldsymbol{\theta}}_{t-1} + \eta_t \nabla \boldsymbol{l}\big(\hat{\boldsymbol{\theta}}_{t-1} \,|\, Y_{t-1}, \{i_{t-1}, j_{t-1}\}, \mathbf{X}_{t-1}\big)$$

for some suitable parameter $\eta_t > 0$ (learning rate). Although the theoretical guarantees shown do not hold for this SGD variant, we do not see much of a difference regarding the regret in our experiments by using the SGD variant instead of maximizing the likelihood on the entire history of data in each time step (see Section E). However, we see a clear advantage of the former over the latter with respect to the cumulative elapsed time to make a choice. It is worth mentioning that the computational costs of COLSTIM's arm selection (lines 7–15) is quite low due to the simple form of the two maximization problems to be solved.

4. Experiments

In this section, we present experimental results for our learning algorithm for the contextual dueling bandits setting for the two most prominent LST models, namely the Bradley-Terry-Luce (BTL) model and the Thurstone-Mosteller (TM) model. We compare our approach with Double-Thompson Sampling (DTS) (Wu & Liu, 2016) and Self-Sparring (SS) with independent beta priors for each arm (Sui et al., 2017), which are state-of-the art algorithms for the non-contextual dueling bandits setting, as well as Maximum-Informative-Pair (MaxInP) (Saha, 2021), which is suitable for the contextual dueling bandits setting under the contextualized BTL model. Moreover, we include the random choice strategy (Random), which is choosing the pair of arms in each time step uniformly at random. For DTS and SS, we used the same hyperparameters as in the corresponding experiments, while for MaxInP, we simply use $t_0 = dn$ and $\eta = \sqrt{d\log(T)}$, as the hyperparameters are not reported in the experiments by Saha (2021). Note that our choices are simplified versions of the parameters for which the theoretical guarantees hold. For COLSTIM, we use $\tau = t_0$ and $c_1 = C_{\text{thresh}} = \eta$, as τ has a similar role as t_0 and c_1 or C_{thresh} have a similar role as η, respectively. For the coupling probability p_t of COLSTIM, we use a simplified version of the one derived in Theorem 3.2, namely $p_t = \min\big(1, \frac{d}{\sqrt{t-\tau}}\log(dT)\big)$. In every experiment, the performances of the algorithms are measured in terms of cumulative average regret (cf. (8)), averaged across 100 runs and reported with their standard deviation. Moreover, for both MaxInP and COLSTIM, we use the SGD-based variant described in Section 3.2 with a fixed learning rate of $\eta_t = 1/2$. For a comparison of the SGD-based variant and the "full MLE" variants see Section E. We omit SUP-COLSTIM and Sta'D (Saha, 2021) since it is known that algorithms adopting the stage-wise approach of Auer (2002) tend to perform poorly in numerical simulations. All these experiments were conducted on a machine featuring an Intel(R) Xeon E5-2670 @2.6GHz with 16 cores and 64 GB of RAM.

4.1. Contextual Setting

Recall that a problem instance P in our setting is specified by the number of available arms n, the dimension of the context vectors d, the perturbation distribution G^* and the weight parameter $\boldsymbol{\theta}^*$, such that we accordingly write $P = P(n, d, G, \boldsymbol{\theta}^*)$. Following Saha (2021), for fixed n, d, G^*, we distinguish three problem scenarios with respect to the ℓ_2-norm of the weight parameter $\boldsymbol{\theta}^*$:

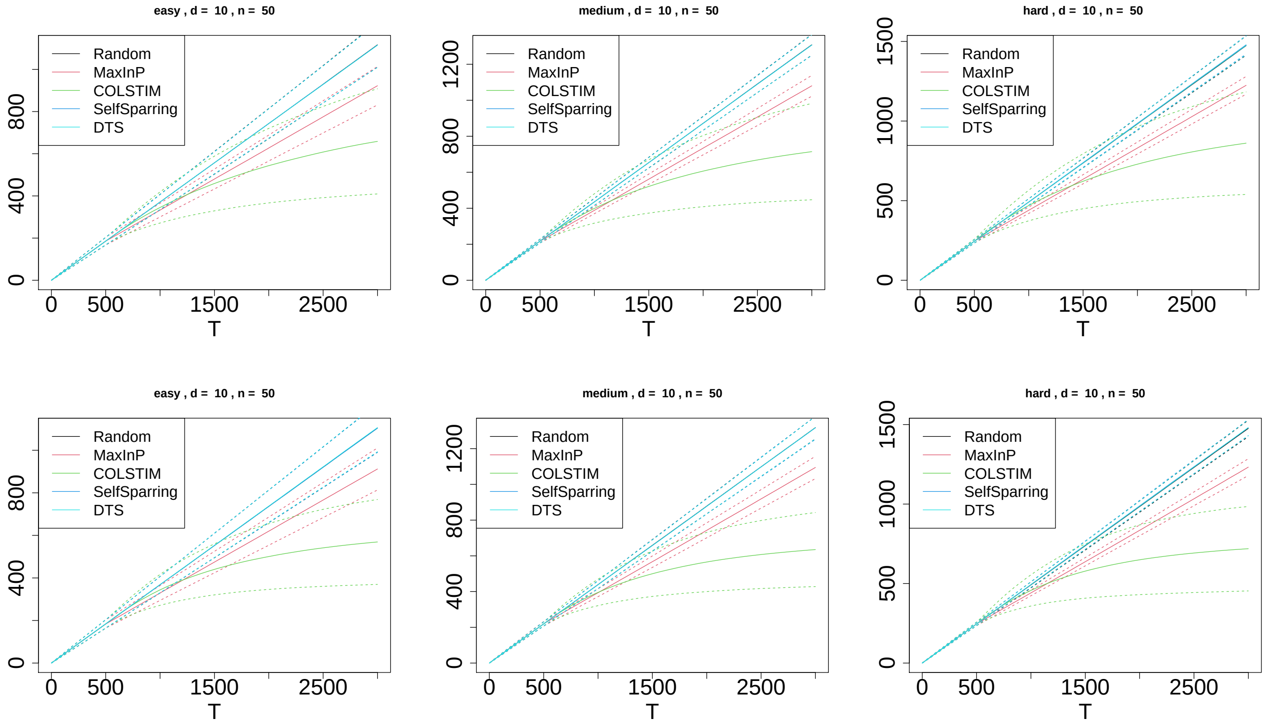

Figure 1. Top panel: Averaged cumulative regret of the different methods on $E(d,n,G^*)$ (left), $M(d,n,G^*)$ (middle) and $H(d,n,G^*)$ (right) for G^* being standard Gumbel. Bottom panel: Averaged cumulative regret of the different methods on $E(d,n,G^*)$ (left), $M(d,n,G^*)$ (middle) and $H(d,n,G^*)$ (right) for G^* being the standard normal distribution.

- $E(d,n,G^*) = \bigcup_{\|\boldsymbol{\theta}^*\| \le 1/\sqrt{d}} P(n,d,G^*,\boldsymbol{\theta}^*)$;
- $M(d,n,G^*) = \bigcup_{1/\sqrt{d} \le \|\boldsymbol{\theta}^*\| \le 1} P(n,d,G^*,\boldsymbol{\theta}^*)$;
- $H(d,n,G^*) = \bigcup_{1 \le \|\boldsymbol{\theta}^*\| \le \sqrt{d}} P(n,d,G^*,\boldsymbol{\theta}^*)$.

which we refer to as the easy, medium and hard problem scenario, respectively.

Table 1. Averaged cumulative runtimes (in seconds) and the corresponding standard deviations (in brackets) of the different methods for the different problem scenarios.

G^* = Gumbel	$E(d,n,G^*)$	$M(d,n,G^*)$	$H(d,n,G^*)$
Random	0.19 (0.02)	0.18 (0.01)	0.18 (0.01)
MaxInP	164.58 (6.58)	155.54 (4.92)	155.51 (3.49)
CoLSTIM	7.44 (0.51)	7.08 (0.41)	6.98 (0.17)
SS	9.92 (0.66)	9.43 (0.56)	9.39 (0.22)
DTS	52.01 (2.63)	49.73(2.04)	50.15 (1.08)
G^* = Gaussian	$E(d,n,G^*)$	$M(d,n,G^*)$	$H(d,n,G^*)$
Random	0.19 (0.02)	0.18 (0.02)	0.18 (0.02)
MaxInP	159.33 (6.00)	157.70 (8.54)	156.56 (7.48)
CoLSTIM	7.09 (0.38)	7.05 (0.51)	7.03 (0.48)
SS	9.49 (0.50)	9.47 (0.67)	9.47 (0.60)
DTS	50.94 (2.45)	50.19 (2.99)	49.93 (2.69)

In each run on one of the scenarios, the weight parameters are sampled uniformly at random from the corresponding subset of the ℓ_2-Ball, while the context vectors are sampled uniformly at random from $\mathcal{B}_1(2)$, i.e., the unit ℓ_2-Ball, for each time step t within one run.

The results for G^* being the standard Gumbel distribution are illustrated in the top panel of Figure 1, while the bottom panel of Figure 1 shows the results for G^* being the standard normal distribution, both for $n = 50$ and $d = 10$. Our CoLSTIM method using the corresponding perturbation distribution $G = G^*$ outperforms the other methods in all scenarios, while MaxInP performs only slightly better than the naïve random selection strategy. Unsurprisingly, the non-contextual methods DTS and SS are hardly distinguishable from the latter.

Table 1 reports the averaged cumulative runtimes and the corresponding standard deviations of each of the considered methods for making the choice of the pairs (i.e., the MLE step for CoLSTIM and MaxInP are neglected). The results confirm the discussion in Section 1 regarding the computational efficiency of CoLSTIM's choice mechanism, which, quite interestingly, is competitive to the non-contextual methods.

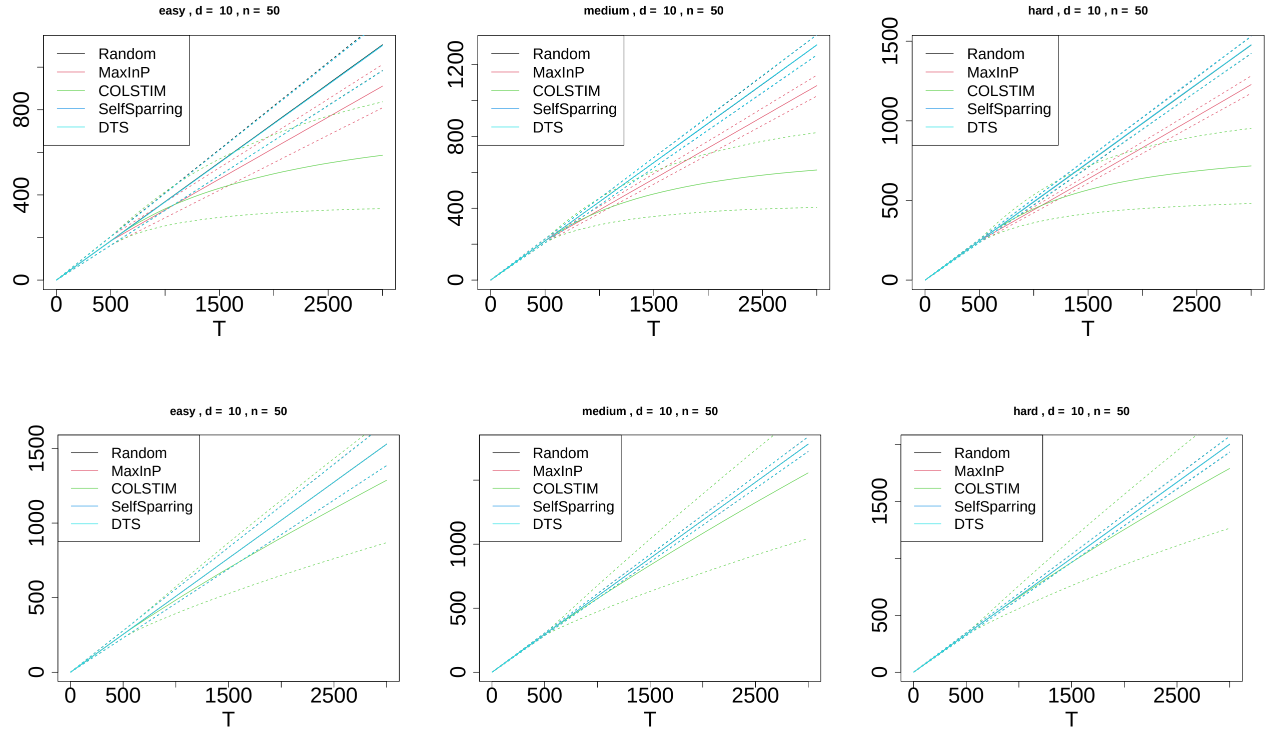

Figure 2. Top panel: Averaged cumulative regret of the different methods on $E(d, n, G^*)$ (left), $M(d, n, G^*)$ (middle) and $H(d, n, G^*)$ (right) for G^* being standard Gumbel. Bottom panel: Averaged cumulative regret of the different methods on $E(d, n, G^*)$ (left), $M(d, n, G^*)$ (middle) and $H(d, n, G^*)$ (right) for G^* being the normal distribution with zero and variance 5. Here, CoLSTIM is used with G being the standard normal distribution (misspecification).

4.2. Misspecification

Given the results of the previous section, one may wonder how sensitive our CoLSTIM method is to misspecification of the perturbation distribution G, i.e., how it performs when $G \neq G^*$? To this end, we repeat the experiments for the easy, medium and hard problem instances with G^* being the standard Gumbel distribution, while the CoLSTIM method is used with G being the standard normal distribution. Note that MaxInP is designed for exactly this case, in which the perturbation distribution of the underlying CoLST feedback model is the standard Gumbel distribution, so that MaxInP is in a favorable position. The top panel in Figure 2 illustrates the result for this setting, which shows a similar picture as Figure 1, where no misspecification is present.

Finally, we also consider the case where there is an incorrect specification of the parameters in the correct parametric distribution family. More specifically, the bottom panel shows the result of the experiments for the easy, medium and hard problem instances with G^* being the normal distribution with zero mean and variance 5 and CoLSTIM still uses the standard normal distribution for G. Although our algorithm now performs slightly worse and also has a higher variance, it still performs better than the other algorithms as they still do not show a convergence behavior at all, and our algorithm will probably outperform them eventually for longer time horizons T.

5. Related Work

So far, a contextualized extension of preference-based bandits with pairwise comparisons (i.e., dueling bandits) has only been studied explicitly by Dudík et al. (2015) and recently by Saha (2021). In the former work, it is assumed that the learner has access to a space of learning policies from which the contextualized von Neumann winner should be found. This learning setting is different from ours, due to the absence of a learning policy space. Saha (2021) assumes a latent linear utility score for each arm as well as a logit link function for modeling the feedback governed by the pairwise preference probabilities. This modeling assumption is a special case of the linear stochastic transitivity models considered in this paper and corresponds to the (contextualized) Bradley-Terry-Luce model. Nevertheless, a lower bound for the regret in this special case of order $\Omega(\sqrt{dT})$ is derived, which is also a lower bound for our more general setting. Further, two learning algorithms are suggested and theoretically analyzed, which both essentially use the pair of arms having the highest uncertainty in each round. Our proposed learning algorithm, however, uses the idea of randomized learning strategies, which have gained popularity in reward-based bandit problems in recent years (see Vaswani et al. (2020) for a detailed overview). This can be attributed to the fact that, unlike the well-known Thompson sampling algorithm, no closed posterior distribution is needed, but only a sampling distribution.

Finally, it should be mentioned that the utility-based dueling bandits learning scenario with infinitely many arms, where each arm is a d-dimensional point, is closely related to the learning scenario considered here. In fact, in the latter scenario, if only K many arms, which can vary, are allowed to be selected in each learning round, this corresponds to the contextual learning scenario considered here. However, not all learning algorithms are directly amenable to this adaptation (Yue & Joachims, 2009), or if they are, they no longer have theoretical guarantees (Ailon et al., 2014; Kumagai, 2017), or even did not have them before (González et al., 2017).

6. Conclusion and Future Work

We studied the contextualized version of the dueling bandits problem under linear stochastic transitivity models and proposed an algorithm for effective learning in this setting. The algorithm is inspired by randomized learning strategies for numerical bandit problems and has both satisfactory regret bounds as well as excellent numerical performance. In addition, we have also shown a lower bound on the weak regret, which is of the same order as for the average regret. This indicates that stronger theoretical guarantees in the worst case sense cannot be obtained for weak regret learners.

For future work, an extension to more general action sets than pairs of arms would be interesting from both a theoretical and a practical point of view as considered by recent works (Brost et al., 2016; Saha & Gopalan, 2018; 2019; Bengs & Hüllermeier, 2020; Agarwal et al., 2020). Context-dependent random utility models (Train, 2009) could be a practically meaningful counterpart for linear stochastic transitivity models. Another interesting question would be whether one could replace the linearity assumption on the latent utility values with more general assumptions such as a kernel-based correlation as considered by Sui et al. (2017) for the non-contextual dueling bandits setting. Although our experiments show that our proposed method is robust to misspecification of the perturbation distribution and the corresponding comparison function, one could imagine learning this perturbation distribution online as well. For this purpose, the work by Oliveira et al. (2018) could be relevant, which, however, deals with the batch learning scenario. Yet another interesting question would be to investigate whether Assumption **(A1)** can be relaxed to an assumption like in (Li et al., 2017), which is also used by MaxInP (Saha, 2021). Last but not least, it would certainly be worthwhile to investigate the performance of our suggested algorithm in real-world applications, e.g., in realtime algorithm configuration or online learning-to-rank problems for which preference-based bandit algorithms have been used before (Brost et al., 2016; Schuth et al., 2016; Oosterhuis et al., 2016; Zhao & King, 2016; Schäfer & Hüllermeier, 2018; El Mesaoudi-Paul et al., 2020).

Acknowledgments

The authors would like to thank the Paderborn Center for Parallel Computation (PC2) for the use of the OCuLUS cluster.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2312–2320, 2011.

Agarwal, A., Johnson, N., and Agarwal, S. Choice bandits. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18399–18410, 2020.

Ailon, N., Karnin, Z., and Joachims, T. Reducing dueling bandits to cardinal bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 856–864, 2014.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2002.

Bengs, V. and Hüllermeier, E. Preselection bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 778–787, 2020.

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

Brost, B., Seldin, Y., Cox, I. J., and Lioma, C. Multi-dueling bandits and their application to online ranker evaluation. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 2161–2166, 2016.

Cattelan, M. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27 (3):412–433, 2012.

Cheng, W., Hüllermeier, E., and Dembczynski, K. J. Label ranking methods based on the Plackett-Luce model. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 215–222, 2010.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 2011.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pp. 563–587, 2015.

El Mesaoudi-Paul, A., Weiß, D., Bengs, V., Hüllermeier, E., and Tierney, K. Pool-based realtime algorithm configuration: A preselection bandit approach. In *International Conference on Learning and Intelligent Optimization*, pp. 216–232. Springer, 2020.

Elo, A. E. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 23, pp. 586–594, 2010.

González, J., Dai, Z., Damianou, A., and Lawrence, N. D. Preferential Bayesian optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1282–1291, 2017.

Herbrich, R., Minka, T., and Graepel, T. Trueskill: A Bayesian skill rating system. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 569–576, 2006.

Hunter, D. R. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

Jin, T., Xu, P., Gu, Q., and Farnoud, F. Rank aggregation via heterogeneous Thurstone preference models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 4353–4360, 2020.

Kim, B. and Tewari, A. On the optimality of perturbations in stochastic and adversarial multi-armed bandit problems. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2695–2704, 2019.

Kumagai, W. Regret analysis for continuous dueling bandit. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1488–1497, 2017.

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

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2071–2080. PMLR, 2017.

Oliveira, I. F., Ailon, N., and Davidov, O. A new and flexible approach to the analysis of paired comparison data. *The Journal of Machine Learning Research*, 19(1): 2458–2486, 2018.

Oosterhuis, H., Schuth, A., and de Rijke, M. Probabilistic multileave gradient descent. In *Proceedings of European Conference on Information Retrieval (ECIR)*, pp. 661–668, 2016.

Popescu, P. G., Dragomir, S., Sluşanschi, E. I., and Stanaşila, O. N. Bounds for Kullback-Leibler divergence. *Electronic Journal of Differential Equations*, 2016, 2016.

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

Saha, A. and Gopalan, A. Battle of bandits. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 805–814, 2018.

Saha, A. and Gopalan, A. Combinatorial bandits with relative feedback. pp. 983–993, 2019.

Schäfer, D. and Hüllermeier, E. Dyad ranking using Plackett–Luce models based on joint feature representations. *Machine Learning*, 107(5):903–941, 2018.

Schuth, A., Oosterhuis, H., Whiteson, S., and de Rijke, M. Multileave gradient descent for fast online learning to rank. In *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 457–466, 2016.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. Multidueling bandits with dependent arms. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Train, K. E. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 654–663, 2013.

Vaswani, S., Mehrabian, A., Durand, A., and Kveton, B. Old dog learns new tricks: Randomized UCB for bandit problems. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1988–1998, 2020.

Wu, H. and Liu, X. Double Thompson sampling for dueling bandits. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 649–657, 2016.

Xue, B., Wang, G., Wang, Y., and Zhang, L. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2936–2942, 2020.

Yue, Y. and Joachims, T. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1201–1208, 2009.

Zhao, T. and King, I. Constructing reliable gradient exploration for online learning to rank. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1643–1652, 2016.

A. List of Symbols

The following table contains a list of symbols that are frequently used in the main paper as well as in the following supplementary material.

	Basics
$\prec, \succ$	preference relation for objects, i.e., $o \succ o'$ (or $o \prec o'$) iff object o is (not) preferred over object o'
$1_{[\cdot]}$	indicator function
$\mathbb{N}$	set of natural numbers (without 0), i.e., $\mathbb{N} = \{1, 2, 3, \dots\}$
$\mathbb{R}$	set of real numbers
$[n]$	the set $\{1, 2, \dots, n\}$ for some $n \in \mathbb{N}$
$\boldsymbol{x}, \boldsymbol{z}$	d – dimensional (column) vectors
$A^\top$	transpose of a matrix $A \in \mathbb{R}^{d \times d'}$
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle$	inner product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$
$\boldsymbol{I}_d$	$d \times d$ identity matrix
$\mathbf{0_d}$	$d \times d$ zero matrix
$\lVert \boldsymbol{x} \rVert$	the Euclidean norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$, i.e., $\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$
$\lVert \boldsymbol{x} \rVert_A$	weighted norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$ for some positive semi-definite matrix $A \in \mathbb{R}^{d \times d}$, i.e., $\sqrt{\boldsymbol{x}^\top A \boldsymbol{x}}$
$\mathcal{B}_r(p)$	ℓ_p ball of radius $r \geq 0$, i.e., $\mathcal{B}_r(p) = \{\mathbf{x} \mid \lVert \mathbf{x} \rVert_p \leq r\}$, where $\lVert \cdot \rVert_p$ denotes the standard ℓ_p-norm
	Modelling related
n	number of arms
$\mathcal{A} = [n]$	set of arms
T	the time horizon
$\mathcal{X}$	the context space (subset of $\mathbb{R}^d$)
$\boldsymbol{x}_{t,i}$	(d – dimensional) context vector related to arm i at time step $t \in [T]$ (element in $\mathcal{X}$)
$\mathbf{X}_t$	($d \times n$ – dimensional) context matrix at time step t, i.e., $\mathbf{X}_t = (\boldsymbol{x}_{t,1} \dots \boldsymbol{x}_{t,n})$
$\boldsymbol{z}_{t,i,j}$	the contrast vector of arm i and j at time step t, i.e., $\boldsymbol{z}_{t,i,j} = \boldsymbol{x}_{t,i} - \boldsymbol{x}_{t,j}$
$\mathbf{Z}_t$	($d \times \binom{n}{2}$– dimensional) contrast matrix at time step t, i.e., $\mathbf{Z}_t = (\boldsymbol{z}_{t,1,2}\, \boldsymbol{z}_{t,1,3} \dots \boldsymbol{z}_{t,1,n}\, \boldsymbol{z}_{t,2,3} \dots \boldsymbol{z}_{t,n-1,n})$
$\boldsymbol{\theta}^*$	ground truth weight parameter of the underlying CoLST model (see (6))
G^*, F^*	ground truth perturbation distribution and corresponding comparison function of the underlying CoLST model (see (6))
$S_t = \{i_t, j_t\}$	selected pair of arms at time step t
Y_t	(binary) feedback for the selected pair of arms S_t, i.e., $Y_t = 1_{[i_t \succ j_t]}$
u_i^*	mapping from $\mathbb{R}^{d \times n}$ to $\mathbb{R}$ according to $u_i^*(\mathbf{X}) = \boldsymbol{x}_i^\top \boldsymbol{\theta}^*$, where $\boldsymbol{x}_i$ is the ith column of $\mathbf{X}$
$i^*(t) = i_t^*$	optimal arm at time step t, i.e., the arm with the highest context-dependent utility at time t $i^*(t) = \arg\max_{i \in \mathcal{A}} u_i^*(\mathbf{X}_t)$
$r_t^w(S_t, \mathbf{X}_t)$	instantaneous weak regret at time step t for selecting S_t i.e., $r_t^w(S_t, \mathbf{X}_t) = u_{i^*(t)}^*(\mathbf{X}_t) - \max_{k \in S_t} u_k^*(\mathbf{X}_t)$
$r_t^a(S_t, \mathbf{X}_t)$	instantaneous average regret at time step t for selecting S_t i.e., $r_t^a(S_t, \mathbf{X}_t) = \frac{2u_{i^*(t)}^*(\mathbf{X}_t) - u_{i_t}^*(\mathbf{X}_t) - u_{j_t}^*(\mathbf{X}_t)}{2}$
R_T^w, R_T^a	cumulative weak or average regret till time T for selecting $(S_t)_{t \in [T]}$ (see (8))
	CoLSTIM related
$\hat{\boldsymbol{\theta}}_t$	maximum-likelihood estimate of the weight parameter of the underlying CoLSTIM (see (9))
$c_1 > 0$	confidence width parameter of CoLSTIM (hyperparameter of CoLSTIM)
$C_{\text{thresh}} > 0$	threshold parameter of CoLSTIM (hyperparameter of CoLSTIM)
$\tau > 0$	pure exploration length of CoLSTIM (hyperparameter of CoLSTIM)
$p_t \in [0, 1]$	coupling probability (hyperparameter of CoLSTIM)
G, F	perturbation distribution and comparison function used by CoLSTIM (hyperparameter of CoLSTIM)
$\mathbf{M}_t$	Gram matrix at time step t, i.e., $\mathbf{M}_t = \sum_{s=1}^{t-1} \boldsymbol{z}_{s,i_s,j_s} \boldsymbol{z}_{s,i_s,j_s}^\top$
$\tilde{\epsilon}_{t,i}$	sampled perturbation variable for arm i at time step t, i.e., $\tilde{\epsilon}_{t,i} \sim G$ (see line 7 of Alg. 1)
$\epsilon_{t,i}$	trimmed sampled perturbation variable for arm i at time step t, i.e., $\epsilon_{t,i} = \min(C_{\text{thresh}}, \max(-C_{\text{thresh}}, \tilde{\epsilon}_{t,i}))$
$\hat{u}_{t,i}(\mathbf{X}_t)$	estimated utility of arm i in time step t, i.e., $\hat{u}_{t,i}(\mathbf{X}_t) = \boldsymbol{x}_{t,i}^\top \hat{\boldsymbol{\theta}}_t$
$\tilde{u}_{t,i}(\mathbf{X}_t)$	perturbed estimated utility of arm i in time step t, i.e., $\tilde{u}_{t,i}(\mathbf{X}_t) = \hat{u}_{t,i}(\mathbf{X}_t) + \epsilon_{t,i} \lVert \boldsymbol{x}_{t,i} \rVert_{\mathbf{M}_t^{-1}}$
	Sup-CoLSTIM related
$c_1, C_{\text{thresh}}, \tau, p_t, G, F$	hyperparameters of Sup-CoLSTIM (same meaning as for CoLSTIM)
s, S	stages in Sup-CoLSTIM, maximal number of stages (set to $\lfloor \log_2(T) \rfloor$)
$A_t^{(s)}$	arms which are “sufficiently accurately estimated promising arms” at stage s and time step t
$\Psi^{(0)}$	all time steps, where the choice has been made at some stage $s \in [S]$ according to CoLSTIM’s choice mechansim
$\Psi^{(s)}$	all time steps, where the choice has been made at stage $s \in [S]$, but not according to CoLSTIM’s choice mechansim
$\hat{\boldsymbol{\theta}}_t^{(s)}$	maximum-likelihood estimate of the weight parameter using only observations from the time steps in $\Psi^{(s)}$
$\mathbf{M}_t^{(s)}$	Gram matrix at time step t and stage s i.e., $\mathbf{M}_t^{(s)} = \sum_{l \in \Psi^{(s)}} \boldsymbol{z}_{l,i_l,j_l} \boldsymbol{z}_{l,i_l,j_l}^\top$
$\tilde{\epsilon}_{t,i}$, $\epsilon_{t,i}$	(trimmed) perturbation variables (same meaning as for CoLSTIM)
$\hat{u}_{t,i}^{(s)}(\mathbf{X}_t)$	estimated utility of arm i in stage s at time t, i.e., $\hat{u}_{t,i}^{(s)}(\mathbf{X}_t) = \boldsymbol{x}_{t,i}^\top \hat{\boldsymbol{\theta}}_t^{(s)}$
$\tilde{u}_{t,i}^{(s)}(\mathbf{X}_t)$	perturbed estimated utility of arm i in stage s at time t, i.e., $\tilde{u}_{t,i}^{(s)}(\mathbf{X}_t) = \hat{u}_{t,i}^{(s)}(\mathbf{X}_t) + \epsilon_{t,i} \lVert \boldsymbol{x}_{t,i} \rVert_{(\mathbf{M}_t^{(s)})^{-1}}$

B. Lower Bound: Proof of Theorem 3.1

Theorem 3.1 *For any learning algorithm $\mathcal{A}$ for the contextual dueling bandit problem under linear stochastic transitivity models in dimension d, there exists an instance of the problem characterized by a weight vector $\boldsymbol{\theta}^* \in \mathcal{B}_1(2)$, context space $\mathcal{X} \subseteq \mathcal{B}_1(\infty)$, and a comparison function $F : \mathbb{R} \to [0,1]$, such that the the expected weak regret incurred by $\mathcal{A}$ in any $T > \max(d, 16)$ rounds is*

$$\mathbb{E}\big[R_T^w((S_t^{\mathcal{A}})_{t\in[T]})\big] = \Omega\Big(d\sqrt{T}\Big).$$

Further, if we restrict $\mathcal{X} \subseteq \mathcal{B}_1(2)$, then

$$\mathbb{E}\big[R_T^w((S_t^{\mathcal{A}})_{t\in[T]})\big] = \Omega\Big(\sqrt{dT}\Big).$$

Proof. **Case 1. $\mathcal{X} \subseteq \mathcal{B}_1(\infty)$:**

Our proof is based on a similar line of reasoning as that proposed by Lattimore & Szepesvári (2020) (Chapter 24) for the analysis of linear bandit algorithms. We assume the context space $\mathcal{X}$ to be $\{-1,1\}^d \subset \mathcal{B}_1(\infty)$, and let $\Theta = \big\{-\frac{1}{\sqrt{T}}, \frac{1}{\sqrt{T}}\big\}^d$ be the set of possible unknown weight vectors $\boldsymbol{\theta}^* \in \Theta$. Note since $d < T$, we have $\|\boldsymbol{\theta}\|_2 \le 1$ for any $\boldsymbol{\theta} \in \Theta$, and hence $\Theta \subset \mathcal{B}_1(2)$. Assume the perturbation distribution G of the underlying CoLST is the standard Gumbel distribution, i.e., F corresponds to the BTL model (see Examples in Section 1). Fix any algorithm, and suppose $(\boldsymbol{\ell}_1, \mathbf{r}_1), \ldots (\boldsymbol{\ell}_T, \mathbf{r}_T)$ be the context vectors of the pairs chosen by the algorithm (learner) for T rounds, i.e., $\boldsymbol{\ell}_t := \mathbf{x}_{t,i_t}$ and $\mathbf{r}_t := \mathbf{x}_{t,j_t}$ for all $t \in [T]$. Now, for any t, let us denote by $\mathbf{k}_t \in \{\boldsymbol{\ell}_t, \mathbf{r}_t\}$ the context vector of the chosen arms with the higher utility, i.e.,

$$\mathbf{k}_t = \begin{cases} \boldsymbol{\ell}_t, & \text{if } \boldsymbol{\ell}_t^\top\boldsymbol{\theta}^* > \mathbf{r}_t^\top\boldsymbol{\theta}^*, \\ \mathbf{r}_t, & \text{else.} \end{cases}$$

Denote by $\text{sgn}(x)$ the sign function for any $x \in \mathbb{R}$. Let the sequence of context information matrices $(\mathbf{X}_t)_{t\in[T]}$ be such that there exists exactly one arm $i^* \in [n]$ such that $x_{t,i}^{i^*} = \text{sgn}(\theta_i^*)$, $\forall i \in [d]\, \forall t \in [T]$, where $x_{t,i}^{i^*}$ is the i^{th} component of i^*'s context vector $\boldsymbol{x}_{t,i^*}$ at time t, and θ_i^* is the i^{th} component of $\boldsymbol{\theta}^*$. Thus, the best arm is i^* for each time step t. For sake of brevity, let us denote its corresponding context vector by $\boldsymbol{x}^*$.

For any $\boldsymbol{\theta} \in \Theta$, we denote by $\mathbb{P}_{\boldsymbol{\theta}}$ the measure on (preference) outcomes induced by the interaction of a fixed algorithm $\mathcal{A}$ and the contextual dueling bandit instance parametrized by $\boldsymbol{\theta}$ for F being the BTL model. Also, denote by $\mathbb{E}_{\boldsymbol{\theta}^*}[\cdot]$ the expected regret of $\mathcal{A}$ under the problem instance, induced by the model parameter $\boldsymbol{\theta}^*$.

Writing k_{ti} for the i^{th} component of $\mathbf{k}_t$, note that the expected cumulative weak regret of the algorithm for T rounds can be written as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^*}[R_T^w(\mathcal{A})] &= \mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^T \left(\mathbf{x}_*^\top\boldsymbol{\theta}^* - \mathbf{k}_t^\top\boldsymbol{\theta}^*\right)\right] = \mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^T\sum_{i=1}^d (\text{sgn}(\theta_i^*) - k_{ti})\theta_i^*\right] \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^T\sum_{i=1}^d \mathbb{E}_{\boldsymbol{\theta}^*}\Big[\text{sgn}(\theta_i^*) \ne \text{sgn}(k_{ti})\Big] = \frac{1}{\sqrt{T}}\sum_{i=1}^d \mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^T \text{sgn}(\theta_i^*) \ne \text{sgn}(k_{ti})\right] \\
&\ge \frac{1}{\sqrt{T}}\frac{T}{2}\sum_{i=1}^d \mathbb{P}_{\boldsymbol{\theta}^*}\left(\sum_{t=1}^T (\text{sgn}(\theta_i^*) \ne \text{sgn}(k_{ti})) \ge \frac{T}{2}\right) \\
&= \frac{\sqrt{T}}{2}\sum_{i=1}^d \mathbb{P}_{\boldsymbol{\theta}^*}(i)\,, \qquad (11)
\end{aligned}$$

where $\mathbb{P}_{\boldsymbol{\theta}}(i) := \mathbb{P}_{\boldsymbol{\theta}}\Big(\sum_{t=1}^T (\text{sgn}(\theta_i) \ne \text{sgn}(k_{ti})) \ge \frac{T}{2}\Big)$ for any $\boldsymbol{\theta} \in \Theta$. So the only task left is to suitably lower bound the quantity $\sum_{i=1}^d \mathbb{P}_{\boldsymbol{\theta}^*}(i)$. For this we appeal to the Bretagnolle-Huber inequality (see Theorem 14.2 in (Lattimore & Szepesvári, 2020)), which is used for analyzing the lower bound of linear bandit algorithms as well.

For every $\boldsymbol{\theta} \in \Theta$, and $i \in [d]$, we denote by $\boldsymbol{\theta}^i \in \Theta$ such that $\theta_j^i = \theta_j, \forall j \in [d] \setminus \{i\}$ and $\theta_i^i = -\theta_i$. Let us define $\mathbb{P}_{\boldsymbol{\theta}^i}^c(i) := \mathbb{P}_{\boldsymbol{\theta}^i}\Big(\sum_{t=1}^T (\text{sgn}(\theta_i^i) \ne \text{sgn}(k_{ti})) \le \frac{T}{2}\Big)$. But interestingly note that $\mathbb{P}_{\boldsymbol{\theta}^i}^c(i) = \mathbb{P}_{\boldsymbol{\theta}}(i)$, and now applying

Bretagnolle-Huber inequality we have that for any $\boldsymbol{\theta} \in \Theta$:

$$\mathbb{P}_{\boldsymbol{\theta}}(i) + \mathbb{P}^c_{\boldsymbol{\theta}^i}(i) = 2\mathbb{P}_{\boldsymbol{\theta}}(i) \geq \frac{1}{2}\exp(-KL(\mathbb{P}_{\boldsymbol{\theta}}||\mathbb{P}_{\boldsymbol{\theta}^i}))\,,$$

where $KL(P, Q)$ denotes the KL-divergence between two probability measures P and Q. Recalling that $\mathbb{P}_{\boldsymbol{\theta}}$ denotes the measure on (preference) outcomes induced by the contextual dueling bandit instance $\boldsymbol{\theta} \in \Theta$, we can further use the chain rule of KL-divergence to get

$$2\mathbb{P}_{\boldsymbol{\theta}}(i) \geq \frac{1}{2}\exp(-KL(\mathbb{P}_{\boldsymbol{\theta}}||\mathbb{P}_{\boldsymbol{\theta}^i})) = \frac{1}{2}\exp\Bigg(-\sum_{t=1}^{T} KL\Big(\mathbb{P}_{\boldsymbol{\theta}}(Y_t)||\mathbb{P}_{\boldsymbol{\theta}^i}(Y_t)\Big)\Bigg), \tag{12}$$

Now, since $Y_t \sim \text{Ber}\big(F(\mathbf{z}_t^\top \boldsymbol{\theta})\big)$ with $\mathbf{z}_t = (\boldsymbol{\ell}_t - \mathbf{r}_t)$, we have that $KL\big(\mathbb{P}_{\boldsymbol{\theta}}(Y_t)||\mathbb{P}_{\boldsymbol{\theta}^i}(Y_t)\big)$ is the KL-divergence between two Bernoulli distributions. We bound this by using the following upper bound:

$$KL(\mathbf{p}||\mathbf{q}) \leq \sum_{y \in \mathcal{Y}} \frac{p^2(y)}{q(y)} - 1, \tag{13}$$

where $\mathbf{p}$ and $\mathbf{q}$ are two probability mass functions on some finite set $\mathcal{Y}$ (Popescu et al., 2016, Theorem 1.4).

Recall that we assume a BTL model, so that G is the standard Gumbel distribution and consequently F is the cumulative distribution function of the the standard logistic distribution, so that

$$\mathbb{P}_{\boldsymbol{\theta}}(Y_t) = \text{Ber}\bigg(\frac{e^{\boldsymbol{\ell}_t^\top \boldsymbol{\theta}}}{e^{\boldsymbol{\ell}_t^\top \boldsymbol{\theta}} + e^{\mathbf{r}_t^\top \boldsymbol{\theta}}}\bigg), \qquad \forall t \in [T].$$

For sake of brevity, let us denote $\boldsymbol{\theta}_{\boldsymbol{\ell}} = e^{\boldsymbol{\ell}_t^\top \boldsymbol{\theta}}$ and $\boldsymbol{\theta}_{\mathbf{r}} = e^{\mathbf{r}_t^\top \boldsymbol{\theta}}$; similarly $\boldsymbol{\theta}^i_{\boldsymbol{\ell}} = e^{\boldsymbol{\ell}_t^\top \boldsymbol{\theta}^i}$ and $\boldsymbol{\theta}^i_{\mathbf{r}} = e^{\mathbf{r}_t^\top \boldsymbol{\theta}^i}$. Further note that $\boldsymbol{\theta}^i_{\boldsymbol{\ell}} = \boldsymbol{\theta}_{\boldsymbol{\ell}} e^{-\frac{2}{\sqrt{T}}\text{sgn}(\boldsymbol{\theta}_i)\ell_{ti}}$, and $\boldsymbol{\theta}^i_{\mathbf{r}} = \boldsymbol{\theta}_{\mathbf{r}} e^{-\frac{2}{\sqrt{T}}\text{sgn}(\boldsymbol{\theta}_i)r_{ti}}$. Further, abbreviating $\phi := -\frac{2}{\sqrt{T}}\text{sgn}(\theta_i)\ell_{ti}$ and $\psi := -\frac{2}{\sqrt{T}}\text{sgn}(\theta_i)r_{ti}$ we finally get:

$$\begin{aligned}
KL\Big(\mathbb{P}_{\boldsymbol{\theta}}(Y_t)||\mathbb{P}_{\boldsymbol{\theta}^i}(Y_t)\Big) &= KL\Bigg(\text{Ber}\bigg(\frac{\boldsymbol{\theta}_{\boldsymbol{\ell}}}{\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}}}\bigg)||\text{Ber}\bigg(\frac{\boldsymbol{\theta}^i_{\boldsymbol{\ell}}}{\boldsymbol{\theta}^i_{\boldsymbol{\ell}} + \boldsymbol{\theta}^i_{\mathbf{r}}}\bigg)\Bigg) \\
&\overset{(13)}{\leq} \frac{\Big(1 - \frac{\boldsymbol{\theta}_{\boldsymbol{\ell}}}{\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}}}\Big)^2}{\Big(1 - \frac{e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}}}{e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}} + e^\psi \boldsymbol{\theta}_{\mathbf{r}}}\Big)} + \frac{\Big(\frac{\boldsymbol{\theta}_{\boldsymbol{\ell}}}{\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}}}\Big)^2}{\frac{e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}}}{e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}} + e^\psi \boldsymbol{\theta}_{\mathbf{r}}}} - 1 \\
&= \frac{\boldsymbol{\theta}_{\mathbf{r}}^2 (e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}} + e^\psi \boldsymbol{\theta}_{\mathbf{r}})}{(\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}})^2 e^\psi \boldsymbol{\theta}_{\mathbf{r}}} + \frac{\boldsymbol{\theta}_{\boldsymbol{\ell}}^2 (e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}} + e^\psi \boldsymbol{\theta}_{\mathbf{r}})}{(\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}})^2 e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}}} - 1 \\
&= \frac{(e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}} + e^\psi \boldsymbol{\theta}_{\mathbf{r}})(\boldsymbol{\theta}_{\mathbf{r}}^2 e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\boldsymbol{\ell}}^2 e^\psi \boldsymbol{\theta}_{\mathbf{r}})}{(\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}})^2 e^\psi \boldsymbol{\theta}_{\mathbf{r}} e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}}} - 1 \\
&= \frac{e^\phi \boldsymbol{\theta}_{\boldsymbol{\ell}} + e^\psi \boldsymbol{\theta}_{\mathbf{r}}}{(\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}})^2} \frac{e^\psi \boldsymbol{\theta}_{\boldsymbol{\ell}} + e^\phi \boldsymbol{\theta}_{\mathbf{r}}}{e^\phi e^\psi} - 1 \\
&= \frac{\boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\mathbf{r}}}{(\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}})^2} \frac{\big(e^\phi - e^\psi\big)^2}{e^\phi e^\psi} \\
&\leq \frac{1}{4} \cdot \frac{48}{T} = \frac{12}{T},
\end{aligned} \tag{14}$$

where the last inequality (14) follows from the fact that $\frac{\boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\mathbf{r}}}{(\boldsymbol{\theta}_{\boldsymbol{\ell}} + \boldsymbol{\theta}_{\mathbf{r}})^2} = \frac{1}{\Big(\sqrt{\frac{\boldsymbol{\theta}_{\boldsymbol{\ell}}}{\boldsymbol{\theta}_{\mathbf{r}}}} + \sqrt{\frac{\boldsymbol{\theta}_{\mathbf{r}}}{\boldsymbol{\theta}_{\boldsymbol{\ell}}}}\Big)^2} \leq \frac{1}{2^2} = \frac{1}{4}$, and the second term

$\frac{\left(e^{\phi}-e^{\psi}\right)^2}{e^{\phi}e^{\psi}}$ can shown to be upper bounded by $\frac{48}{T}$ for all $T \geq 16$ as follows:

$$\begin{aligned}
\frac{\left(e^{\phi}-e^{\psi}\right)^2}{e^{\phi}e^{\psi}} &= \frac{(e^{\phi-\psi}-1)^2}{e^{\phi-\psi}} \leq \frac{(e^{4/\sqrt{T}}-1)^2}{e^{4/\sqrt{T}}} && \text{since } \phi-\psi \in \{-\frac{4}{\sqrt{T}}, 0, \frac{4}{\sqrt{T}}\} \\
&\leq (e^{4/\sqrt{T}}-1)^2 && \text{assuming } (T \geq 16) \implies (e^{4/\sqrt{T}} \geq 1) \\
&\leq \left((e-1)\frac{4}{\sqrt{T}}\right)^2 && (\text{using: } e^x - 1 \leq (e-1)x \text{ for } x \in [0,1]) \\
&= \frac{16(e-1)^2}{T} \leq \frac{48}{T}.
\end{aligned}$$

Using this in (12) we further get that:

$$\mathbb{P}_{\boldsymbol{\theta}}(i) \geq \frac{1}{4}\exp\left(-T\frac{12}{T}\right) = \frac{\exp(-12)}{4}.$$

Finally, since this holds for any $\boldsymbol{\theta} \in \Theta$, and $|\Theta| = 2^d$, averaging over all $\boldsymbol{\theta} \in \Theta$ we get:

$$\frac{1}{|\Theta|}\sum_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{d}\mathbb{P}_{\boldsymbol{\theta}}(i) \geq \frac{d\exp(-12)}{4},$$

which implies that there exists at least one $\boldsymbol{\theta} \in \Theta$, say $\tilde{\boldsymbol{\theta}}$, such that $\sum_{i=1}^{d}\mathbb{P}_{\tilde{\boldsymbol{\theta}}}(i) > \frac{d\exp(-12)}{4}$. The claim now simply follows from (11) with the choice of $\boldsymbol{\theta}^* = \tilde{\boldsymbol{\theta}}$.

Case 2. $\mathcal{X} \subseteq \mathcal{B}_1(2)$**:**

For the second part of the proof, it requires us to restrict $\mathcal{X} \subseteq \mathcal{B}_1(2)$. To achieve this, we use the context space $\mathcal{X} = \big\{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\big\}^d$. Note in this case the expected cumulative weak regret of any algorithm $\mathcal{A}$ for T rounds becomes

$$\mathbb{E}_{\boldsymbol{\theta}^*}[R_T^w(\mathcal{A})] = \mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^{T}\left(\mathbf{x}_*^\top\boldsymbol{\theta}^* - \mathbf{k}_t^\top\boldsymbol{\theta}^*\right)\right] = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}(\mathrm{sgn}(\theta_i^*) - k_{ti})\theta_i^*\right] \geq \frac{\sqrt{T}}{2\sqrt{d}}\sum_{i=1}^{d}\mathbb{P}_{\boldsymbol{\theta}^*}(i)\ ,$$

which gives the $1/\sqrt{d}$ scaling in the regret bound. The rest of the proof follows same as analyzed for **Case 1** before yielding the desired weak-regret lower bound for this case. □

C. Proofs of Regret Upper Bounds for CoLSTIM

For some fixed constants $c_1, c_2 > 0$ define the MLE concentration event

$$A_{\mathrm{MLE}} = \{\forall\{i,j\} \subset [n], t > \tau\ :\ |\hat{u}_{t,i}(\mathbf{X}_t) - u_i^*(\mathbf{X}_t) + u_j^*(\mathbf{X}_t) - \hat{u}_{t,j}(\mathbf{X}_t)| \leq c_1\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}}\}$$

and the (time-dependent) perturbed estimated utility concentration event

$$A_{\mathrm{conc},t} = \{\forall\{i,j\} \subset [n]\ :\ |\,\tilde{u}_{t,i}(\mathbf{X}_t) - \tilde{u}_{t,j}(\mathbf{X}_t) - \hat{u}_{t,i}(\mathbf{X}_t) + \hat{u}_{t,j}(\mathbf{X}_t)| \leq c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}}\}.$$

Moreover, define the initial concentration event

$$A_{\mathrm{init}} = \{\forall t > \tau\ :\ |\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*| \leq 1 \text{ and } \mathbf{M}_t \geq \boldsymbol{I}_d\}.$$

For sake of convenience, we will denote the average regret for selecting pairs (i_t, j_t) at time t by

$$\Delta_{i_t,j_t} = \frac{u_{i^*(t)}^*(\mathbf{X}_t) - u_{i_t}^*(\mathbf{X}_t) + u_{i^*(t)}^*(\mathbf{X}_t) - u_{j_t}^*(\mathbf{X}_t)}{2}.$$

C.1. Proof of Theorem 3.2 and Corollary 3.3

We use the following theorem (proven in Section C.3) to prove Theorem 3.2 as well as Corollary 3.3.

Theorem C.1. *Let $(S_t)_{t\in[T]}$ denote the selected pairs of arms by* CoLSTIM *and let $R_T^a(C) = R_T^a((S_t)_{t\in[T]})$ be its corresponding cumulative average regret. Assume there exist constants $c_1 > 0$ and $c_2 > 0$ such that for some $p_1, p_2, p_{3,t} \in [0,1]$ it holds that $\mathbb{P}(A_{\text{MLE}}) \geq 1 - p_1$, $\mathbb{P}(A_{\text{init}}) \geq 1 - p_2$, and for any given $t > \tau$ and for any possible history $\mathcal{H}_{t-1}$ before the start of round t, we have $\mathbb{P}(A_{\text{conc},t} \,|\, \mathcal{H}_{t-1}) \geq 1 - p_{3,t}$. Then, for any constant $c_3 \geq \sqrt{2\,d}$ it holds that*

$$\mathbb{E}\left[R_T^a(\text{CoLSTIM})\right] \leq \Delta_{max}\tau + \Delta_{max}\sum_{t=\tau+1}^{T} p_{3,t} + \Delta_{max}(p_1+p_2)T + \frac{c_3}{2}\Big(3\,c_1 + c_2\Big)\sqrt{T\log\left(\frac{2T}{d}\right)},$$

where $\Delta_{max} = 2\sqrt{2}$. Moreover, if $\sum_{t=\tau+1}^{T}\lambda_{\min}^{-1/2}(\mathbf{M}_t) \leq c\sqrt{T}$, where c is some positive constant and $\lambda_{\min}(A)$ denotes the smallest eigenvalue of a square matrix A, then the previous inequality holds for any constant $c_3 \geq \sqrt{2}c$.

Proof of Theorem 3.2. By the choice of the initial exploration length τ and Assumption **(A1)** we can infer that the smallest eigenvalue of the Gram matrix $\mathbf{M}_t$ is at least $\rho\tau = \max\left(1, \frac{d\log(T/d)+2\log(T)}{\mu^2}\right)$, so that following the lines of proof of Theorem 4 in (Vaswani et al., 2020) we have that $\mathbb{P}(A_{\text{init}}) \geq 1 - p_2$, holds for $p_2 = 1/T$. Similarly, by the choice of c_1 and following the lines of proof of Theorem 4 in (Vaswani et al., 2020) we have that $\mathbb{P}(A_{\text{MLE}}) \geq 1 - p_1$ holds for $p_1 = 1/T$.

It remains to derive a suitable choice for $p_{3,t}$. For this purpose, fix $t > \tau$ and $\mathcal{H}_{t-1}$ arbitrary such that we can leave out the conditioning on $\mathcal{H}_{t-1}$ in the following. With this,

$$\begin{aligned}
\mathbb{P}(A^{\complement}_{\text{conc},t}) &= \mathbb{P}(\exists\{i,j\} \subset [n] : |\,\tilde{u}_{t,i}(\mathbf{X}_t) - \tilde{u}_{t,j}(\mathbf{X}_t) - \hat{u}_{t,i}(\mathbf{X}_t) + \hat{u}_{t,j}(\mathbf{X}_t)| > c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}}) \\
&= \mathbb{P}(\exists\{i,j\} \subset [n] : |\,\tilde{u}_{t,i}(\mathbf{X}_t) - \tilde{u}_{t,j}(\mathbf{X}_t) - \hat{u}_{t,i}(\mathbf{X}_t) + \hat{u}_{t,j}(\mathbf{X}_t)| > c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}} \,|\, B_t = 1)\mathbb{P}(B_t = 1) \\
&\quad + \mathbb{P}(\exists\{i,j\} \subset [n] : |\,\tilde{u}_{t,i}(\mathbf{X}_t) - \tilde{u}_{t,j}(\mathbf{X}_t) - \hat{u}_{t,i}(\mathbf{X}_t) + \hat{u}_{t,j}(\mathbf{X}_t)| > c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}} \,|\, B_t = 0)\mathbb{P}(B_t = 0) \\
&\leq p_t + \mathbb{P}(\exists\{i,j\} \subset [n] : |\,\tilde{u}_{t,i}(\mathbf{X}_t) - \tilde{u}_{t,j}(\mathbf{X}_t) - \hat{u}_{t,i}(\mathbf{X}_t) + \hat{u}_{t,j}(\mathbf{X}_t)| > c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}} \,|\, B_t = 0) \\
&= p_t + \mathbb{P}(\exists\{i,j\} \subset [n] : |\,\epsilon_{t,i}\|\boldsymbol{x}_{t,i}\|_{\mathbf{M}_t^{-1}} - \epsilon_{t,j}\|\boldsymbol{x}_{t,j}\|_{\mathbf{M}_t^{-1}}| > c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}} \,|\, B_t = 0) \\
&\leq p_t + \mathbb{P}(\exists\{i,j\} \subset [n] : |\,\epsilon_{t,i}\|\boldsymbol{x}_{t,i}\|_{\mathbf{M}_t^{-1}} - \epsilon_{t,j}\|\boldsymbol{x}_{t,j}\|_{\mathbf{M}_t^{-1}}| > c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}} \,|\, B_t = 0)
\end{aligned}$$

Note that conditioned on $B_t = 0$, we have that all perturbation variables $(\epsilon_{t,i})_{i\in[n]}$ are the same, i.e., for each $i \in [n]$ it holds that $\epsilon_{t,i} = \epsilon = \min(C_{\text{thresh}}, \max(-C_{\text{thresh}}, \tilde{\epsilon}))$ for some $\tilde{\epsilon} \sim G$. Thus,

$$\begin{aligned}
\mathbb{P}(A^{\complement}_{\text{conc},t}) &\leq p_t + \mathbb{P}(\exists\{i,j\} \subset [n] : |\,\epsilon\,| \cdot \big|\|\boldsymbol{x}_{t,i}\|_{\mathbf{M}_t^{-1}} - \|\boldsymbol{x}_{t,j}\|_{\mathbf{M}_t^{-1}}\big| > c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}} \,|\, B_t = 0) \\
&\leq p_t + \mathbb{P}(\,|\,\epsilon\,| > c_2 \,|\, B_t = 0) \\
&= p_t,
\end{aligned}$$

where we used that by the reverse triangle inequality it holds that

$$\big|\|\boldsymbol{x}_{t,i}\|_{\mathbf{M}_t^{-1}} - \|\boldsymbol{x}_{t,j}\|_{\mathbf{M}_t^{-1}}\big| \leq \|\boldsymbol{x}_{t,i} - \boldsymbol{x}_{t,j}\|_{\mathbf{M}_t^{-1}} = \|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}}$$

for any $i, j \in [n]$ and that $|\,\epsilon\,| \leq C_{\text{thresh}} < c_2$. As a consequence, we can set

$$p_{3,t} = p_t = \min\left(1, \frac{\sqrt{2d}}{2\sqrt{t-\tau}}\Big(3\,c_1 + c_2\Big)\sqrt{\log\left(\frac{2T}{d}\right)}\right)$$

such that $\mathbb{P}(A_{\text{conc},t} \,|\, \mathcal{H}_{t-1}) \geq 1 - p_{3,t}$ and

$$\sum_{t=\tau+1}^{T} p_{3,t} \leq \frac{\sqrt{2d}}{2}\Big(3\,c_1 + c_2\Big)\sqrt{\log\left(\frac{2T}{d}\right)}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq \frac{\sqrt{2d}}{2}\Big(3\,c_1 + c_2\Big)\sqrt{T\log\left(\frac{2T}{d}\right)} = O(d\sqrt{T}\log(T)). \qquad (15)$$

Thus, using the fact that with the choices of c_1 and τ we can use $p_1 = p_2 = 1/T$, while by the choice of c_2, C_{thresh} and p_t, we can set $p_{3,t} = p_t$, we obtain the claim by virtue of Theorem C.1, by using (15) as well as $c_2 \leq c_1 = O(\sqrt{d\log(T)})$ and $\tau = o(\sqrt{dT})$. □

Proof of Corollary 3.3. Use the same choices of c_1, c_2 and $p_1, p_2, p_{3,t} \in [0,1]$ as in the proof of Theorem 3.2, but use the second statement of Theorem C.1. □

C.2. Technical Lemmas

The following lemma is an adaptation of Lemma 11 from (Abbasi-Yadkori et al., 2011) for our setting.

Lemma C.2. *Let $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_t \in \mathbb{R}^d$ be such that $\max_{s \in [t]} \|\boldsymbol{z}_s\|^2 \leq 2$ and $\mathbf{M}_{t+1} = \sum_{s=1}^t \boldsymbol{z}_s \boldsymbol{z}_s^\top$. Further, let $\tau < t$ be such that $\mathbf{M}_{\tau+1} \geq I_d$ holds, then*

$$\sum_{s=\tau+1}^{t} \sqrt{2} \wedge \|\boldsymbol{z}_s\|_{\mathbf{M}_s^{-1}} \leq \sqrt{2\, d\, t \log\left(\frac{2t}{d}\right)}.$$

Proof. Note that $x \leq 2\log(1+x)$ holds for any $x \in [0,2]$, so that

$$\sum_{s=\tau+1}^{t} (\sqrt{2} \wedge \|\boldsymbol{z}_s\|_{\mathbf{M}_s^{-1}})^2 \leq \sum_{s=\tau+1}^{t} \|\boldsymbol{z}_s\|^2_{\mathbf{M}_s^{-1}} \leq \sum_{s=\tau+1}^{t} 2\log(1 + \|\boldsymbol{z}_s\|^2_{\mathbf{M}_s^{-1}}) = 2\log\left(\prod_{s=\tau+1}^{t} 1 + \|\boldsymbol{z}_s\|^2_{\mathbf{M}_s^{-1}}\right). \tag{16}$$

Denoting by $\det(A)$ the determinant of a square matrix A, we obtain by the matrix determinant lemma

$$\begin{aligned}
\det(\mathbf{M}_{t+1}) = \det(\mathbf{M}_t + \boldsymbol{z}_t \boldsymbol{z}_t^\top) &= \det(\mathbf{M}_t)(1 + \boldsymbol{z}_t^\top \mathbf{M}_t^{-1} \boldsymbol{z}_t) \\
&= \det(\mathbf{M}_{\tau+1}) \prod_{s=\tau+1}^{t} (1 + \boldsymbol{z}_s^\top \mathbf{M}_s^{-1} \boldsymbol{z}_s) \\
&= \det(\mathbf{M}_{\tau+1}) \prod_{s=\tau+1}^{t} 1 + \|\boldsymbol{z}_s^\top\|^2_{\mathbf{M}_s^{-1}} \\
&\geq \prod_{s=\tau+1}^{t} 1 + \|\boldsymbol{z}_s^\top\|^2_{\mathbf{M}_s^{-1}},
\end{aligned} \tag{17}$$

where we used in the inequality that $\mathbf{M}_{\tau+1} \geq I_d$ holds. Further, by the determinant-trace inequality, i.e., $\det(A)^{1/d} \leq \frac{\operatorname{tr}(A)}{d}$ for any positive definite matrix $A \in \mathbb{R}^{d \times d}$, we obtain

$$d\ \det(\mathbf{M}_{t+1})^{1/d} \leq \operatorname{tr}(\mathbf{M}_{t+1}) = \operatorname{tr}\left(\sum_{s=1}^{t} \boldsymbol{z}_s \boldsymbol{z}_s^\top\right) = \sum_{s=1}^{t} \operatorname{tr}\left(\boldsymbol{z}_s \boldsymbol{z}_s^\top\right) = \sum_{s=1}^{t} \|\boldsymbol{z}_s\|^2 \leq 2t, \tag{18}$$

where $\operatorname{tr}(A)$ denotes the trace of a matrix A. Combining (16), (17) and 18 we obtain

$$\sum_{s=\tau+1}^{t} (\sqrt{2} \wedge \|\boldsymbol{z}_s\|_{\mathbf{M}_s^{-1}})^2 \leq 2\log\left(\det(\mathbf{M}_{t+1})\right) \leq 2d\log\left(\frac{2t}{d}\right).$$

Finally, by the Cauchy-Schwarz inequality we obtain

$$\sum_{s=\tau+1}^{t} \sqrt{2} \wedge \|\boldsymbol{z}_s\|_{\mathbf{M}_s^{-1}} \leq \sqrt{(t-\tau) \sum_{s=\tau+1}^{t} (\sqrt{2} \wedge \|\boldsymbol{z}_s\|_{\mathbf{M}_s^{-1}})^2} \leq \sqrt{2\, d\, t \log\left(\frac{2t}{d}\right)}.$$

□

Lemma C.3. *Let c_1, c_2 and $p_1, p_2, p_{3,t} \in [0,1]$ be as in Theorem C.1. Then, for any round $t > \tau$ and any history $\mathcal{H}_{t-1}$, we have*

$$\mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\text{MLE}} \cap A_{\text{init}}} \,|\, \mathcal{H}_{t-1}] \leq \Delta_{max} p_{3,t} + \frac{1}{2}\Big(3c_1 + c_2\Big)\mathbb{E}[\|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} 1_{A_{\text{init}}} \,|\, \mathcal{H}_{t-1}],$$

where $c_1 > 0$ and $c_2 > 0$.

Proof of Lemma C.3. Fix t and $\mathcal{H}_{t-1}$ arbitrary such that both events A_{MLE} and A_{init} in the indicator function are true. In this way, we can leave out the conditioning on $\mathcal{H}_{t-1}$ in the following. We first bound the (conditional) expected value as follows:

$$\begin{aligned}\mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\mathrm{MLE}} \cap A_{\mathrm{init}}}] &= \mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\mathrm{MLE}} \cap A_{\mathrm{init}}} 1_{A_{\mathrm{conc},t}}] + \mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\mathrm{MLE}} \cap A_{\mathrm{init}}} 1_{A^{\complement}_{\mathrm{conc},t}}] \\ &\leq \mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\mathrm{MLE}} \cap A_{\mathrm{init}}} 1_{A_{\mathrm{conc},t}}] + \Delta_{max} \mathbb{P}(A^{\complement}_{\mathrm{conc},t}) \\ &\leq \mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\mathrm{MLE}} \cap A_{\mathrm{init}}} 1_{A_{\mathrm{conc},t}}] + \Delta_{max} p_{3,t}.\end{aligned}$$

On the event A_{MLE} it holds that

$$\begin{aligned}\langle \boldsymbol{z}_{t,i_t^*,i_t}, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t \rangle &= u^*_{i_t^*}(\mathbf{X}_t) - \hat{u}_{t,i_t^*}(\mathbf{X}_t) + \hat{u}_{t,i_t}(\mathbf{X}_t) - u^*_{i_t}(\mathbf{X}_t) \\ &\leq |u^*_{i_t^*}(\mathbf{X}_t) - \hat{u}_{t,i_t^*}(\mathbf{X}_t) + \hat{u}_{t,i_t}(\mathbf{X}_t) - u^*_{i_t}(\mathbf{X}_t)| \\ &\leq c_1 \|\boldsymbol{z}_{t,i_t^*,i_t}\|_{\mathbf{M}_t^{-1}}\end{aligned} \tag{19}$$

and similarly $\langle \boldsymbol{z}_{t,i_t^*,j_t}, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t \rangle \leq c_1 \|\boldsymbol{z}_{t,i_t^*,j_t}\|_{\mathbf{M}_t^{-1}}$. Further, note that

$$\boldsymbol{z}_{t,i_t^*,j_t} = \boldsymbol{x}_{t,i_t^*} - \boldsymbol{x}_{t,j_t} = \boldsymbol{x}_{t,i_t^*} - \boldsymbol{x}_{t,i_t} + \boldsymbol{x}_{t,i_t} - \boldsymbol{x}_{t,j_t} = \boldsymbol{z}_{t,i_t^*,i_t} + \boldsymbol{z}_{t,i_t,j_t} \tag{20}$$

and by definition of j_t it holds for all $i \in [n]$ that

$$\langle \boldsymbol{z}_{t,i,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + c_1 \, \|\boldsymbol{z}_{t,i,i_t}\|_{\mathbf{M}_t^{-1}} \leq \langle \boldsymbol{z}_{t,j_t,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + c_1 \, \|\boldsymbol{z}_{t,j_t,i_t}\|_{\mathbf{M}_t^{-1}} = \langle \boldsymbol{z}_{t,j_t,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + c_1 \, \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} \tag{21}$$

Finally, by the definition of i_t it holds for all $i \in [n]$ that

$$\langle \boldsymbol{x}_{t,i}, \hat{\boldsymbol{\theta}}_t \rangle + \epsilon_{t,i} \|\boldsymbol{x}_{t,i}\|_{\mathbf{M}_t^{-1}} - \big(\langle \boldsymbol{x}_{t,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + \epsilon_{t,i_t} \|\boldsymbol{x}_{t,i_t}\|_{\mathbf{M}_t^{-1}} \big) = \langle \boldsymbol{z}_{t,i,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + \epsilon_{t,i} \|\boldsymbol{x}_{t,i}\|_{\mathbf{M}_t^{-1}} - \epsilon_{t,i_t} \|\boldsymbol{x}_{t,i_t}\|_{\mathbf{M}_t^{-1}} \leq 0 \, . \tag{22}$$

With these considerations,

$$\begin{aligned}2\Delta_{i_t,j_t} &= u^*_{i^*(t)}(\mathbf{X}_t) - u^*_{i_t}(\mathbf{X}_t) + u^*_{i^*(t)}(\mathbf{X}_t) - u^*_{j_t}(\mathbf{X}_t) \\ &= \langle \boldsymbol{z}_{t,i_t^*,i_t}, \boldsymbol{\theta}^* \rangle + \langle \boldsymbol{z}_{t,i_t^*,j_t}, \boldsymbol{\theta}^* \rangle \\ &= \langle \boldsymbol{z}_{t,i_t^*,i_t}, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t \rangle + \langle \boldsymbol{z}_{t,i_t^*,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + \langle \boldsymbol{z}_{t,i_t^*,j_t}, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t \rangle + \langle \boldsymbol{z}_{t,i_t^*,j_t}, \hat{\boldsymbol{\theta}}_t \rangle \\ &\overset{(19)}{\leq} c_1 \big(\|\boldsymbol{z}_{t,i_t^*,i_t}\|_{\mathbf{M}_t^{-1}} + \|\boldsymbol{z}_{t,i_t^*,j_t}\|_{\mathbf{M}_t^{-1}} \big) + \langle \boldsymbol{z}_{t,i_t^*,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + \langle \boldsymbol{z}_{t,i_t^*,j_t}, \hat{\boldsymbol{\theta}}_t \rangle \\ &\overset{(20)}{\leq} c_1 \big(2 \, \|\boldsymbol{z}_{t,i_t^*,i_t}\|_{\mathbf{M}_t^{-1}} + \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} \big) + 2 \langle \boldsymbol{z}_{t,i_t^*,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + \langle \boldsymbol{z}_{t,i_t,j_t}, \hat{\boldsymbol{\theta}}_t \rangle \\ &\overset{(21)}{\leq} 3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} + 2 \langle \boldsymbol{z}_{t,j_t,i_t}, \hat{\boldsymbol{\theta}}_t \rangle + \langle \boldsymbol{z}_{t,i_t,j_t}, \hat{\boldsymbol{\theta}}_t \rangle \\ &= 3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} + \langle \boldsymbol{z}_{t,j_t,i_t}, \hat{\boldsymbol{\theta}}_t \rangle \\ &\overset{(22)}{\leq} 3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} + \epsilon_{t,i_t} \|\boldsymbol{x}_{t,i_t}\|_{\mathbf{M}_t^{-1}} - \epsilon_{t,j_t} \|\boldsymbol{x}_{t,j_t}\|_{\mathbf{M}_t^{-1}} \\ &\leq 3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} + c_2 \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}},\end{aligned}$$

where the last inequality holds, since on $A_{\mathrm{conc},t}$ we have that

$$\epsilon_{t,i_t} \|\boldsymbol{x}_{t,i_t}\|_{\mathbf{M}_t^{-1}} - \epsilon_{t,j_t} \|\boldsymbol{x}_{t,j_t}\|_{\mathbf{M}_t^{-1}} \leq c_2 \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}}.$$

Thus,

$$\mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\mathrm{MLE}} \cap A_{\mathrm{init}}}] \leq \frac{1}{2} \mathbb{E} \left[\big(3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} + c_2 \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} \big) 1_{A_{\mathrm{MLE}} \cap A_{\mathrm{init}}} 1_{A_{\mathrm{conc},t}} \right] + \Delta_{max} p_{3,t},$$

from which we can conclude the claim. □

C.3. Proof of Theorem C.1

Proof of Theorem C.1. Let $\Delta_{max} = 2\sqrt{2}$ then

$$2\Delta_{i_t,j_t} = u^*_{i^*(t)}(\mathbf{X}_t) - u^*_{i_t}(\mathbf{X}_t) + u^*_{i^*(t)}(\mathbf{X}_t) - u^*_{j_t}(\mathbf{X}_t) \leq \Delta_{max}.$$

Indeed, by assumption $\max\left(\|\boldsymbol{x}_{t,i}\|^2, \|\boldsymbol{x}_{t,j}\|^2\right) \leq 1$ so that $\|\boldsymbol{z}_{t,i,j}\|^2 \leq 2$ holds for all $i, j \in [n]$. Further, by assumption $\|\boldsymbol{\theta}^*\| \leq 1$, so that by the Cauchy-Schwarz inequality

$$\begin{aligned} u^*_{i^*(t)}(\mathbf{X}_t) - u^*_{i_t}(\mathbf{X}_t) + u^*_{i^*(t)}(\mathbf{X}_t) - u^*_{j_t}(\mathbf{X}_t) &\leq |u^*_{i^*(t)}(\mathbf{X}_t) - u^*_{i_t}(\mathbf{X}_t) + u^*_{i^*(t)}(\mathbf{X}_t) - u^*_{j_t}(\mathbf{X}_t)| \\ &\leq \sqrt{\|\boldsymbol{z}_{t,i_t*,i_t}\|^2\|\boldsymbol{\theta}^*\|^2} + \sqrt{\|\boldsymbol{z}_{t,i_t*,j_t}\|^2\|\boldsymbol{\theta}^*\|^2} \\ &\leq 2\sqrt{2}. \end{aligned}$$

With this, we get

$$\begin{aligned} \mathbb{E}[R^a_T(\text{CoLSTIM})] &= \sum_{t=1}^{T} \mathbb{E}\left(\Delta_{i_t,j_t}\right) \\ &\leq \Delta_{max}\tau + \sum_{t=\tau+1}^{T} \mathbb{E}[\Delta_{i_t,j_t}] \\ &\leq \Delta_{max}\tau + \Delta_{max}T\left(\mathbb{P}\big(A^{\mathsf{C}}_{\text{init}}\big) + \mathbb{P}\big(A^{\mathsf{C}}_{\text{MLE}}\big)\right) + \sum_{t=\tau+1}^{T} \mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\text{MLE}}\cap A_{\text{init}}}] \\ &= \Delta_{max}\tau + \Delta_{max}T\left(\mathbb{P}\big(A^{\mathsf{C}}_{\text{init}}\big) + \mathbb{P}\big(A^{\mathsf{C}}_{\text{MLE}}\big)\right) + \sum_{t=\tau+1}^{T} \mathbb{E}\left[\mathbb{E}[\Delta_{i_t,j_t} 1_{A_{\text{MLE}}\cap A_{\text{init}}} \,|\, \mathcal{H}_{t-1}]\right], \end{aligned}$$

where the last equality is due to the tower property of the expected value. Using Lemma C.3, we obtain

$$\begin{aligned} \mathbb{E}[R^a_T(\text{CoLSTIM})] \leq\ & \Delta_{max}\tau + \Delta_{max}\sum_{t=1}^{T} p_{3,t} + \Delta_{max}T\left(\mathbb{P}\big(A^{\mathsf{C}}_{\text{init}}\big) + \mathbb{P}\big(A^{\mathsf{C}}_{\text{MLE}}\big)\right) \\ & + \frac{1}{2}\Big(3\,c_1 + c_2\Big)\sum\nolimits_{t=\tau+1}^{T} \mathbb{E}\left[\mathbb{E}[\|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} 1_{A_{\text{init}}} \,|\, \mathcal{H}_{t-1}]\right]. \end{aligned}$$

On A_{init} it holds for any $i, j \in [n]$ and for each time step $t > \tau$ that

$$\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}} = \sqrt{\boldsymbol{z}^\top_{t,i,j}\mathbf{M}_t^{-1}\boldsymbol{z}_{t,i,j}} \leq \sqrt{\boldsymbol{z}^\top_{t,i,j}\boldsymbol{z}_{t,i,j}} \leq \sqrt{2}.$$

since $\|\boldsymbol{z}_{t,i,j}\|^2 \leq 2$ holds by assumption (on the context vectors $(\boldsymbol{x}_{t,i})_{i\in[n]}$) and $\mathbf{M}_t \geq \boldsymbol{I}_d$ holds on A_{init}. Thus, we can use Lemma C.2 to obtain

$$\begin{aligned} \mathbb{E}[R^a_T(\text{CoLSTIM})] \leq\ & \Delta_{max}\tau + \Delta_{max}\sum_{t=\tau+1}^{T} p_{3,t} + \Delta_{max}T\left(\mathbb{P}\big(A^{\mathsf{C}}_{\text{init}}\big) + \mathbb{P}\big(A^{\mathsf{C}}_{\text{MLE}}\big)\right) \\ & + \frac{c_3}{2}\Big(3\,c_1 + c_2\Big)\sqrt{T\log\left(\frac{2T}{d}\right)}, \end{aligned}$$

for any $c_3 \geq \sqrt{2\,d}$. This lets us infer the first statement of the theorem, since $\mathbb{P}\big(A^{\mathsf{C}}_{\text{init}}\big) \leq p_2$ and $\mathbb{P}\big(A^{\mathsf{C}}_{\text{MLE}}\big) \leq p_1$ by assumption of the theorem.

For the second statement note that by the Cauchy-Schwarz inequality

$$\sum_{t=\tau+1}^{T} \sqrt{2} \wedge \|\boldsymbol{z}_{t,i_t,j_t}\|_{\mathbf{M}_t^{-1}} \leq \sum_{t=\tau+1}^{T} \sqrt{\boldsymbol{z}^\top_{t,i_t,j_t}\mathbf{M}_t^{-1}\boldsymbol{z}_{t,i_t,j_t}} \leq \sum_{t=\tau+1}^{T} \|\boldsymbol{z}_{t,i_t,j_t}\|\|\mathbf{M}_t^{-1/2}\| \leq \sqrt{2}\sum_{t=\tau+1}^{T} \lambda_{\text{min}}^{-1/2}(\mathbf{M}_t),$$

where we used the fact that the Euclidean matrix norm is consistent with the Euclidean vector norm. Thus, using the assumption on the smallest eigenvalue of $\mathbf{M}_t$, we obtain

$$\begin{aligned}\mathbb{E}[R_T^a(\text{CoLSTIM})] \le \Delta_{max}\tau + \Delta_{max}\sum_{t=\tau+1}^{T} p_{3,t} + \Delta_{max}T\left(\mathbb{P}(A_{\text{init}}^{\complement}) + \mathbb{P}(A_{\text{MLE}}^{\complement})\right) \\ + \frac{c_3}{2}\Big(3\,c_1 + c_2\Big)\sqrt{T\log\left(\frac{2T}{d}\right)},\end{aligned}$$

for any $c_3 = \sqrt{2}c$. □

D. Proof of Regret Upper Bounds for Sup-CoLSTIM

For some fixed constants $c_1, c_2 > 0$ define the MLE concentration event

$$A_{\text{MLE}}^{(S)} = \{\forall t > \tau, \forall s \in [S], \forall\{i,j\} \subset A_t^{(s)} \,:\, |\hat{u}_{t,i}^{(s)}(\mathbf{X}_t) - u_i^*(\mathbf{X}_t) + u_j^*(\mathbf{X}_t) - \hat{u}_{t,j}^{(s)}(\mathbf{X}_t)| \le c_1\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}}\}$$

and the time-dependent perturbed estimated utility concentration event

$$A_{\text{conc},t}^{(S)} = \{\forall s \in [S],\, \forall\{i,j\} \subset A_t^{(s)} :\, |\,\tilde{u}_{t,i}(\mathbf{X}_t) - \tilde{u}_{t,j}(\mathbf{X}_t) - \hat{u}_{t,i}^{(s)}(\mathbf{X}_t) + \hat{u}_{t,j}^{(s)}(\mathbf{X}_t)| \le c_2\|\boldsymbol{z}_{t,i,j}\|_{\mathbf{M}_t^{-1}}\}.$$

Moreover, define the initial concentration event

$$A_{\text{init}}^{(S)} = \{\forall t > \tau,\, \forall s \in [S] :\, |\hat{\boldsymbol{\theta}}_t^{(s)} - \boldsymbol{\theta}^*| \le 1 \text{ and } \mathbf{M}_t^{(s)} \ge \boldsymbol{I}_d\}.$$

As in Section C we denote the average regret for selecting pairs (i_t, j_t) at time t again by

$$\Delta_{i_t,j_t} = \frac{u_{i^*(t)}^*(\mathbf{X}_t) - u_{i_t}^*(\mathbf{X}_t) + u_{i^*(t)}^*(\mathbf{X}_t) - u_{j_t}^*(\mathbf{X}_t)}{2}.$$

Lemma D.1. *For all time steps $t \in [T]$ and any stages $s \in [S]$, given any realization of chosen arm-pairs $(i_t, j_t)_{t\in\Psi^{(s)}}$, the corresponding preference observations $(Y_t)_{t\in\Psi^{(s)}}$ are independent Bernoulli distributed random variables with Y_t having success probability $F^*(u_{i_t}^*(\mathbf{X}_t) - u_{j_t}^*(\mathbf{X}_t))$.*

Proof. The proof is similar to Lemma 4 in (Li et al., 2017) or Lemma 14 in (Saha, 2021). □

Lemma D.2. *Consider some time step $t \in [T]\backslash[\tau]$ and suppose that (i_t, j_t) is chosen at stage $s_t \in [S]$. Then, on the event $A_{\text{MLE}}^{(S)}$ it holds that $i_t^* \in A_t^{(s)}$ for all $s \le s_t$. Moreover, on $A_{\text{MLE}}^{(S)} \cap A_{\text{conc},t}^{(S)}$ it holds that $\Delta_{i_t,j_t} \le \begin{cases} \frac{2}{\sqrt{T}}, & \textit{if } t \in \Psi^{(0)}, \\ \frac{4}{2^{s_t}}, & \textit{else.}\end{cases}$*

Proof. The proof of the first part (i.e., $i_t^* \in A_t^{(s)}\ \forall s \le s_t$) is analogous to Lemma 6 in (Li et al., 2017) or Part-1 in Lemma 6 in (Saha, 2021), as $A_{\text{MLE}}^{(S)}$ corresponds to the set $\mathcal{E}_X$ (Li et al., 2017) or $\mathcal{E}$ (Saha, 2021).

Next, let us write for sake of brevity $s = s_t$ and let us assume that $t \in \Psi^{(0)}$, i.e., it holds that

$$w_{i,j}^{(s)}(\mathbf{X}_t) = c_1\|\boldsymbol{z}_{t,i,j}\|_{(\mathbf{M}_t^{(s)})^{-1}} \le 1/\sqrt{T}, \quad \forall i,j \in A_t^{(s_t)}. \tag{23}$$

On the event $A_{\text{MLE}}^{(S)}$ it holds that

$$\begin{aligned}\langle \boldsymbol{z}_{t,i_t^*,i_t}, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t^{(s)}\rangle &= u_{i_t^*}^*(\mathbf{X}_t) - \hat{u}_{t,i_t^*}^{(s)}(\mathbf{X}_t) + \hat{u}_{t,i_t}^{(s)}(\mathbf{X}_t) - u_{i_t}^*(\mathbf{X}_t) \\ &\le |u_{i_t^*}^*(\mathbf{X}_t) - \hat{u}_{t,i_t^*}^{(s)}(\mathbf{X}_t) + \hat{u}_{t,i_t}^{(s)}(\mathbf{X}_t) - u_{i_t}^*(\mathbf{X}_t)| \\ &\le c_1\|\boldsymbol{z}_{t,i,j}\|_{(\mathbf{M}_t^{(s)})^{-1}}\end{aligned} \tag{24}$$

and similarly $\langle \boldsymbol{z}_{t,i_t^*,j_t}, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t \rangle \le c_1 \|\boldsymbol{z}_{t,i_t^*,j_t}\|_{\mathbf{M}_t^{-1}}$, where we used that $i_t^* \in A_t^{(s)}$ for all $s \le s_t$ from the first part of the lemma. On $A_{\text{conc},t}^{(S)}$ we have that

$$\epsilon_{t,i_t} \|\boldsymbol{x}_{t,i_t}\|_{(\mathbf{M}_t^{(s_t)})^{-1}} - \epsilon_{t,j_t} \|\boldsymbol{x}_{t,j_t}\|_{(\mathbf{M}_t^{(s_t)})^{-1}} \le c_2 \|\boldsymbol{z}_{t,i_t,j_t}\|_{(\mathbf{M}_t^{(s_t)})^{-1}}. \tag{25}$$

Further, since $t \in \Psi^{(0)}$ the choice of j_t implies for all $i \in A_t^{(s_t)}$ that

$$\langle \boldsymbol{z}_{t,i,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + c_1 \|\boldsymbol{z}_{t,i,i_t}\|_{(\mathbf{M}_t^{(s_t)})^{-1}} \le \langle \boldsymbol{z}_{t,j_t,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{(\mathbf{M}_t^{(s_t)})^{-1}} \le \langle \boldsymbol{z}_{t,j_t,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + 1/\sqrt{T}, \tag{26}$$

while the choice of i_t implies for all $i \in A_t^{(s_t)}$ that

$$\begin{aligned} &\langle \boldsymbol{x}_{t,i}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \epsilon_{t,i} \|\boldsymbol{x}_{t,i}\|_{(\mathbf{M}_t^{(s_t)})^{-1}} - \big(\langle \boldsymbol{x}_{t,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \epsilon_{t,i_t} \|\boldsymbol{x}_{t,i_t}\|_{(\mathbf{M}_t^{(s_t)})^{-1}} \big) \\ &\qquad = \langle \boldsymbol{z}_{t,i,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \epsilon_{t,i} \|\boldsymbol{x}_{t,i}\|_{(\mathbf{M}_t^{(s_t)})^{-1}} - \epsilon_{t,i_t} \|\boldsymbol{x}_{t,i_t}\|_{(\mathbf{M}_t^{(s_t)})^{-1}} \le 0\,. \end{aligned} \tag{27}$$

With these considerations,

$$\begin{aligned} 2\Delta_{i_t,j_t} &= u_{i^*(t)}^*(\mathbf{X}_t) - u_{i_t}^*(\mathbf{X}_t) + u_{i^*(t)}^*(\mathbf{X}_t) - u_{j_t}^*(\mathbf{X}_t) \\ &= \langle \boldsymbol{z}_{t,i_t^*,i_t}, \boldsymbol{\theta}^* \rangle + \langle \boldsymbol{z}_{t,i_t^*,j_t}, \boldsymbol{\theta}^* \rangle \\ &= \langle \boldsymbol{z}_{t,i_t^*,i_t}, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \langle \boldsymbol{z}_{t,i_t^*,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \langle \boldsymbol{z}_{t,i_t^*,j_t}, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \langle \boldsymbol{z}_{t,i_t^*,j_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle \\ &\overset{(24)}{\le} c_1 \big(\|\boldsymbol{z}_{t,i_t^*,i_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} + \|\boldsymbol{z}_{t,i_t^*,j_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} \big) + \langle \boldsymbol{z}_{t,i_t^*,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \langle \boldsymbol{z}_{t,i_t^*,j_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle \\ &\overset{(20)}{\le} c_1 \big(2\, \|\boldsymbol{z}_{t,i_t^*,i_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} + \|\boldsymbol{z}_{t,i_t,j_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} \big) + 2\langle \boldsymbol{z}_{t,i_t^*,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \langle \boldsymbol{z}_{t,i_t,j_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle \\ &\overset{(26)}{\le} 3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} + 2\langle \boldsymbol{z}_{t,j_t,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle + \langle \boldsymbol{z}_{t,i_t,j_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle \\ &= \; 3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} + \langle \boldsymbol{z}_{t,j_t,i_t}, \hat{\boldsymbol{\theta}}_t^{(s_t)} \rangle \\ &\overset{(27)}{\le} 3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} + \epsilon_{t,i_t} \|\boldsymbol{x}_{t,i_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} - \epsilon_{t,j_t} \|\boldsymbol{x}_{t,j_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} \\ &\overset{(25)}{\le} \; 3c_1 \|\boldsymbol{z}_{t,i_t,j_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} + c_2 \|\boldsymbol{z}_{t,i_t,j_t}\|_{(\mathbf{M}_t^{(s)})^{-1}} \\ &\overset{(23)}{\le} \frac{4}{\sqrt{T}}, \end{aligned}$$

where we used in the last inequality that by choice $c_2 \le c_1$. This shows the case $t \in \Psi^{(0)}$.

Finally, showing $\Delta_{i_t,j_t} \le \frac{4}{2^{s_t}}$ if $t \notin \Psi^{(0)}$ is analogous to Part-3 of Lemma 6 in (Saha, 2021), as the choice of the pair (i_t, j_t) is the same as for Sta'D in this case[4]. Note that we do not need to condition on the event $A_{\text{conc},t}^{(S)}$ for this case.

□

Lemma D.3. *On the event $A_{\text{init}}^{(S)}$ it holds for any $s \in [S]$ that* $\sqrt{|\Psi^{(s)}|} \le c_1 2^s \sqrt{2\,d\,\log(4T^2/d)}$.

Proof. The proof is analogous to Lemma 7 in (Saha, 2021), as the event $A_{\text{init}}^{(S)}$ ensures that the smallest eigenvalue of any stage-dependent Gram matrix $\mathbf{M}_t^{(s)}$ is at least 1[5]. □

Proof of Theorem 3.4. By the choice of the initial exploration length $\tau = d + \max\{{}^{d^2 \log(T)}/{}_{\mu^2 \rho}, d/\rho\}$ and Assumption **(A1)** we can infer that the smallest eigenvalue of the Gram matrix $\mathbf{M}_t$ is at least $\rho\tau = \max\left(1, \frac{d\log(T/d)+2\log(T)}{\mu^2}\right)$, so that following the lines of proof of Lemma 4 in (Saha, 2021) we have that $\mathbb{P}\big(A_{\text{init}}^{(S)}\big) \ge 1 - 1/T$. Moreover, due to

[4] Our $\Psi^{(0)}$ corresponds to ϕ^c in (Saha, 2021).

[5] The stage-dependent Gram matrix $\mathbf{M}_t^{(s)}$ is denoted by V_t^s in (Saha, 2021).

the choice of $c_1 = \frac{3}{2\mu}\sqrt{2\log(3nT^2)}$ as well as Lemma D.1 (corresponds to Lemma 14 in (Saha, 2021)) it holds that $\mathbb{P}\big(A^{(S)}_{\text{MLE}}\big) \geq 1 - 1/T$. As in the proof of Theorem 3.2 it is straightforward to show that $\mathbb{P}(A^{(S)}_{\text{conc},t} \,|\, \mathcal{H}_{t-1}) \geq 1 - p_t$ and

$$\sum_{t=\tau+1}^{T} p_t \leq \frac{\sqrt{2d}}{2}\Big(3\,c_1 + c_2\Big)\sqrt{\log\left(\frac{2T}{d}\right)}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq \frac{\sqrt{2d}}{2}\Big(3\,c_1 + c_2\Big)\sqrt{T\log\left(\frac{2T}{d}\right)} = O(\sqrt{dT\log(n)}\log(T)). \quad (28)$$

With this, we get

$$\begin{aligned}
\mathbb{E}[R^a_T(\textsc{Sup-CoLSTIM})] &= \sum_{t=1}^{\tau}\mathbb{E}[\Delta_{i_t,j_t}] + \sum_{t\in\Psi^{(0)}}\mathbb{E}[\Delta_{i_t,j_t}] + \sum_{s=1}^{S}\sum_{t\in\Psi^{(s)}}\mathbb{E}[\Delta_{i_t,j_t}] \\
&\leq \Delta_{max}\tau + \Delta_{max}T\left(\mathbb{P}\big((A^{(S)}_{\text{init}})^{\complement}\big) + \mathbb{P}\big((A^{(S)}_{\text{MLE}})^{\complement}\big)\right) + \sum_{t=\tau+1}^{T} p_t \\
&\quad + \sum_{t\in\Psi^{(0)}}\mathbb{E}\Big[\mathbb{E}[\Delta_{i_t,j_t}\mathbb{1}_{A^{(S)}_{\text{init}}\cap A^{(S)}_{\text{MLE}}\cap A^{(S)}_{\text{conc},t}} \,|\, \mathcal{H}_{t-1}]\Big] \\
&\quad + \sum_{s=1}^{S}\sum_{t\in\Psi^{(s)}}\mathbb{E}\Big[\mathbb{E}[\Delta_{i_t,j_t}\mathbb{1}_{A^{(S)}_{\text{init}}\cap A^{(S)}_{\text{MLE}}\cap A^{(S)}_{\text{conc},t}} \,|\, \mathcal{H}_{t-1}]\Big] \\
&\leq \Delta_{max}\tau + 2\Delta_{max} + \sum_{t=\tau+1}^{T} p_t + \frac{2|\Psi^{(0)}|}{\sqrt{T}} + 4\sum_{s=1}^{S}\frac{|\Psi^{(s)}|}{2^s} && \text{(Lemma D.2)} \\
&\leq \Delta_{max}\tau + 2\Delta_{max} + \sum_{t=\tau+1}^{T} p_t + \frac{2|\Psi^{(0)}|}{\sqrt{T}} + 4c_1\sqrt{2\,d\,\log(4T^2/d)}\sum_{s=1}^{S}\sqrt{|\Psi^{(s)}|} \\
& && \text{(Lemma D.3)} \\
&\leq \Delta_{max}\tau + 2\Delta_{max} + \sum_{t=\tau+1}^{T} p_t + 2\sqrt{T} + 4c_1\sqrt{2\,d\,\log(4T^2/d)}\sqrt{T\log(T)}. \\
& && \text{(Cauchy-Schwarz inequality and } |\Psi^{(s)}| \leq T\text{)}
\end{aligned}$$

Using (28) and noting that $c_1 = O(\sqrt{\log(nT)})$ as well as $\tau = o(\sqrt{dT})$, we can conclude the proof. □

E. SGD vs. Full Maximum-Likelihood-Estimation

Figure 3 illustrates the difference in regret for the easy problem scenario $E(10, 50, G^*)$ for G^* being the standard Gumbel distribution. There is not much of a difference between the resulting regret curves for MaxInP, while the full MLE variant of CoLSTIM performs slightly better than its SGD variant. Moreover, the full MLE variant of CoLSTIM has a smaller fluctuation in the sense that its range of the standard error is smaller than of its SGD variant.

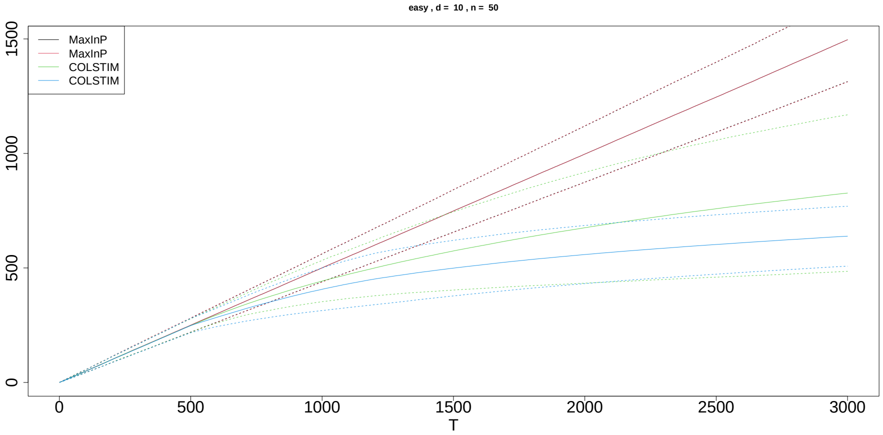

Figure 3. Averaged cumulative regret of the SGD variants of CoLSTIM and MaxInP and their full MLE variants on $E(d, n, G^*)$ for G^* being standard Gumbel.

However, regarding the average elapsed runtimes (in seconds), we see a huge difference between the SGD variants and the full MLE variants as Table E shows.

Table 2. Averaged cumulative runtimes and the corresponding standard deviations (in brackets) of the SGD variants of CoLSTIM and MaxInP and their full MLE variants on $E(d, n, G^*)$ for G^* being standard Gumbel.

	avg. runtimes (std)
MaxInP-SGD	158.00 (4.04)
MaxInP-MLE	910.13 (113.99)
CoLSTIM-SGD	7.01 (0.15)
CoLSTIM-MLE	747.04 (113.63)