CONTEXTUAL RANKING AND MATCHING. OPTIMAL REGRET UNDER LST

Hafedh El Ferchichi
CREST, ENSAE, IP Paris
Fairplay joint team

Matthieu Lerasle
CREST, ENSAE, IP Paris
Fairplay joint team

Vianney Perchet
CREST, ENSAE, Criteo AI LAB
Fairplay joint team

Abstract

We address the problem of online matchmaking with contextual information. In each round, a perfect matching between a varying set of players – with different strengths – is selected, and the outcomes of the comparisons of the chosen pairs are observed. We assume that matching players incurs dissatisfaction proportional to the "strength gap", thereby incentivising the pairing of players with closely matched strengths. Additionally, we assume that the strength of each player can be inferred from some available contextual information through the contextualised linear stochastic transitivity model **(LST)**. We propose an algorithm that performs matchmaking by selecting pairs of maximum informativeness among admissible pairs and prove that its regret is optimal up to logarithmic factors.

1 INTRODUCTION

Background and Motivation: *Learning-to-Rank* is a fundamental problem in machine learning with wide-ranging applications, including information retrieval, recommendation systems, and matchmaking for competitive gaming platforms (Cao et al., 2007; Bengs et al., 2021; Guo et al., 2020; Morik et al., 2020). A particularly important class of learning-to-rank problems is based on *pairwise comparisons*, where the relative preferences or strengths of players or items are inferred from direct interactions. These comparisons often exhibit inherent randomness. For example, in games such as chess, where Elo scores are often used (Elo, 1978), even if two players compete repeatedly, the results may vary due to psychological, environmental, or more generally uncontrolled stochastic factors. In particular, when the "skill gap" between players is small, the probability of winning approaches 1/2, as reflected in models such as Elo (Elo, 1978; Herbrich et al., 2006; Minka et al., 2018). This stochasticity makes the ranking both challenging and realistic.

Given these considerations, ranking from noisy comparisons has attracted substantial interest in recent years from both the theoretical and applied communities (Minka et al., 2018; Bengs et al., 2021; Gong et al., 2020; El Ferchichi et al., 2024b). It underpins key systems such as Microsoft's TrueSkill (Herbrich et al., 2006; Minka et al., 2018), which uses the track record of each player to infer rankings and proposes future matchups on large-scale gaming platforms.

A significant challenge lies in reconciling the dual objectives of ranking accurately and generating engaging matches, two complementary but often conflicting goals. On the one hand, the ranking process requires the exploration of a wide range of comparisons to improve statistical efficiency. Pairing two players in isolation, for example, offers limited insight into their relative standing within the broader population. However, such exploration can lead to matchups between players with markedly different skill levels. Although these unbalanced comparisons may be informative for ranking, they often lack engagement value for the participants, as the results become highly predictable, with one player consistently outperforming the other. Recognizing this trade-off, platforms strive to maintain accurate rankings while simultaneously generating matchups between players of comparable skill, balancing informativeness with user engagement (Minka et al., 2018). These considerations have been echoed in the framework of "active ranking" (Falahatgar et al., 2017b,a; Zoghi et al., 2017; Szörényi et al., 2015; Saha and Gopalan, 2019; Ren et al., 2019; El Ferchichi et al., 2024a) and more broadly in the framework of dueling/preference-based bandits (Yue et al., 2012; Bengs et al., 2021; Saha,

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

2021).

However, a common limitation among active ranking algorithms (Falahatgar et al., 2017a; Zoghi et al., 2017; Szörényi et al., 2015; Ren et al., 2019; El Ferchichi et al., 2024a) is their exclusive reliance on pairwise comparison outcomes, without the ability to incorporate contextual information that could facilitate the ranking process. This constraint manifests itself in two significant ways: (1) These methods presuppose the continuous availability of all players, as this is a requirement of the sorting algorithms on which they rely. (2) A further limitation is the assumption of static skill levels, even though players generally improve over time as they accumulate experience. To address these limitations, we adopt the framework of *Linear Stochastic Transitivity* **(LST)** models, which encompasses several well-established models such as Bradley-Terry-Luce **(BTL)** (Bradley and Terry, 1952; Hunter, 2004) or Thurston-Mosteller model (Henery, 1992; Thurstone, 1994). These are notably used in ranking systems like Elo (Elo, 1978) and TrueSkill (Minka et al., 2018; Herbrich et al., 2006). **(LST)** models characterize the stochastic process underlying the outcome of a pairwise comparison between players by assuming that each one possesses a latent skill value and that comparisons are biased in favor of the player with the larger skill value.

Another notable gap in the literature is the absence of ranking formulations within the contextual dueling bandit framework (Dudík et al., 2015; Saha, 2021; Bengs et al., 2022). Existing algorithms in this domain focus exclusively on identifying the best arm—albeit under various solution concepts—without addressing the broader ranking problem. Moreover, they typically select only one pair of players per round, rather than a full matching. This limitation is particularly problematic in settings such as online video games, where neglecting players for several consecutive rounds can lead to disengagement (Minka et al., 2018). To mitigate this issue, we impose the constraint that, at each round t, all k_t players connected to the platform must be matched. For simplicity, we assume that k_t is even. Concretely, the algorithm selects a perfect matching over the k_t active players—leveraging both their contextual features and past interaction data—observes the outcomes of all duels and then proceeds to the next round.

Assuming that players generally prefer compelling duels, against opponents with comparable skill levels, we presume that the quality of a match peaks if opponents have the same skill levels, incurring a cost of 0; this cost increases linearly with the skill gap. Consequently, the cost of a matching is the sum of the costs associated with its pairs.

Organization of the paper: Section 2 formalizes the model and summarizes the contributions, while Section 3 reviews prior work.

Section 4 introduces Maximum-Informative Matching (MIM), an adaptation of **MaxInP** (Saha, 2021), and provides regret guarantees that are $\tilde{O}(\sqrt{d})$ from optimal in some regimes.

Section 5 presents an optimal (up to logarithmic factors) algorithm, Sup-Matchmaking, together with a matching lower bound. The approach builds on (Auer, 2002) and highlights why these ideas extend to our setting but not to broader combinatorial problems.

Proofs and additional details are deferred to the appendix.

2 PROBLEM SETTING AND NOTATIONS

Notations: For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we refer to its ℓ^2-norm as $||\boldsymbol{x}||$, its transpose as $\boldsymbol{x}^\intercal$. Its weighted ℓ^2 norm, given by a positive definite matrix $\boldsymbol{A}$ is denoted by $||\boldsymbol{x}||_{\boldsymbol{A}} = \sqrt{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle}$. For convenience, all the notations used throughout this paper are summarized in the Appendix A.

Model: Consider a scenario where at each iteration $t \in \{1, \dots, T\}$, P_t players/items are available (for simplicity, P_t is assumed to be even), each associated with an unknown ranking denoted $(r(1), \dots, r(P_t))$. In each round t, the algorithm picks a perfect matching M_t, which is a partition of the available players divided into pairs, that is, $M_t = \{\{a_{t,1}, a_{t,2}\}, \dots, \{a_{t,P_t-1}, a_{t,P_t}\}\}$ where $\{a_{t,1}, \dots, a_{t,P_t}\} = \{1, 2, \dots, P_t\}$. Hereafter, a "perfect matching" will simply be referred to as a "matching", with size $k_t := P_t/2$. When two players/items i and j are paired, the result "i beats j" is observed with probability $p(i, j)$ and "j beats i" is observed with probability $p(j, i) = 1 - p(i, j)$. Note that this implicitly disregards the possibility of ties. These probabilities are assumed to satisfy the *Linear Stochastic Transitivity (LST)* model (Szörényi et al., 2015; Bengs et al., 2021): Each player i is assigned a latent score or strength $\mu_i \in \mathbb{R}$, and the outcome of a comparison between players i and j is modeled as a Bernoulli random variable with success probability

$$p(i, j) = F(\mu_i - \mu_j), \tag{1}$$

where $F : \mathbb{R} \to [0, 1]$ is an increasing *comparison function* satisfying $F(-x) = 1 - F(x)$, and $F(\infty) = 1$. The players are thus ranked by their strength; the rank of i is smaller than the rank of j - $r(i) < r(j)$ - whenever i is stronger than j - $\mu_i > \mu_j$. The results of these comparisons are assumed to be independent (Bengs et al.,

2021; Feige et al., 1994; Falahatgar et al., 2017a,b; Saha and Gopalan, 2019; Heckel et al., 2019).

Example of Parametric Models: Bradley-Terry-Luce: Among the popular choices is the Bradley-Terry-Luce model (**BTL**) (Bradley and Terry, 1952; Chetrite et al., 2017; Chen et al., 2022; Gao et al., 2023), which is a well-established parametric model for defining $p(i, j)$, also related to the widely used Elo ranking system (Elo, 1978). In this model, the comparison function is $F(x) = \frac{1}{1+e^{-x}}$.

These models capture the intuition that players with similar skill levels produce uncertain outcomes, while matches between players with very different skills are highly predictable. It also implies important structuring properties such as transitivity and locality (Bengs et al., 2021). We proceed under the assumption that this function F is known.

Linear Contextual Models (CoLST): To model the additional available information about each player, the *contextualised LST* **(CoLST)** model assumes that each player i is associated at time t with an observed time dependent feature vector $x_{t,i} \in \mathbb{R}^d$, and that their latent strength is a linear function of the form:

$$\mu_{t,i} = \langle \boldsymbol{\theta}^*, x_{t,i} \rangle, \tag{2}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is an unknown parameter vector to be estimated (Schäfer and Hüllermeier, 2018; Saha, 2021; Bengs et al., 2022). In this model, the pairwise comparison probability becomes:

$$p_t(i, j) = F\left(\langle \boldsymbol{\theta}^*, x_{t,i} - x_{t,j} \rangle\right). \tag{3}$$

In the following, let $z_{t,i,j} = x_{t,i} - x_{t,j}$ denote the contrast vector between i and j and at time t and let $c_t(i, j) = \langle \boldsymbol{\theta}^*, z_{t,i,j} \rangle$ denote the "skill gap" between i and j at time t. Specifically, the signs of the skill gaps give the underlying ranking of the players.

Matching Cost: As previously mentioned, efficient matchmaking in video games serves as a compelling example (Minka et al., 2018; Herbrich et al., 2006). We consider the problem of designing a dynamic matchmaking system that selects a matching M_t at each round. Mismatches—where the skill gap is perceived as large—result in player dissatisfaction, arising from games that are either trivially easy or excessively challenging. Hence, not all matchings are equally desirable. Platforms aim to maximize player engagement by proposing matches between players of comparable skills. To model this, we assume that the dissatisfaction or cost incurred by matching i and j is linear in terms of the skill-gap $c_t(i, j)$. Hence, the cost of a matching M is the sum of the costs of each pair in that matching, i.e., $C_t(M) = \sum_{\{i,j\} \in M} |c_t(i, j)|$. This choice of loss links the optimal matching problem with ranking: Once the ranking is recovered, the optimal match is obtained by matching the first player with the second one, the third with the fourth, etc. (See Lemma 4.1).

Problem Statement and Objective: At each round t, k_t players are available, each with their own feature vector $\boldsymbol{x}_{t,i}$. The algorithm chooses a matching M_t and observes the results of the induced comparisons. The goal is to minimize the cumulative cost, that is $\sum_{t=1}^{T} C_t(M_t)$ or, equivalently, the cumulative regret:

$$R_T = \sum_{t=1}^{T} C_t(M_t) - C_t(M_t^*) \tag{4}$$

where $M_t^* = \arg\min_{M \text{ matching}} C_t(M)$ is the matching with minimal cost at time t.

Main Results: Our contributions are threefold:

- We propose two algorithms, Maximum-Informative Matching **(MIM)** 1 and Sup-Matchmaking 2, to adaptively solve the matchmaking problem while minimizing cumulative regret.
  - **(MIM)** achieves a worst-case regret of $\tilde{O}(d\sqrt{K} + d^{3/2}N)$[1], where $K = \sum_{t=1}^{T} k_t$ is the total number of observed pairs, which is optimal for very large values of K, and $N = \max_t k_t$ is the maximum number of pairs at each round.
  - Sup-Matchmaking achieves regret bounds of $\tilde{O}(\sqrt{dK} + dN)$, which is optimal (up to logarithmic factors). It theoretically outperforms **(MIM)** as long as $\log K \leq d$. This algorithm generalizes ideas from **(SupLinRL / SupLinUCB)** (Auer, 2002; Chu et al., 2011) to a combinatorial setting, offering insight into what sort of contextual combinatorial bandit problems these algorithms would generalize to.
- We provide a lower bound on cumulative regret under the **(CoLST)** models, certifying the optimality of Sup-Matchmaking (up to logarithmic terms).

3 RELATED WORKS

Parallel Sorting algorithms: Parallel comparison-based sorting algorithms are attractive in our setting given the intrinsic connection between matching and ranking; see Lemma 4.1). Numerous methods have

[1] $\tilde{O}$ hides logarithmic terms.

been developed to take advantage of parallelism for sorting under different computation models and constraints (Knuth, 1998; Valiant, 1975). Among these algorithms, the "sorting networks" (Batcher, 1968; Ajtai et al., 1983; Chvátal, 1992) stand out because they share the property that no element participates in multiple comparisons in the same round. This is relevant to our setting as players can only play one game at a time. Multiple sorting networks exist for multiple settings, in particular to take into account possible faulty comparisons, where the strongest element does not necessarily prevail (Assaf and Upfal, 1990; Feige et al., 1994; Leighton et al., 1997; El Ferchichi et al., 2024a). However, these algorithms suffer from $\Omega(N \log T/\Delta)$ asymptotic gap-dependent lower bound (Ren et al., 2019). This is problematic, as N is prohibitively large in practice (Minka et al., 2018). When additional information is available and can be effectively leveraged via a generalization model to (approximately) infer the skill levels, sorting algorithms are less interesting, as they solely rely on comparison outcomes.

Contextual Combinatorial Bandits: The matching problem with available contexts can be seen as an example of contextual combinatorial semi-bandits (Wen et al., 2015; Takemura et al., 2021). The combinatorial structure is clear as, for N players, there are $\frac{N!}{2^{N/2}(N/2)!} \simeq (\frac{N}{e})^{N/2}$ possible matchings. However, the feedback is quite different here: In the contextual combinatorial bandit problem, better choices made by the algorithm yield more informative samples on θ^*. This is analogous to the discrepancy between dueling bandits (Bengs et al., 2021) and standard multi-armed bandits (Lattimore and Szepesvári, 2020) and ultimately the contextual versions of these two problems (Bengs et al., 2021; Saha, 2021; Bengs et al., 2022)

Dueling Bandits: The dueling bandits framework (Bengs et al., 2021) extends the classical multi-armed bandit problem to settings where actions are selected in pairs and only relative feedback between pairs of players is observed, rather than absolute reward. In each round, the algorithm selects a pair of arms and observes which is preferred, typically under stochastic (Bengs et al., 2022) or adversarial noise models (Saha et al., 2021). In this framework, a variety of objectives have been considered, ranging from best-arm related objectives (differing mainly through the optimality concept, including Borda winner(Saha et al., 2021) Copeland winner (Zoghi et al., 2015), etc.) to the top k ranking (Braverman et al., 2016; Mohajer et al., 2017) and the total ranking (Ailon et al., 2008; Ren et al., 2019; El Ferchichi et al., 2024a). For an extensive review of the subject, we refer to (Bengs et al., 2021). Perhaps closest to our work are (Bengs et al., 2022; Saha, 2021), where a dueling bandit framework is considered with the notable addition of contexts, following the Linear Stochastic Transitivity **(LST)** model. In those works, the goal is to sequentially pick pairs containing the best arm, while minimizing the regret. Algorithms to minimize regret are provided (**(SupCoLSTIM / ColSTIM)** by (Bengs et al., 2022) and **(Sta'D / MaxInP)** by (Saha, 2021)) and subsequently analyzed, showing their optimality. In many aspects, their work generalizes that of (Li et al., 2017), notably the algorithms **(SupCB-GLM / UBC-GLM)**, to the setting of dueling bandits. Similarly, we build on the ideas of (Auer, 2002; Chu et al., 2011; Li et al., 2017; Saha, 2021; Bengs et al., 2022) to design efficient algorithms for the matchmaking problem. In particular, this paper is the first work that extends the ideas of **(SupCB-GLM)** to a combinatorial setting.

4 MAXIMUM-INFORMATIVE MATCHING (MIM)

4.1 Link between Ranking and Matching

In this section, we start by highlighting the link between ranking players and eliciting optimal matching, as attested by the following lemma (see the proof in Appendix C):

Lemma 4.1. *If $\mu_1 > \mu_2 > \ldots > \mu_k$, then the optimal matching is*

$$M = \{\{1,2\},\{3,4\},\ldots,\{k-1,k\}\}. \tag{5}$$

Lemma 4.1 shows that once the ranking is known, the optimal matching is obtained by pairing adjacent players. When the ranking is uncertain, this translates into uncertainty over which pairs may belong to the optimal matching. In particular, if two players i and j are statistically indistinguishable, they may occupy adjacent positions in some plausible ranking and could therefore be matched together. However, adjacency in the ranking does not necessarily correspond to small gaps. For example, consider $\mu_1 > \mu_2 > \mu_3 > \ldots > \mu_k$ with $\Delta_{1,2} \gg \Delta_{2,3}$: players 1 and 2 remain adjacent and must be matched in the optimal solution, despite being easily distinguishable. More generally, if $r(i)$ is even, then i is paired with the weakest player stronger than itself. This observation suggests that identifying the optimal matching under uncertainty amounts to determining which players may be adjacent in the ranking, which is precisely the structure exploited by MIM This applies symmetrically for uneven ranks. This rationale is reflected in the lines 16 to 26 of **MIM** 1. These observations are the guiding principles for the design of **(MIM)** 1.

4.2 Maximally Informative Matching (MIM)

In this section, we introduce Algorithm 1, termed Maximum-Informative Matching **(MIM)**. **(MIM)** is an intuitive algorithm for contextual matchmaking, analogous in spirit to CombLinUCB (Wen et al., 2015) and C^2 UBC (Qin et al., 2014), but adapted to the setting considered in this work. The basic idea behind **(MIM)** is to maintain an estimate $\hat{\boldsymbol{\theta}}_t$ and a confidence set for the true parameter $\boldsymbol{\theta}^*$. These constructions are based on maximum likelihood estimation **(MLE)** techniques for generalized linear models **(GLM)**, which are standard in the linear bandit literature (Bengs et al., 2022; Li et al., 2017; Filippi et al., 2010). Then, these estimates are used in a way that deviates from CombLinUCB and C^{2}UCB. Instead of considering optimistic estimates of costs for each pair and then constructing a minimal cost matching according to those estimates, **(MIM)** selects a matching M_t that is "*maximally informative*", i.e. maximizes $\sum_{\{i,j\}\in M_t} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}$, where the design matrix $\boldsymbol{\Sigma_t} = \sum_{s=1}^{t}\sum_{\{i,j\}\in M_s} \boldsymbol{z}_{s,i,j}\boldsymbol{z}_{s,i,j}$. The main difference is that the original algorithms equate all non-distinguishable pairs, since their optimistic cost is 0, while **(MIM)** identifies pairs that are *optimistically* in the optimal matching, then construct a matching using these pairs. This favors exploration and so the quality of the estimate $\hat{\boldsymbol{\theta}}_t$.

Notice that adapting CombLinUCB or C^{2}UCB is not straightforward as these are designed for standard combinatorial bandit problems where the signal increases as the quality of the decisions improves, whereas our setting could be seen as an instance of "*combinatorial dueling bandits*". It is well known that algorithms from standard bandits require non-trivial adaptations for them to work in the dueling bandit setting (Saha, 2021; Bengs et al., 2022, 2021; Yue et al., 2012).

A separate but important characteristic of **(MIM)** is that it allows the size of the available players k_t to vary over time. However, the maximum number of players available in each round $N = \max\{k_t, t \in [T]\}$ is required. This is not a significant limitation given the motivating example of video game matchmaking, as platforms have finite and known physical restrictions on the maximum number of simultaneously connected players.

Main Idea Behind MIM: Algorithm 1 starts with a number of exploration rounds τ. τ is calibrated so that the estimation of the costs of the pairs is precise with high probability. Then, instead of acting greedily on the estimates of the costs, we devise a strategy that relies on a simple observation: if for a given pair $\{i, j\}$, it is not possible to decide which has a higher skill level, then that pair could potentially be in the optimal matching (Line 10, Algorithm 1). However, some pairs are easier to distinguish than others. Since an easily distinguishable pair could still be in the optimal matching, the rule (line 10, Algorithm 1) will ultimately eliminate it, so, through (Line 16 to 27, Algorithm 1) the algorithm retrieves the edges that potentially are in this situation. Later, Algorithm 1 picks the maximally informative matching, i.e., $\arg\max \sum_{\{i,j\}\in\mathcal{M}} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}$.

4.3 Theoretical Analysis

To analyze the performance of Algorithm 1, we make the following mild technical assumptions:

Technical Assumptions

Assumption 4.2. *$\|\boldsymbol{x}_{t,i}\| \le 1$ for all $t \le T$ and $i \in [k_t]$. There exists a basis $\{\boldsymbol{e}_i\}_{i=1}^d \subset \{\boldsymbol{x}_{t,i}\}_{i=1}^{k_t}$, for all t, such that $\rho I_d \preceq \sum_{i,j=1}^d \boldsymbol{e}_i\boldsymbol{e}_i^\intercal$ for some $\rho > 0$, that is, the feature vectors span $\mathbb{R}^d$.*

This assumption is commonplace in the literature on contextual bandits (Filippi et al., 2010; Li et al., 2017; Saha, 2021; Bengs et al., 2022). It is only needed in the initialization phases of Algorithms 1 and 2: $M_1, \dots, M_\tau$ are chosen to ensure that the design matrix $\boldsymbol{\Sigma_t}$ is invertible for all t in Algorithm 1 and $\boldsymbol{\Sigma_t^{(s)}}$ is invertible for all t, s in Algorithm 2 ($\boldsymbol{\Sigma_t^{(s)}}$ is the design matrix of stage s; see A for definition). The remainder of the analysis does not rely on Assumption 4.2. An alternative strategy is to use regularization (Abbasi-Yadkori et al., 2011; Takemura et al., 2021) and initialize $\boldsymbol{\Sigma}_0 = \lambda I_d$.

Assumption 4.3 (Λ-smoothness)**.** *The derivative of F, denoted by F', is Λ-Lipschitz for some $\Lambda > 0$. Moreover,*

$$\kappa := \inf\{F'(\langle\hat{\theta}, x\rangle); \|x\|_2 \le 1, \|\hat{\theta} - \theta^*\| < 1\} > 0 \quad (6)$$

where θ^* is the true unknown weight vector.

Notice that Assumption 4.3 implies that F is L-Lipschitz for some $L \le \kappa + \Lambda$. This assumption is also standard in the literature on generalized linear bandits (Li et al., 2017; Saha, 2021; Bengs et al., 2022). It is notably satisfied by the logistic function σ, corresponding to the **(BTL)** model, for $\Lambda = \frac{1}{4}$ and $\kappa = \sigma(1)(1 - \sigma(1))$.

Let

$$K = \sum_{t=1}^{T} k_t$$
$$\tau = \max\{(d\log(T) + 2\log(T)/\kappa^2\rho, 1/\rho\}$$
$$c = \frac{1}{2\kappa}\sqrt{d\log(NT/d) + 2\log(NT)}$$

The following theorem quantifies the performance of **(MIM)**.

Theorem 4.4. *The regret incurred by 1 is upper-bounded by*

$$\mathbb{E}\left[R_T\right] = O\left(\frac{dL}{\kappa^2}\sqrt{K}\log(NT) + d^{\frac{3}{2}}N\log\left(\frac{NT}{d}\right)\right) \quad (7)$$

Ideas of Proof: The proof relies mainly on results on maximum likelihood estimation **(MLE)** for generalized linear bandits (Li et al., 2017), to show that the choices of τ and c ensure such that the estimates $\langle \boldsymbol{z}_{t,i,j}, \hat{\boldsymbol{\theta}}_t\rangle$ and the *true* cost $\langle \boldsymbol{z}_{t,i,j}, \boldsymbol{\theta}^*\rangle$ are "close". It also relies on the following adaptation of the self-normalized bound (Abbasi-Yadkori et al., 2011; Takemura et al., 2021) to account for the variability of k_t given in Lemma 4.5. The proof of Lemma 4.5 is postponed to Appendix E. The detailed proof of Theorem 4.4 is postponed to Appendix D.

Lemma 4.5. *Let* $\{\boldsymbol{z}_{t,i,j}\}_{t\in[\tau+1,\dots,T],\{i,j\}\in M_t}$ *such that* $\|\boldsymbol{z}_{i,j}\| \leq 1$ *and* $\boldsymbol{\Sigma}_t = \sum_{s=1}^{t-1}\sum_{\{i,j\}\in M_t}\boldsymbol{z}_{s,i,j}\boldsymbol{z}_{s,i,j}^\intercal$ *such that* $\Sigma_{\tau+1} \succeq I_d$. *Let* $\tilde{k}_t = \#M_t$ *and* $\tilde{K} = \sum_{t=\tau+1}^{T}\tilde{k}_t$ *Then*

$$\sum_{t=\tau}^{T}\sum_{\{i,j\}\in M_t}\min\left(\frac{1}{\sqrt{\tilde{k}_t}}, \|\boldsymbol{z}_{t,i,j}\|_{\Sigma_t^{-1}}\right) \leq \sqrt{2d\tilde{K}\log\frac{2\tilde{K}}{d}}. \quad (8)$$

Discussion of the Result: This result provides an adaptation of ideas from contextual linear bandits to the setting of matchmaking. Two conclusions may be drawn from the regret bound: **(1)** It does not matter (asymptotically) how k_t are distributed, as long as $N = O(\sqrt{K})$. Below that point, the regret behaves as if the pairs were selected one at a time, as we get the same regret bounds as (Saha, 2021; Bengs et al., 2021). **(2)** Similar to the GLM-Bandits case (Li et al., 2017), we expect this bound to be optimal (up to logarithmic terms) for very large values of K, as it would match the minimax lower bound for contextual bandits problems where the action set is the unit ℓ^2-ball (Dani et al., 2008). However, this bound is known to be suboptimal for small values of K, since a bound of $\tilde{O}(\sqrt{dK})$ is achievable for stochastic contextual bandits in general (Auer, 2002; Chu et al., 2011; Li et al., 2017), contextual dueling bandits in particular (Saha, 2021; Bengs et al., 2022).

In the next section, we extend the ideas of (Auer, 2002) to obtain an algorithm that achieves a sublinear regret bound in terms of the dimension, of order $O(\sqrt{dK}(\log NT)^{3/2})$, at the price of an additional logarithmic term.

Algorithm 1 MAXIMUM-INFORMATIVE MATCHING **(MIM)**

Input: $\tau > 0$, $c > 0$;
Initialization: Pick τ Matchings $M_1, \dots, M_\tau$, observe the outcome $\{Y_{t,\{i,j\}}, \{i,j\} \in M_t\}$, $\forall t \in [\tau]$;
Set $\boldsymbol{\Sigma}_{\tau+1} = \sum_{t=1}^{\tau}\sum_{\{i,j\}\in M_t}\boldsymbol{z}_{t,i,j}\boldsymbol{z}_{t,i,j}$;
for $t = \tau+1, 2, \dots, T$ **do**
  Set $\boldsymbol{E}_t = \boldsymbol{0}_{k_t}$, $\boldsymbol{W}_t = \boldsymbol{0}_{k_t}$, $\boldsymbol{O} = 0_{k_t}$;
  Compute MLE $\hat{\boldsymbol{\theta}}_t$;
  Set $\boldsymbol{W}_{t,i,j} = ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}$;
  for $i, j \in [k_t]$ **do**
    if $|\langle\hat{\boldsymbol{\theta}}_t, \boldsymbol{z}_{t,i,j}\rangle| \leq c\boldsymbol{W}_{t,i,j}$ **then**
      Set $\boldsymbol{E}_{t,i,j}, \boldsymbol{E}_{t,j,i} = 1, 1$;
    else
      Set $\boldsymbol{O}_{t,i,j} = \text{sign}(\langle\hat{\boldsymbol{\theta}}, \boldsymbol{z}_{t,i,j}\rangle)$;
      Set $\boldsymbol{O}_{t,j,i} = \text{sign}(\langle\hat{\boldsymbol{\theta}}, \boldsymbol{z}_{t,j,i}\rangle)$;
    end if
  end for
  Set $V_t = \emptyset$;
  for $i \in [k_t]$ **do**
    if $\boldsymbol{E}_{t,i,j} = 0$, $\forall j \in [k_t]$ **then**
      $r_i = \text{RANK}(i)$;
      $V_t = V_t \cup \{i\}$;
      if r_i is even **then**
        $\text{LN}(i, V_t, \boldsymbol{E}_t, \boldsymbol{O}_t)$;
      else
        $\text{RN}(i, V_t, \boldsymbol{E}_t, \boldsymbol{O}_t)$;
      end if
    end if
  end for
  $M_t = \text{PICKMATCHINGMIM}(V_t, \boldsymbol{E}_t, \boldsymbol{W}_t)$;
  $\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t + \sum_{\{i,j\}\in M_t}\boldsymbol{z}_{t,i,j}\boldsymbol{z}_{t,i,j}^\intercal$;
end for

5 SUP-MATCHMAKING

5.1 High-level Intuitions behind SUP-MATCHMAKING

(MIM) (Algorithm 1) is of order $\sqrt{d}$ shy of the lower bound proved in Theorem 5.3. In classical stochastic linear bandit, asymptotically optimal algorithms (up to logarithmic terms) have been developed successfully (Auer, 2002; Li et al., 2017; Saha, 2021; Bengs et al., 2022). However, generalizing these algorithms to generic combinatorial semi-bandits has remained elusive due to dependency issue in the observations, which we have been able to circumvent. The goal of SUP-MATCHMAKING 2 is to generalize **(SupLinUCB)** (Li et al., 2017) to the problem of matching. Essentially, SUP-MATCHMAKING 2 outperforms **(MIM)** 1 because it relies on stronger concentration guarantees of the estimated skill gaps, see Lemma E.2. This is achieved by building phases as in (Auer and Ortner,

2010; Lattimore and Szepesvári, 2020). Adapting this construction to contextual (generalized) linear bandits is far from trivial.

Algorithm Sup-Matchmaking 2 would best be described as multi-scale filtering: The algorithm maintains distinct collections of samples $(\Psi^{(s)})_{s=1,2,\dots}$, such that for each scale s, the samples stored in $\Psi^{(s)}$ depend only on the previous phases $s-1, s-2, \dots$, making them independent conditionally on $\Psi^{(s-1)}, \Psi^{(s-2)}, \dots, \Psi^{(1)}$. These samples are used to build sequentially partial matchings $M_t^{(s)}$, using only the remaining pairs that are in phase s, so the choice of $M_t^{(s)}$ is only affected by $M_t^{(s-1)}, M_t^{(s-2)}, \dots, M_t^{(1)}$. The subroutine PickMatchingSM 7 aggregates $M_t^{(1)}, M_t^{(2)}, \dots, M_t^{(S)}$ and, to build a perfect matching, adds $M_t^{(S+1)}$ by matching all the elements left out, making sure that $M_t^{(1)} \cup M_t^{(2)} \cup \dots \cup M_t^{(S)} \cup M_t^{(S+1)}$ is a perfect match. Theorem C.3 shows that $M_t^{(S+1)}$ is not an arbitrarily bad matching, showing that errors made by choosing a pair of cost c would only deteriorate the remaining possibilities by $\frac{L}{\kappa}c$. We postponed the statement and proof of Theorem C.3 to Appendix C to comply with the space limit.

5.2 Theoretical Analysis

Let

$$\tau = \max\{(d\log(T) + 2\log T)/\kappa^2\rho,\ \ 1/\rho\}$$
$$c = \frac{3}{2\mu}\sqrt{\log(3NT)}.$$

The following theorem quantifies the performance of Sup-Matchmaking 2.

Theorem 5.1. *The regret incurred by Algorithm 2 is upper-bounded by*

$$\mathbb{E}[R_T] = O\left(\frac{L}{\kappa^2}\left[\sqrt{dK}(\log NT)^{\frac{3}{2}} + dN(\log NT)^{\frac{5}{2}}\right]\right) \quad (9)$$

Ideas of Proof: Sup-Matchmaking 2 starts with the same initialization as **(MIM)** (Algorithm 1). The improvement $\sqrt{d}$ with respect to **(MIM)** comes from the choice of a confidence parameter c that is $O(\sqrt{d})$ smaller. This is possible because the samples inside a phase $\Psi^{(s)}$ are "independent", as observed in the seminal work (Auer, 2002). The extension of this central result in our setting is achieved through PickMatchingSM (Algorithm 7), as shown in the following lemma, whose proof is postponed to E

Lemma 5.2. *For all time steps $t \in [T]$ and any stages $s \in [S]$, $\Psi^{(s)}(t)$ the corresponding preference observations $(Y_{(t,(i,j))})_{(t,(i,j))\in\Psi^{(s)}(t)}$ are independent Bernoulli-distributed random variables with $Y_{(t,(i,j))}$ having the probability of success $F(\langle \boldsymbol{z}_{t,i,j}, \boldsymbol{\theta}^*\rangle)$.*

It remains to show that the regret incurred by the matching picked by function PickMatching is not "too costly". In fact, we show that the regret depends linearly on the estimation error made on the chosen pairs (Lemma C.3). This is detailed in Appendix C. The remaining steps of the proof are close to those of **(SupCB-GLM)** (Li et al., 2017) and are postponed to Appendix E.

Discussion: The upper bound on the regret of Algorithm 2 indicates, as for Algorithm 1, that the performance does not deteriorate (asymptotically) for $N = O(\sqrt{K/d})$. Within that boundary, the algorithm 2 is also oblivious to the distribution of k_t. Note that the logarithmic term suggests that algorithm 2 would only outperform algorithm 1 as long as $\log K = O(d)$, which is expected, as for infinite action sets, a linear scale of regret in terms of dimension d is inevitable (Dani et al., 2008).

Lower Bound: To certify the optimality of Sup-Matchmaking, we also provide a worst-case lower bound for the regret of any algorithm in our setting. The proof of Theorem 5.3 is postponed to Appendix F.

Theorem 5.3. *Let N, d denote integers such that $d > 3$ and $N > 4d$. There is an instance $\left\{\{\boldsymbol{x}_{t,i}\}_{t\in[T],i\in[N]}, \boldsymbol{\theta}^*\right\}$ where, for all $t \in [T]$ and $i \in [N]$, $\|x_{t,i}\| \leq 1$ and $\|\boldsymbol{\theta}^*\| \leq 1$, $k_t = N, \forall t \in [T]$ such that $K = NT$. We have that*

$$\mathbb{E}(R_T) = \Omega(\sqrt{dK}). \quad (10)$$

The main intuition of the proof is to consider a problem in low dimension, with $d = 2$ and $N = 4$ where the regret scales as $\sqrt{T}$. Using an additional dimension, it is possible to prove that the regret must scale as $\sqrt{NT}$ for larger values of N, by creating $N/4$ independent instances. Finally, by further considering $d/3$ independent problems, we prove that the final regret scales as $\sqrt{dNT}$ as claimed.

6 CONCLUSION

We studied the problem of online matchmaking under the contextual *linear stochastic transitivity* model. We provide two algorithms Maximum-Informative Matching 1 and Sup-Matchmaking to solve this problem, each optimal in a certain regime. The contributions of this work are two-fold: first, we show that **(MIM)** achieves optimal regret for arbitrarily large K. Second, we generalized the ideas of (Auer, 2002), through Sup-Matchmaking 2, to a particular combinatorial setting, achieving a regret of $\tilde{O}(\sqrt{dK})$. We also provided a lower bound that certifies the optimality.

Algorithm 2 SUP-MATCHMAKING

Input: $\tau > 0$, $c > 0$;
Initialization: Pick τ Matchings $M_1, \dots, M_\tau$, observe the outcome $\{Y_{t,\{i,j\}}, \{i,j\} \in M_t\}$, $\forall t \in [\tau]$;
Set $S = \lfloor \log_2 \sqrt{NT} \rfloor$, $\Psi^{(\infty)} = \emptyset, \Psi^{(1)} = \ldots = \Psi^{(S)} = \{(t, \{i,j\}), \{i,j\} \in M_t,\ t \in [\tau]\}$;
for $t = \tau + 1, 2, \dots, T$ **do**
    Observe context vectors $\boldsymbol{X}_t = (\boldsymbol{x}_{t,1} \dots, \boldsymbol{x}_{t,n})$;
    Set $s = 1$ and $A_t^{(s)} = [N]$;
    Set $\omega_t(i,j) = 0$ for all $i, j \in [N]$;
    Set $O_t(i,j) = 0$ for all $i, j \in [N]$;
    while TRUE **do**
        Compute MLE $\hat{\boldsymbol{\theta}}_t^{(s)}$ using only data in $\Psi^{(s)}$;
        Set $\Sigma_t^{(s)} = \sum_{(l,\{i,j\}) \in \Psi^{(s)}} \boldsymbol{z}_{l,i,j} \boldsymbol{z}_{l,i,j}^\intercal$;
        Set $c_{t,i,j}^{(s)} = \beta(\delta) \|\boldsymbol{z}_{t,i,j}\|_{(\Sigma_t^{(s)})^{-1}} \forall j \in A_t^{(s)}$;
        if $c_{t,i,j}^{(s)} \le 1/\sqrt{NT}, \quad \forall \{i,j\} \in A_t^{(s)}$ **then**
            ORDER$(c_t^{(s)}, w_t, O_t)$;
            Break;
        else
            $A_t^{(s+1)} \leftarrow \text{PRUNE}(A_t^{(s)}, c_t^{(s)}, \boldsymbol{z}_t^{(s)}, \hat{\boldsymbol{\theta}}_t^{(s)})$;
        end if
        if $A_t^{(s+1)} = \emptyset$ **then**
            Break;
        end if
        $s \leftarrow s + 1$;
    end while
end for
$M_t = \text{PICKMATCHING}(w_t, O_t)$;
Query the pairs in M_t;
for $\{i,j\} \in M_t$ **do**
    $\Psi^{(\omega(i,j))} = \Psi^{(\omega(i,j))} \cup \{(t, \{i,j\})\}$;
end for

Acknowledgements

Vianney Perchet acknowledges the support of the ANR through the grant DOOM (ANR-23- CE23-0002) and through the PEPR IA FOUNDRY project (ANR-23-PEIA-0003)

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):1–27, 2008.

Miklós Ajtai, János Komlós, and Endre Szemerédi. An 0 (n log n) sorting network. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 1–9, 1983.

Shay Assaf and Eli Upfal. Fault tolerant sorting network. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 275–284. IEEE, 1990.

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

Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Kenneth E Batcher. Sorting networks and their applications. In *Proceedings of the April 30–May 2, 1968, spring joint computer conference*, pages 307–314, 1968.

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

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pages 1764–1786. PMLR, 2022.

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

Mark Braverman, Jieming Mao, and S Matthew Weinberg. Parallel algorithms for select and partition with noisy comparisons. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 851–862, 2016.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

Pinhan Chen, Chao Gao, and Anderson Y Zhang. Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775–1805, 2022.

Raphael Chetrite, Roland Diel, and Matthieu Lerasle. The number of potential winners in bradley–terry model in random environment. *The Annals of Applied Probability*, 2017.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Vaŝek Chvátal. Lecture notes on the new aks sorting network. Technical report, Rutgers University, 1992.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, number 101, pages 355–366, 2008.

S.S. Dragomir, M.L. Scholz, and J. Sunde. Some upper bounds for relative entropy and applications. *Computers & Mathematics with Applications*, 39(9-10): 91–100, 2000.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR, 2015.

Hafedh El Ferchichi, Matthieu Lerasle, and Vianney Perchet. Active ranking and matchmaking, with perfect matchings. In *ICML 2024-The Forty-First International Conference on Machine Learning*, 2024a.

Hafedh El Ferchichi, Matthieu Lerasle, and Vianney Perchet. Active ranking and matchmaking, with perfect matchings. In *Forty-first International Conference on Machine Learning*, 2024b.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York, 1978. ISBN 0668047216.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. *Advances in Neural Information Processing Systems*, 30, 2017a.

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Maximum selection and ranking under noisy comparisons. In *International Conference on Machine Learning*, pages 1088–1096. PMLR, 2017b.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Chao Gao, Yandi Shen, and Anderson Y Zhang. Uncertainty quantification in the bradley–terry–luce model. *Information and Inference: A Journal of the IMA*, 12(2):1073–1140, 2023.

Linxia Gong, Xiaochuan Feng, Dezhi Ye, Hao Li, Runze Wu, Jianrong Tao, Changjie Fan, and Peng Cui. Optmatch: Optimized matchmaking via modeling the high-order interactions on the arena. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2300–2310, 2020.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067, 2020.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, and Martin J Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 2019.

Robert J Henery. An extension to the thurstone-mosteller model for chess. *Journal of the Royal Statistical Society Series D: The Statistician*, 41(5): 559–567, 1992.

Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006.

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

Donald E Knuth. *The Art of Computer Programming: Sorting and Searching, volume 3*. Addison-Wesley Professional, 1998.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tom Leighton, Yuan Ma, and C Greg Plaxton. Breaking theθ (n log2 n) barrier for sorting with faults. *Journal of Computer and System Sciences*, 54(2): 265–304, 1997.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

Tom Minka, Ryan Cleven, and Yordan Zaykov. Trueskill 2: An improved bayesian skill rating system. *Technical Report*, 2018.

Soheil Mohajer, Changho Suh, and Adel Elmahdy. Active learning for top-k rank aggregation from noisy comparisons. In *International Conference on Machine Learning*, pages 2488–2497. PMLR, 2017.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 429–438, 2020.

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

Wenbo Ren, Jia Kevin Liu, and Ness Shroff. On sample complexity upper and lower bounds for exact

ranking from noisy comparisons. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Aadirupa Saha and Aditya Gopalan. Active ranking with subset-wise preferences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3312–3321. PMLR, 2019.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, pages 9235–9244. PMLR, 2021.

Dirk Schäfer and Eyke Hüllermeier. Dyad ranking using plackett–luce models based on joint feature representations. *Machine Learning*, 107:903–941, 2018.

Balázs Szörényi, Róbert Busa-Fekete, Adil Paul, and Eyke Hüllermeier. Online rank elicitation for plackett-luce: A dueling bandits approach. *Advances in neural information processing systems*, 28, 2015.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9791–9798, 2021.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 101(2):266, 1994.

Leslie G Valiant. Parallelism in comparison problems. *SIAM Journal on Computing*, 4(3):348–355, 1975.

Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. *arXiv preprint arXiv:1910.04928*, 2019.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122. PMLR, 2015.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. Copeland dueling bandits. *Advances in neural information processing systems*, 28, 2015.

Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. Online learning to rank in stochastic click models. In *International conference on machine learning*, pages 4199–4208. PMLR, 2017.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

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	
$\mathbb{1}_{[\cdot]}$	Indicator function.
$\mathbb{N}$	Set of natural numbers, i.e., $\mathbb{N} = \{0, 1, 2, 3, \dots\}$.
$\mathbb{R}$	Set of real numbers.
$\mathbb{M}_k$	Set of real square matrices of k rows/columns.
$[n]$	The set $\{1, 2, \dots, n\}$ for some $n \in \mathbb{N}$.
$\boldsymbol{x}, \boldsymbol{z}$	d-dimensional vectors.
$\boldsymbol{A}^\intercal$	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.
$\boldsymbol{0}_d$	$d \times d$ zero matrix.
$\|\boldsymbol{x}\|$	The Euclidean norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$, i.e., $\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$.
$\|\boldsymbol{x}\|_{\boldsymbol{A}}$	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., $\sqrt{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle}$.
Modelling related	
k_t	The number of pairs sampled at round t.
N	Maximum number of players per round.
T	The time horizon (number of rounds).
K	The total number of players: $K = \sum_{t=1}^{T} k_t$.
$\mathcal{X}$	The context space (subset of $\mathbb{R}^d$).
$\boldsymbol{x}_{t,i}$	(d – dimensional) context vector related to player i at time step $t \in [T]$ (element in $\mathcal{X}$).
$\boldsymbol{z}_{t,i,j}$	The contrast vector of players i and j at time step t, i.e., $\boldsymbol{z}_{t,i,j} = \boldsymbol{x}_{t,i} - \boldsymbol{x}_{t,j}$.
$\boldsymbol{\theta}^*$	Ground truth weight vector of the underlying model.
$M_t = \{\{i_{t,1}, i_{t,2}\}, \dots, \{i_{t,k_t-1}, i_{t,k_t}\}\}$	Selected matching of players at time step t.
F	Comparison function defining the **(CoLST)** model.
κ	Minimum of F' on working set: $\kappa := \inf\{F'(\langle \hat{\theta}, x \rangle); \|x\|_2 \leq 1, \|\hat{\theta} - \theta^*\| < 1\} > 0$.
Λ	F' is $\Lambda-$Lipschitz.
L	F is $L-$Lipschitz.
$Y_{t,(i,j)}$	(Binary) Feedback for the selected couple $\{i, j\}$, i.e., $Y_t = 1_{[i_{t,k} \succ j_{t,k}]} \sim \text{Ber}(F(\langle \boldsymbol{\theta}^*, z_{t,i,j} \rangle))$.
$\mu_{t,i}$	Strength of player i: $\mu_{t,i} = \langle \boldsymbol{x}_{t,i}, \boldsymbol{\theta}^* \rangle$.
$c_t(i,j)$	Cost of matching i with j; $c_t(i,j) = \|\mu_{t,i} - \mu_{t,j}\|$.
$c_t(M)$	Cost of selecting a matching $M = \{\{i_1, i_2\}, \dots, \{i_{k_t-1}, i_{k_t}\}\}$ at time step t : $C_t(M) = \sum_{\{i,j\} \in M} c(i,j)$.
M_t^*	Optimal matching at time step t, i.e., the matching with the lowest cost at time t : $M_t^* = \arg\min_{M \ matching} C_t(M)$.
R_T	Cumulative regret up to time T for selecting $(M_t)_{t \in [T]}$: $R_T = \sum_{t=1}^{T} r_t$.
Algorithm related	
c	Confidence width parameter of Algorithms 1 and 2 (hyperparameter of Algorithms 1 and 2).
τ	Pure exploration length for Algorithms 1 and 2.
$\hat{\boldsymbol{\theta}}_t$	Maximum-likelihood estimate estimator of $\boldsymbol{\theta}^*$ (Algorithm 1).
$\boldsymbol{\Sigma}_t$	Gram matrix at round t: $\boldsymbol{\Sigma}_t = \sum_{t=1}^{T} \sum_{\{i,j\} \in M_t} \boldsymbol{z}_{t,i,j} \boldsymbol{z}_{t,i,j}^\intercal$.
$\boldsymbol{W}_t$	Square matrix of size k_t. Contains the uncertainty: $\boldsymbol{W}_{t,i,j} = \|\|z_{t,i,j}\|\|_{\boldsymbol{\Sigma}_t^{-1}}$
$\boldsymbol{E}_t$	Adjacency matrix: if $\boldsymbol{E}_{t,i,j} = 1$, Algorithm 1 can pick $\{i, j\}$.
$\boldsymbol{O}_t$	Ordering matrix: if $\boldsymbol{O}_{t,i,j} = 1$ (resp. $\boldsymbol{O}_{t,i,j} = -1$) then Algorithm 1 is confident i is stronger(resp. weaker) than j.
r_i	The estimated rank of i.
s	Stage-index in Algorithm 2.
S	Maximal number of stages (set to $\lfloor \log_2(\sqrt{NT}) \rfloor$).
$\omega_t(i,j)$	Stage at which edge $\{i, j\}$ is at round t.
$\Psi^{(s)}$	All samples corresponding to stage $s \in [S]$.
$\hat{\boldsymbol{\theta}}_t^{(s)}$	Maximum-likelihood estimate of the weight vector $\boldsymbol{\theta}^*$ using only observations from $\Psi^{(s)}$.
$\boldsymbol{\Sigma}_t^{(s)}$	Gram matrix at time step t and stage s i.e., $\boldsymbol{\Sigma}_t^{(s)} = \sum_{(l,\{i_l,j_l\}) \in \Psi^{(s)}} \boldsymbol{z}_{l,i_l,j_l} \boldsymbol{z}_{l,i_l,j_l}^\intercal$.
$\hat{\mu}_{t,i}^{(s)}$	Estimated strength of player i in stage s at time t, i.e., $\hat{\mu}_{t,i}^{(s)} = \langle \hat{\boldsymbol{\theta}}_t^{(s)}, \boldsymbol{x}_{t,i} \rangle$.
$c_{t,i,j}^{(s)}$	Uncertainty on $\hat{\mu}_{t,i}^{(s)} - \hat{\mu}_{t,j}^{(s)}$. $c_{t,i,j}^{(s)} := \beta(\delta) \|\|\boldsymbol{z}_{t,i,j}\|\|_{(\boldsymbol{\Sigma}_t^{(s)})^{-1}}$.

B Omitted Pseudo-codes

Algorithm 3 LN

Input: $i \in [k_t], V_t \in [k_t],\ \boldsymbol{E}_t, \boldsymbol{O}_t, \boldsymbol{W}_t \in \mathbb{M}_{k_t}$;
$L = \{j \in [k_t], \boldsymbol{O}_{t,i,j} = -1\}$;
for $j \in L$ **do**
  if There is $k \in L$ s.t. $\boldsymbol{O}_{t,j,k} = 1$ **then** ▷ j is confidently not the weakest element of L.
    $L \leftarrow L \setminus \{j\}$;
  end if
end for
for $j \in L$ **do**
  $\boldsymbol{E}_{t,i,j}, \boldsymbol{E}_{t,i,j} = 1, 1$; ▷ Update Adjacency matrix.
end for

Algorithm 4 ORDER

Input: $z_t \in \mathbb{M}_{k_t}$, $\hat{\boldsymbol{\theta}}_t^{(s)} \in \mathbb{R}^d$, $\boldsymbol{O}_t, w_t \in \mathbb{R}^d$
for $\{i,j\} \in A_t^{(s)}$ **do**
  if $\langle \boldsymbol{z}_{t,i,j}, \hat{\boldsymbol{\theta}}_t^{(s)} \rangle \geq 1/\sqrt{NT}$ **then**
    $O_t(i,j) = 1$;
  else if $\langle \boldsymbol{z}_{t,i,j}, \hat{\boldsymbol{\theta}}_t^{(s)} \rangle \leq -1/\sqrt{NT}$ **then**
    $O_t(i,j) = -1$;
  else
    $\omega_t(i,j) = S$;
  end if
end for

Algorithm 5 PRUNE

Input: $A_t^s \subset \{\{i,j\}, i,j \in [k_t], i \neq j\}, c_t^{(s)} \in \mathbb{M}_{k_t}, (\boldsymbol{z}_{t,i,j}^{(s)})_{i,j \in [k_t]}, \hat{\boldsymbol{\theta}}_t^{(s)} \in \mathbb{R}^d$;
$A_t^{(s+1)} = \emptyset$;
for $\{i,j\} \in A_t^{(s)}$ **do**
  if $c_{t,i,j}^{(s)} > 1/2^s$ **then**
    $\omega_t(i,j) = s$;
  else if $\langle \boldsymbol{z}_{t,i,j}, \hat{\boldsymbol{\theta}}_t^{(s)} \rangle \geq 1/2^s$ **then**
    $O_t(i,j) = 1$;
  else if $\langle \boldsymbol{z}_{t,i,j}, \hat{\boldsymbol{\theta}}_t^{(s)} \rangle \leq -1/2^s$ **then**
    $O_t(i,j) = -1$;
  else
    $A_t^{(s+1)} = A_t^{(s+1)} \cup \{i,j\}$;
  end if
end for ▷ Proceed to next stage.

Algorithm 6 RN

Input: $i \in [k_t], V_t \in [k_t],\ \boldsymbol{E}_t, \boldsymbol{O}_t, \boldsymbol{W}_t \in \mathbb{M}_{k_t}$;
$R = \{j \in [k_t], \boldsymbol{O}_{t,i,j} = 1\}$;
for $j \in R$ **do**
  if There is $k \in R$ s.t. $\boldsymbol{O}_{t,k,j} = 1$ **then** ▷ j is confidently not the strongest element of R.
    $R \leftarrow R \setminus \{j\}$;
  end if
end for
for $j \in R$ **do**
  $\boldsymbol{E}_{t,i,j}, \boldsymbol{E}_{t,i,j} = 1, 1$; ▷ Update Adjacency matrix.
end for

Algorithm 7 PickMatchingSM

Input: $\omega_t, O_t \in \mathbb{M}_N(\mathbb{R}), S = \lfloor \log_2 \sqrt{NT} \rfloor$;
$V = [N]$;
for $s \in [S]$ **do**
    pick a maximal matching $M_t^{(s)}$ on V using edges $\{i, j\}$ such that $\omega_t(i, j) = s$; ▷ 2-approx of best possible.;
    $V = V \setminus M_t^{(s)}$;
end for
Set $R =$ Ranking of V following O_t;
Pick $M_t^{(S+1)}$ Match elements of V according to the ranking R;
return $M_t^{(1)} \cup M_t^{(2)} \cup \ldots \cup M_t^{(S+1)}$;

Algorithm 8 PickMatchingMIM

Input: $V_t \subset [k_t], \boldsymbol{E}_t, \boldsymbol{W}_t \in \mathbb{M}_{k_t}$;
Let $G = ([k_t], \boldsymbol{E}_t, \boldsymbol{W}_t)$ a weighted graph; ▷ $\boldsymbol{E}_t$ the adjacency matrix and $\boldsymbol{W}_t$ the weight matrix.
Pick M_t a maximal weight matching on G;
return M_t;

C Technical lemmas

We start by proving the Lemma 4.1. For convenience, we restate the lemma here:

Lemma C.1. *If $\mu_1 > \mu_2 > \ldots > \mu_k$, then the optimal matching is*

$$M^* = \{\{1, 2\}, \{3, 4\}, \ldots, \{k-1, k\}\}. \tag{11}$$

Proof. Take any two *crossing* edges in an arbitrary matching, say $\{i, k\}$ and $\{j, \ell\}$ with $\mu_i < \mu_j < \mu_k < \mu_\ell$. Replace them by the *non-crossing* pair $\{i, j\}$ and $\{k, \ell\}$. The change in total weight is

$$\begin{aligned}\Delta &= (\mu_j - \mu_i) + (\mu_\ell - \mu_k) - (\mu_k - \mu_i) - (\mu_\ell - \mu_j) \\ &= 2(\mu_j - \mu_k) \;\le\; 0.\end{aligned}$$

Hence each *uncrossing* step never increases the total cost; iterating until no crossings remain yields the neighbour-pairing M^*, which is therefore optimal. □

C.1 Matching

In this part, we fix $k_t = N$ and $x_{t,i} = x_i$, omitting the time dependency for the remainder of this section. Let E be a subset of $[N]$. The optimal matching on E according to the cost c will be denoted as $\mathbf{OPT}(E)$. Recall that the cost function implies that the optimal matching on a given set is the one obtained through ranking the elements of the set according to their strength then matching them from weakest to strongest.

W.l.o.g, we assume that if $i \le j$, then element i is stronger than element j. The following lemma holds.

Lemma C.2. *Let $i, j \in E \subset [N]$. Then*

$$c(\boldsymbol{OPT}(E \setminus \{i, j\})) \le c(\boldsymbol{OPT}(E)) + \frac{L}{\kappa} c(i, j). \tag{12}$$

Algorithm 9 RANK

Input: i;
Set $r = 1$;
for $j \in [k_t]$ **do**
    if $O_{t,j,i} = 1$ **then**
        $r = r + 1$; ▷ j is stronger that i.
    end if
end for
return r;

Proof. W.l.o.g., assume that $E = [N]$ and that $i \in \{1,2\}$ and $j \in \{N-1, N\}$. Recall that $\mathbf{OPT}([N]) = \{\{1,2\},\{3,4\}\dots\{N-1,N\}\}$. We have that $\mathbf{OPT}([N] \setminus \{i,j\}) = \{A, \{4,5\},\{6,7\},\dots,\{N-4,N-3\}, B\}$, where

$$A = \begin{cases} \{1,3\}, & \text{if } i = 2. \\ \{2,3\}, & \text{if } i = 1. \end{cases} \tag{13}$$

$$B = \begin{cases} \{N-2, N\}, & \text{if } j = N-1. \\ \{N-2, N-1\}, & \text{if } j = N. \end{cases} \tag{14}$$

In all cases, through the triangle inequality, we get that:

$$c(\mathbf{OPT}([N] \setminus \{i,j\})) \leq c(1,2) + \underbrace{c(2,3) + \left(\sum_{k=2}^{(N-4)/2} c(2k, 2k+1)\right) + c(N-2, N-1)}_{\alpha} + c(N-1, N). \tag{15}$$

Recall that Assumption 4.1 implies

$$(F(\mu_1) + F(\mu_2) + \ldots + F(\mu_k)) \leq \frac{L}{\kappa} F(\mu_1 + \mu_2 + \ldots + \mu_k). \tag{16}$$

Hence

$$\alpha \leq \frac{L}{\kappa} c(i,j). \tag{17}$$

In conclusion, it holds that:

$$c\left(\mathbf{OPT}\left([N] \leq \{i,j\}\right)\right) \leq c(1,2) + c(N-1,N) + \frac{L}{\kappa} c(i,j) \leq c(\mathbf{OPT}([N])) + \frac{L}{\kappa} c(i,j). \tag{18}$$

□

The previous lemma essentially implies that the optimal matching constrained to containing a given pair $\{i,j\}$ is not significantly more costly than the optimal matching free of any constraint. The degradation is proportional to the cost of the imposed pair $\{i,j\}$. The following result is a generalization of Lemma C.2 when multiple pairs are imposed.

Theorem C.3. *Let $i_1, j_1, i_2, j_2, \dots, i_k, j_k \in [N]$ be distinct elements. Then*

$$c(\boldsymbol{OPT}([N] \setminus \{i_1, j_1, i_2, j_2, \dots, i_l, j_l\}) \leq c(\boldsymbol{OPT}([N]) + \frac{L}{\kappa} \sum_{k=1}^{l} c(i_k, j_k). \tag{19}$$

Proof. The proof follows by recursion using the lemma C.2 □

C.2 Self-Normalized bound

The following lemma is an adaptation of Lemma 11 from (self-normalized bound)(Abbasi-Yadkori et al., 2011; Takemura et al., 2021) for this setting.

Lemma C.4. *Let $\{\boldsymbol{z}_{t,i,j}\}_{t \in [\tau+1,\dots,T], \{i,j\} \in M_t}$ such that $||\boldsymbol{z}_{i,j}|| \leq 1$ and $\boldsymbol{\Sigma}_t = \sum_{s=1}^{t-1} \sum_{\{i,j\} \in M_t} \boldsymbol{z}_{s,i,j} \boldsymbol{z}_{s,i,j}^{\intercal}$ such that $\Sigma_{\tau+1} \succeq I_d$. Let $k_t = \#M_t$ and $K = \sum_{t=\tau+1}^{T} k_t$ Then*

$$\sum_{t=\tau+1}^{T} \sum_{\{i,j\} \in M_t} \min(\frac{1}{\sqrt{k_t}}, ||\boldsymbol{z}_{t,i,j}||_{\Sigma_t^{-1}}) \leq \sqrt{2dK \log \frac{2K}{d}}. \tag{20}$$

Proof. First, we have that:

$$\log\det(\Sigma_t) = \log\det\left(\Sigma_{t-1} + \sum_{\{i,j\}\in M_t} \boldsymbol{z}_{i,j}\boldsymbol{z}_{i,j}^\intercal\right) \tag{21}$$

$$= \log\det\left(\Sigma_{t-1}\right) + \log\det(I_d + \sum_{\{i,j\}\in M_t} (\Sigma_{t-1}^{-1/2}\boldsymbol{z}_{i,j})(\Sigma_{t-1}^{-1/2}\boldsymbol{z}_{i,j})^\intercal) \tag{22}$$

$$= \log\det(\Sigma_{t-1}) + \log\det(\frac{1}{k_t}\sum_{\{i,j\}\in M_t} I_d + k_t(\Sigma_{t-1}^{-1/2}\boldsymbol{z}_{i,j})(\Sigma_{t-1}^{-1/2}\boldsymbol{z}_{i,j})^\intercal) \tag{23}$$

$$\geq \log\det(\Sigma_{t-1}) + \frac{1}{k_t}\sum_{\{i,j\}\in M_t} \log\det(I_d + k_t(\Sigma_{t-1}^{-1/2}\boldsymbol{z}_{i,j})(\Sigma_{t-1}^{-1/2}\boldsymbol{z}_{i,j})^\intercal) \tag{24}$$

$$= 1 + \sum_{t=\tau+1}^{T}\frac{1}{k_t}\sum_{\{i,j\}\in M_t} \log\det(I_d + k_t(\Sigma_{t-1}^{-1/2}\boldsymbol{z}_{i,j})(\Sigma_{t-1}^{-1/2}\boldsymbol{z}_{i,j})^\intercal) \tag{25}$$

$$= 1 + \sum_{t=\tau+1}^{T}\frac{1}{k_t}\sum_{\{i,j\}\in M_t} \log(1 + k_t||\boldsymbol{z}_{i,j}||^2_{\Sigma_t^{-1}}). \tag{26}$$

Since for $0 \leq x \leq 1$, we have that $2\log(1+x) \geq x$, it follows that:

$$\log\det\Sigma_t \geq 1 + \frac{1}{2}\sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} \min(1, k_t||\boldsymbol{z}_{i,j}||^2_{\Sigma_t^{-1}}). \tag{27}$$

Moreover, since $d(\det A)^{1/d} \leq \mathrm{tr}(A)$:

$$\log\det\Sigma_t = d\left(\log d(\det\Sigma_t)^{1/d} - \log d\right) \tag{28}$$

$$\leq d\log\left(\mathrm{tr}(\Sigma_t)/d\right) \tag{29}$$

$$= d\log\left(\frac{1}{d}\sum_{t=\tau+1}^{T}\sum_{\{i,j\},\in M_t} ||\boldsymbol{z}_{i,j}||^2\right) \tag{30}$$

$$\leq d\log\left(\frac{K}{d}\right). \tag{31}$$

$$\tag{32}$$

Hence, it follows that:

$$\sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} \min(\frac{1}{k_t}, ||\boldsymbol{z}_{i,j}||^2_{\Sigma_t^{-1}}) \leq 2d\log\left(\frac{K}{d}\right) \tag{33}$$

and, by Cauchy-Schwarz inequality:

$$\sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} ||\boldsymbol{z}_{i,j}||_{\Sigma_t^{-1}} \leq \sqrt{K}\sqrt{\sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} ||\boldsymbol{z}_{i,j}||^2_{\Sigma_t^{-1}}} \leq \sqrt{2dK\log\frac{2K}{d}}. \tag{34}$$

This concludes the proof. □

Lemma C.5. *Let $\{\boldsymbol{z}_{t,i,j}\}_{t\in[\tau+1,\ldots,T],\{i,j\}\in M_t}$ such that $\boldsymbol{z}_{i,j} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_t = \sum_{s=1}^{t-1}\sum_{\{i,j\}\in M_t}\boldsymbol{z}_{i,j}\boldsymbol{z}_{i,j}^\intercal$ such that $\Sigma_{\tau+1} \succeq I_d$. Let $k_t = \#M_t$ and $K = \sum_{t=\tau+1}^{T} k_t$ Then*

$$\sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} \mathbb{1}(\|\boldsymbol{z}_{t,i,j}\|_{\Sigma_t^{-1}} \geq \frac{1}{\sqrt{k_t}}) \leq 2d\max_t(k_t)\log\left(\frac{K}{d}\right). \tag{35}$$

Proof. It holds that

$$\sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} \mathbb{1}(\|\boldsymbol{z}_{t,i,j}\|_{\Sigma_t^{-1}} \geq \frac{1}{\sqrt{k_t}}) \leq \sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} \min(1, k_t\|\boldsymbol{z}_{t,i,j}\|^2) \leq 2d\max_t(k_t)\log\left(\frac{K}{d}\right). \tag{36}$$

where the second inequality follows from 33. □

D Proof of Theorem 4.4

Let

$$A_{\text{init}} = \{\forall t > \tau : |\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*| \leq 1 \text{ and } \boldsymbol{\Sigma}_t \geq I_d\} \tag{37}$$

be the initial concentration event. The initialization time τ is chosen to guarantee $\mathbb{P}(A_{\text{init}}) \geq 1 - 1/T$. As in (Bengs et al., 2022; Saha, 2021; Vaswani et al., 2019) (mainly the proof of Theorem 4 (Vaswani et al., 2019)), a valid choice is $\tau = \max\left\{(d\log(T/d) + 2\log T)/\kappa^2\rho, 1/\rho\right\}$

Let

$$A_{\text{MLE}} = \{\forall t > \tau, \{i,j\} \subset [k_t] : |\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*, \boldsymbol{z}_{t,i,j}\rangle| \leq \beta(\delta)\|\boldsymbol{z}_{t,i,j}\|_{(\boldsymbol{\Sigma}_t)^{-1}}\} \tag{38}$$

be the MLE concentration event. $\beta(\delta)$ is chosen to ensure that $\mathbb{P}(A_{\text{MLE}}) \geq 1 - \delta$, again for $\delta = 1/T$. Lemma 3 (Li et al., 2017) implies that it's enough to take

$$\beta(\delta) = \frac{1}{2\kappa}\sqrt{d\log\frac{NT}{d} + 2\log(T)}. \tag{39}$$

Now, assuming that these two events A_{init} and A_{MLE} hold, the rest of the proof lies in expressing the regret bound in terms of the estimation error, formalized in the following lemma:

Lemma D.1. *Suppose that A_{init} and A_{MLE} both hold. Let $t \geq \tau+1$ and M_t be the matching chosen by **(MIM)** (Algorithm 1)). Let M_t^* be the optimal matching at round t. Then*

$$C_t(M_t) - C_t(M_t^*) \leq \sum_{\{i,j\}\in M_t} 2\beta(\delta)\frac{L}{\kappa}||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}. \tag{40}$$

Proof. Recall that $M_t = \arg\max_{M \text{ matching on } [k_t]} \sum_{\{i,j\}\in M} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}$. Recall that V_t is the set of elements that are correctly ranked by **(MIM)** (lines 17 to 19). We partition $M_t = M_{t,1} \cup M_{t,2}$, where $M_{t,1}$ is formed by pair $\{i,j\}$ such that either i or j are elements of V_t, whereas $M_{t,2}$ is formed by pair $\{i',j'\}$ such that neither i' nor j' is in V_t, therefore for $\{i',j'\} \in M_{t,2}$, it holds that

$$|\langle \hat{\boldsymbol{\theta}}_t, \boldsymbol{z}_{t,i',j'}\rangle| \leq \beta(\delta)||\boldsymbol{z}_{t,i',j'}||_{\boldsymbol{\Sigma}_t^{-1}}. \tag{41}$$

We note the set of elements in $M_{t,1}$ as $\bar{V}_t$. We note $M_{t,2} = \{\{i_1,j_1\},\dots,\{i_l,j_l\}\}$. Using Theorem C.3, we get that:

$$C_t(\mathbf{OPT}(\bar{V}_t)) = C_t(\mathbf{OPT}([k_t] \setminus \{i_1,j_1,\dots,i_l,j_l\}) \leq C_t(\mathbf{OPT}([k_t]) + \frac{L}{\kappa}\sum_{m=1}^{l} c_t(i_m,j_m). \tag{42}$$

Also, since

$$M_t = \underset{M \text{ matching on } [k_t]}{\arg\max} \sum_{\{i,j\}\in M} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}} \tag{43}$$

then

$$M_{t,1} = \underset{M \text{ matching on } \bar{V}_t}{\arg\max} \sum_{\{i,j\}\in M} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}. \tag{44}$$

Let $\partial V_t = \bar{V}_t \setminus V_t$. As we will show later, it remains to bound the difference of the costs of $M_{t,2}$ and $\mathbf{OPT}(\bar{V}_t)$. To do so, we rely on the following observations:

1. V_t can be correctly ordered. This is by constructions as only elements for which the rank is known are put into V_t (lines 18 and 19 1).
2. Elements of ∂V_t can be inserted correctly (coherently with the correct ranking) between elements of V_t.
3. Let $i,j,k \in V_t$ be three consecutive elements in the correct ranking of V_t, such that $r(i) < r(j) < r(k)$. Let $l = M_{t,2}(j)$ the match of j in $M_{t,2}$. then l is either ranked between i and j or between k and j, that is either $r(i) < r(l) < r(j)$ or $r(j) < r(l) < r(k)$.

The first two points are self-evident. For the third point, it is sufficient to show that $r(i) < r(l) < r(k)$: this is straightforward: if it was not the case, say for example $r(l) < r(i)$, l would have been excluded by LN (3 line 4). A consequence of the third point is that at most 2 elements of ∂V_t can be inserted between 2 consecutive elements of V_t.

Let $i,l \in V_t$ be two consecutive elements of V_t such that $r(i) < r(l)$. The aforementioned observations imply that there is only 3 possible configurations:

1. No elements of ∂V_t is between i and l. It means $r(M_t(i)) < r(i)$ and $r(l) < r(M_t(l))$.

2. There is one element of ∂V_t between i and l, either $r(M_t(i)) < r(i) < r(M_t(l)) < r(l)$ or $r(i) < r(M_t(i)) < r(l) < r(M_t(l))$.

3. There is two elements of ∂V_t between i and l, either $r(i) < r(M_t(i)) < r(M_t(l)) < r(l)$ or $r(i) < r(M_t(i)) < r(M_t(l)) < r(l)$.

If the third situation does not occur for any $i, l \in V_t$, it is necessary that $M_{t,1} = \mathbf{OPT}(\bar{V}_t)$, as the elements of $\bar{V}_t$ can be ranked, yielding no additional cost compared to $\mathbf{OPT}(\bar{V}_t)$. However, if the third situation occurs and the algorithm cannot decide whether $r(M_t(i)) < r(M_t(l)$ or $r(M_t(l)) < r(M_t(i))$, M_t matches i to $M_t(i)$ and l to $M_t(l)$ because

$$||\boldsymbol{z}_{t,i,M_t(i)}||_{\boldsymbol{\Sigma}_t^{-1}} + ||\boldsymbol{z}_{t,l,M_t(l)}||_{\boldsymbol{\Sigma}_t^{-1}} \geq ||\boldsymbol{z}_{t,i,M_t(l)}||_{\boldsymbol{\Sigma}_t^{-1}} + ||\boldsymbol{z}_{t,l,M_t(i)}||_{\boldsymbol{\Sigma}_t^{-1}}. \tag{45}$$

If $r(M_t(i)) < r(M_t(l))$, this choice does not cause any regret. However, if $r(M_t(i)) > r(M_t(l))$, then

$$\begin{aligned}\langle \boldsymbol{\theta}^*, \boldsymbol{z}_{t,i,M_t(i)} \rangle + \langle \boldsymbol{\theta}^*, \boldsymbol{z}_{t,M_t(l),l} \rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{z}_{t,i,M_t(l)} \rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{z}_{t,M_t(i),l} \rangle &= 2 \langle \boldsymbol{\theta}^*, \boldsymbol{z}_{t,M_t(l),M_t(i)} \rangle &(46)\\ &\leq 2\beta(\delta) ||\boldsymbol{z}_{t,M_t(l),M_t(i)}||_{\boldsymbol{\Sigma_t}^{-1}}. &(47)\end{aligned}$$

Here, the last inequality holds because we assume that A_{MLE} holds and that $\{M_t(l), M_t(i)\}$ are not distinguishable. We also have that

$$\begin{aligned}||\boldsymbol{z}_{t,M_t(l),M_t(i)}||_{\boldsymbol{\Sigma_t}^{-1}} &= \frac{1}{2} ||\boldsymbol{z}_{t,M_t(l),i} + \boldsymbol{z}_{t,i,M_t(i)} + \boldsymbol{z}_{t,M_t(l),l} + \boldsymbol{z}_{t,l,M_t(i)}||_{\boldsymbol{\Sigma_t}^{-1}} &(48)\\ &\leq \frac{1}{2} \left[||\boldsymbol{z}_{t,M_t(l),i}||_{\boldsymbol{\Sigma_t}^{-1}} + ||\boldsymbol{z}_{t,l,M_t(i)}||_{\boldsymbol{\Sigma_t}^{-1}} + ||\boldsymbol{z}_{t,i,M_t(i)}||_{\boldsymbol{\Sigma_t}^{-1}} + ||\boldsymbol{z}_{t,l,M_t(l)}||_{\boldsymbol{\Sigma_t}^{-1}} \right] &(49)\\ &\leq ||\boldsymbol{z}_{t,i,M_t(i)}||_{\boldsymbol{\Sigma_t}^{-1}} + ||\boldsymbol{z}_{t,l,M_t(l)}||_{\boldsymbol{\Sigma_t}^{-1}}. &(50)\end{aligned}$$

where the last inequality is a consequence of the selection criteria 45. By combining 47 and 50, we get

$$C_t(M_{t,1}) - C_t(\mathbf{OPT}(\bar{V}_t)) \leq 2\beta(\delta) \sum_{\{i,j\} \in M_{t,2}} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma_t}^{-1}}. \tag{51}$$

To conclude, it remains to bound $c_t(i,j)$ when $\{i,j\} \in M_{t,2}$. We have that

$$\begin{aligned}c_t(i,j) &= \langle \boldsymbol{\theta}^*, \boldsymbol{z}_{t,i,j} \rangle &(52)\\ &= \langle \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}, \boldsymbol{z}_{t,i,j} \rangle + \langle \hat{\boldsymbol{\theta}}, \boldsymbol{z}_{t,i,j} \rangle &(53)\\ &\leq 2\beta(\delta) ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}. &(54)\end{aligned}$$

where the last inequality follows from combining A_{MLE} and 41 (by definition of $M_{t,2}$). Recall that $\frac{L}{\kappa} \geq 1$. It follows

$$\begin{aligned}C_t(M_t) - C_t(M_t^*) &\leq C_t(M_{t,2}) + C_t(M_{t,1}) - C_t(M_t^*) &(55)\\ &\overset{(1)}{\leq} \sum_{\{i,j\} \in M_{t,2}} c_t(i,j) + 2\beta(\delta) \sum_{\{i,j\} \in M_{t,1}} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}\\ &\quad + C_t(\mathbf{OPT}([k_t])) - C_t(M_t^*) + \frac{L}{\kappa} \sum_{\{i,j\} \in M_{t,2}} c_t(i,j) &(56)\\ &\overset{(2)}{\leq} 2\frac{L}{\kappa} \beta(\delta) \sum_{\{i,j\} \in M_t} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}. &(57)\end{aligned}$$

Where the inequality (1) follows from 42 and 51, and inequality (2) follows from combining 54 and the fact that $1 \leq \frac{L}{\kappa}$. This concludes the proof. □

Using the previous lemma D.1, combined with Lemmas C.4 and C.5, we show the following Theorem 4.4.

Theorem D.2. *Let* $\tau = \max\{(d\log(T) + 2\log T)/\kappa^2\rho, 1/\rho\}$ *and* $c = \beta(\delta) = \frac{1}{2\kappa}\sqrt{d\log(NT/d) + 2\log(T)}$. *Let* $K = \sum_{t=1}^{T} k_t$. *Then the regret incurred by 1 is upper-bounded by*

$$\mathbb{E}[R_T] = O\left(\frac{L}{\kappa^2}\left[d\sqrt{K}\log NT + d^{3/2}N\log\left(\frac{NT}{d}\right)\right]\right). \tag{58}$$

Proof. We start by decomposing the regret:

$$\mathbb{E}\left[\sum_{t=1}^{T} C_t(M_t) - C_t(M_t^*)\right] \leq \sum_{t=1}^{\tau} k_t + K\mathbb{P}(A_{\text{MLE}}^c \cup A_{\text{init}}^c) + \mathbb{P}(A_{\text{MLE}} \cap A_{\text{init}}) \sum_{t=\tau+1}^{T} C_t(M_t) - C_t(M_t^*) \tag{59}$$

$$\leq N\tau + 2N + 2\beta(\delta)\frac{L}{\kappa}\sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}}. \tag{60}$$

Using the lemmas C.4 and C.5, we get that

$$\sum_{t=\tau+1}^{T}\sum_{\{i,j\}\in M_t} ||\boldsymbol{z}_{t,i,j}||_{\boldsymbol{\Sigma}_t^{-1}} \leq \sqrt{2dK\log\frac{2K}{d}} + 2dN\log\left(\frac{K}{d}\right). \tag{61}$$

In conclusion, we get that

$$\mathbb{E}\left[\sum_{t=1}^{T} C_t(M_t) - C_t(M_t^*)\right] \leq N\tau + 2N + \frac{L}{\kappa^2}\sqrt{d\log\left(\frac{NT}{d}\right) + 2\log(NT)}\left(\sqrt{2dK\log\frac{2K}{d}} + 2dN\log\left(\frac{K}{d}\right)\right) \tag{62}$$

$$= O\left(\frac{L}{\kappa^2}\left[d\sqrt{K}\log NT + d^{3/2}N\log\left(\frac{NT}{d}\right)\right]\right). \tag{63}$$

because $N\tau = O(Nd)$ □

E Proof of Theorem 5.1

Let

$$A_{\text{init}}^{(S)} = \{\forall t > \tau\,, \forall s \in [S] : |\hat{\boldsymbol{\theta}}_t^{(s)} - \boldsymbol{\theta}^*| \leq 1 \text{ and } \boldsymbol{\Sigma}_t \geq I_d\} \tag{64}$$

be the initial concentration event. The initialization time τ is chosen to guarantee that $\mathbb{P}(A_{\text{init}}^{(S)}) \geq 1 - 1/T$. As in (Bengs et al., 2022; Saha, 2021; Vaswani et al., 2019), a valid choice is $\tau = \max\left\{(d\log(T/d) + 2\log T)/\kappa^2\rho, 1/\rho\right\}$.

Let

$$A_{\text{MLE}}^{(S)} = \{\forall\{i,j\} \subset [n], t > \tau\,, \forall s \in [S] : |\,\langle\hat{\boldsymbol{\theta}}_t^{(s)} - \boldsymbol{\theta}^*, \boldsymbol{z}_{t,i,j}\rangle\,| \leq \beta(\delta)\|\boldsymbol{z}_{i,j}\|_{(\boldsymbol{\Sigma}_t^{(s)})^{-1}}\} \tag{65}$$

be the MLE concentration event. $\beta(\delta)$ is chosen to ensure that $\mathbb{P}(A_{\text{MLE}}^{(S)}) \geq 1 - \delta$. This is an important point as the sophisticated sampling scheme of 2 is designed to allow for a $\tilde{O}(\sqrt{d})$ smaller $\beta(\delta)$. The following lemma holds because the way we pick a matching in Sup-Matchmaking, through the function PickMatchingSM:

Lemma E.1. *For all time steps $t \in [T]$ and any stages $s \in [S]$, $\Psi^{(s)}(t)$ the corresponding preference observations $(Y_{(t,(i,j))})_{(t,(i,j))\in\Psi^{(s)}(t)}$ are independent Bernoulli-distributed random variables with $Y_{(t,(i,j))}$ having the probability of success $F\left(\langle\boldsymbol{z}_{t,i,j}, \boldsymbol{\theta}^*\rangle\right)$.*

Proof. The event $\left\{(t,\{i,j\}) \in \Psi^{(s)}(t)\right\}$ only depends on the results of the samples in $\cup_{\sigma<s}\Psi^{(\sigma)}(t)$ and in $c_{t,a}^{(s)}$ (Algorithm 7, line 4). From the definition of $c_{t,a}^{(s)}$, we know that it depends only on the feature vectors $\boldsymbol{z}_{t,i,j}$, such that $(t,\{i,j\}) \in \Psi_s(t)$. Hence, conditioning on $\cup_{\sigma<s}\Psi^{(\sigma)}(t)$ and on the feature vectors in $\Psi^{(s)}(t)$, the samples $(Y_{(t,(i,j))})_{(t,\{i,j\})\in\Psi^{(s)}(t)}$ are independent, and $\mathbb{E}(Y_{(t,(i,j))}) = F(\langle\boldsymbol{\theta}^*, \boldsymbol{z}_{t,i,j}\rangle)$. □

Lemma E.2. *If $\beta(\delta) = \frac{3}{2\kappa}\sqrt{2\log(3TN/\delta)}$ then*

$$\mathbb{P}(A_{\text{MLE}}^{S}) \geq 1 - \delta. \tag{66}$$

Proof. The proof is analogous to that of Lemma 4 of (Saha, 2021). □

Lemma E.3. *On the event $A_{\text{MLE}}^{(S)}$, the cost associated to the choice of a pair $(i,j) \in M_t$ is upper-bounded as follows:*

$$c_t(i,j) \leq \frac{4}{2^{\omega_t(i,j)}}. \tag{67}$$

Proof. Suppose that $A_{\text{MLE}}^{(s)}$ holds. Since $\{i,j\} \in A_t^{\omega(i,j)}$, then

$$|\langle \hat{\boldsymbol{\theta}}_t^{(\omega(i,j)-1)} - \boldsymbol{\theta}^*, \boldsymbol{z}_{t,i,j} \rangle| \leq c_{t,i,j}^{(\omega(i,j)-1)} \leq 2^{-(\omega(i,j)-1)}. \tag{68}$$

and

$$|\langle \hat{\boldsymbol{\theta}}_t^{(\omega(i,j)-1)}, \boldsymbol{z}_{t,i,j} \rangle| \leq 2^{-(\omega(i,j)-1)} \tag{69}$$

Hence

$$|\langle \boldsymbol{\theta}^*, \boldsymbol{z}_{t,i,j} \rangle| \leq 4 * 2^{-\omega(i,j)}. \tag{70}$$

□

Lemma E.4. *On the event $A_{\text{init}}^{(S)}$ it holds for any $s \in [S]$ that*

$$|\Psi^{(s)}| \leq \beta(\delta) 2^s \left(\sqrt{2d|\Psi^{(s)}| \log\left(\frac{2NT}{d}\right)} + dN \log\left(\frac{NT}{d}\right) \right). \tag{71}$$

Proof. Notice That, by definition of $\Psi^{(s)}$, we have (Algorithm 2, lines 27 and 28)

$$\sum_{(t,\{i,j\}) \in \Psi^{(s)}} c_{t,i,j}^{(s)} \geq \frac{|\Psi^{(s)}|}{2^s}. \tag{72}$$

Moreover

$$\sum_{(t,\{i,j\}) \in \Psi^{(s)}} c_{t,i,j}^{(s)} = \beta(\delta) \sum_{(t,\{i,j\}) \in \Psi^{(s)}} \|\boldsymbol{z}_{t,i,j}\|_{(\boldsymbol{\Sigma}_t^{(s)})^{-1}}. \tag{73}$$

$$\tag{74}$$

Following the notation used in Lemma C.4, let $M_t^{(s)} = \left\{ \{i,j\}, (t,\{i,j\}) \in \Psi^{(s)} \right\}$, such that $k_t^{(s)} = |M_t^{(s)}|$ and $K^{(s)} = |\Psi^{(s)}|$. Hence, by applying the lemmas C.4 and C.5, it follows that

$$\sum_{(t,\{i,j\}) \in \Psi^{(s)}} \|\boldsymbol{z}_{t,i,j}\|_{\boldsymbol{\Sigma}_t^{-1}} \leq \sqrt{2dK^{(s)} \log \frac{2K^{(s)}}{d}} + 2d \max_t(k_t^{(s)}) \log\left(\frac{K^{(s)}}{d}\right). \tag{75}$$

Finally, we get that

$$|\Psi^{(s)}| \leq \beta(\delta) 2^s \left(\sqrt{2d|\Psi^{(s)}| \log\left(\frac{2K^{(s)}}{d}\right)} + dN \log\left(\frac{K^{(s)}}{d}\right) \right). \tag{76}$$

Since it holds trivially that $K \leq NT$, we then have:

$$|\Psi^{(s)}| \leq \beta(\delta) 2^s \left(\sqrt{2d|\Psi^{(s)}| \log\left(\frac{2NT}{d}\right)} + dN \log\left(\frac{NT}{d}\right) \right). \tag{77}$$

□

Again, we restate the result for completeness.

Theorem E.5. *If $\tau = \max\left\{(d\log(T/d) + 2\log T)/\kappa^2\rho, 1/\rho\right\}$ and $c = \frac{3}{\kappa}\sqrt{2\log(3NT)}$, then Algorithm 2 suffers a regret of*

$$\mathbb{E}[R_T] = O\left(\frac{L}{\kappa^2}\left[\sqrt{dNT}(\log T)^{3/2} + dN\,(\log NT)^{5/2}\right]\right). \tag{78}$$

Proof. The regret is upper bounded as follows:

$$\mathbb{E}[\sum_{t=1}^{T}\sum_{\{i,j\}\in M_t} c_t(i,j) - C_t(M_t^*)] \overset{(1)}{\leq} \underbrace{\sum_{t=1}^{\tau} k_t}_{\textbf{(I)}} + \underbrace{NT\mathbb{P}(A_{\text{MLE}}^{(\bar{S})} \cup A_{\text{init}}^{(\bar{S})})}_{\textbf{(II)}}$$

$$+ \mathbb{P}(A_{\text{MLE}}^{(S)} \cap A_{\text{init}}^{(S)})\mathbb{E}\left[\sum_{t=\tau+1}^{T} C_t(M_t) - C_t(M_t^*) | A_{\text{MLE}}^{(S)} \cap A_{\text{init}}^{(S)}\right] \tag{79}$$

$$\leq N\tau + 2N + \underbrace{\mathbb{E}\left[\sum_{t=\tau+1}^{T} C_t(M_t) - C_t(M_t^*) | A_{\text{MLE}}^{(S)} \cap A_{\text{init}}^{(S)}\right]}_{\textbf{(III)}}. \tag{80}$$

where the first inequality follows from the decomposition into **(I)** the initialization cost and **(II)** the cost incurred by the bad event, where estimation does not behave as expected. From this point onward, to control **(III)**, we assume that $A_{\text{MLE}}^{(S)} \cap A_{\text{init}}^{(S)}$ holds. We also use the same notation as PickMatchingSM (algorithm 7).

$$\sum_{t=\tau+1}^{T} C_t(M_t) \overset{(1)}{\leq} \sum_{t=\tau+1}^{T}\sum_{s=1}^{S} 4\frac{|M_t^{(s)}|}{2^s} + C(M_t^{(S+1)}) \tag{81}$$

$$\overset{(2)}{\leq} \sum_{t=\tau+1}^{T}\sum_{s=1}^{S} 4\frac{|M_t^{(s)}|}{2^s} + C_t(\mathbf{OPT}([k_t])) + \frac{L}{\kappa}\sum_{s=1}^{S} 4\frac{|M_t^{(s)}|}{2^s} \tag{82}$$

where inequality (1) is a mere decomposition using the matchings picked by PickMatchingSM, and inequality (2) follows from Theorem C.3 applied on $M_t^{(S+1)}$. Since $\mathbf{OPT}([k_t]) = M_t^*$, it follows

$$\sum_{t=\tau+1}^{T} C_t(M_t) - C_t(M_t^*) \leq 4(1+\frac{L}{\kappa})\sum_{t=\tau+1}^{T}\sum_{s=1}^{S}\frac{|M_t^{(s)}|}{2^s}. \tag{83}$$

To upper bound that quantity, we group the terms by phase-index s:

$$\sum_{t=\tau+1}^{T}\sum_{s=1}^{S}\frac{|M_t^{(s)}|}{2^s} = \sum_{s=1}^{S}\sum_{t=\tau+1}^{T}\frac{|M_t^{(s)}|}{2^s} \leq \sum_{s=1}^{S}\frac{|\Psi^{(s)}|}{2^s}. \tag{84}$$

Then, we make use of Lemma E.4:

$$\sum_{s=1}^{S}\frac{|\Psi^{(s)}|}{2^s} \leq \beta(\delta)\sum_{s=1}^{S}\left(\sqrt{2d|\Psi^{(s)}|\log\left(\frac{2NT}{d}\right)} + dN\log\left(\frac{NT}{d}\right)\right). \tag{85}$$

Recall that $S \leq \log_2\sqrt{NT}$. We bound the second term trivially:

$$\sum_{s=1}^{S} dN\log\left(\frac{K}{d}\right) \leq \frac{dN}{2}\log\left(\frac{NT}{d}\right)\log(NT) \leq dN(\log NT).^2 \tag{86}$$

For the first term, we proceed using Cauchy-Schwarz inequality as follows:

$$\sum_{s=1}^{S}\sqrt{2d|\Psi^{(s)}|\log\left(\frac{2NT}{d}\right)} \leq \sqrt{2d\log\left(\frac{2NT}{d}\right)}\sum_{s=1}^{S}\sqrt{|\Psi^{(s)}|} \tag{87}$$

$$\leq \sqrt{2d\log\left(\frac{2NT}{d}\right)}\sqrt{S\sum_{s=1}^{S}|\Psi^{(s)}|} \tag{88}$$

$$\leq \sqrt{2dK\log(NT)\log\left(\frac{2NT}{d}\right)} \tag{89}$$

$$\leq \sqrt{2dK}\log(NT). \tag{90}$$

where the last inequality stems from the fact that $d \geq 2$. To summarize, recall the fact that $\frac{L}{\kappa} \geq 1$ and that $\beta(\delta) = \frac{3}{2\kappa}\sqrt{2\log(\frac{3NT}{\delta})}$ and that $\delta = \frac{1}{T}$, hence, for $c = \beta(\delta) = \frac{3}{\kappa}\sqrt{2\log 3NT}$, then

$$\mathbb{E}(R_T) \leq N\tau + 2N + 24\frac{L}{\kappa^2} * \left(\sqrt{2dK}(\log NT)^{3/2} + dN(\log NT)^{5/2}\right). \tag{91}$$

Since $N\tau = O(Nd\log T)$, we conclude that:

$$\mathbb{E}(R_T) = O\left(\frac{L}{\kappa^2}\left[\sqrt{dK}(\log NT)^{3/2} + dN(\log NT)^{5/2}\right]\right). \tag{92}$$

□

F Proof of the Lower Bound

Throughout this section, we fix $F = x \mapsto x + 1/2$, so that the probability that i beats j is $1/2 + \langle \boldsymbol{\theta}^*, z_{i,j}\rangle$. Let $\mathrm{KL}(P||Q)$ be the KL divergence of two probability measures P and Q. In particular, for $p, q \in (0,1)$, we use note $\mathrm{KL}(p||q) := \mathrm{KL}(\mathrm{Ber}(p)||\mathrm{Ber}(q))$. Throughout this section, we make use of the following lemma:

Lemma F.1. *Let $p \in [3/8, 5/8]$ and $\epsilon \leq 1/8$ then*

$$\max\{KL(p+\epsilon||p), KL(p||p+\epsilon)\} \leq \frac{16}{3}\epsilon^2. \tag{93}$$

Proof. The lemma is an immediate consequence of the following inequalities applied to $p \in [3/8, 5/8]$ and $\epsilon \in (0, 1/8)$

1. $\mathrm{KL}(p+\epsilon||p) \leq \frac{\epsilon^2}{p(1-p)}$.
2. $\mathrm{KL}(p||p+\epsilon) \leq \frac{\epsilon^2}{(p+\epsilon)(1-p-\epsilon)}$.

The proof of these inequalities, relies on the following result ((Dragomir et al., 2000) Theorem 1):

$$\mathrm{KL}(p||q) \leq \frac{p^2}{q} + \frac{(1-p)^2}{1-q} - 1. \tag{94}$$

For inequality 1, we have:

$$\begin{aligned}
\mathrm{KL}(p+\epsilon||p) &\leq \frac{(p+\epsilon)^2}{p} + \frac{(1-p-\varepsilon)^2}{1-p} - 1 && (95)\\
&\leq \frac{(1-p)(p+\epsilon)^2 + p(1-p-\epsilon)^2 - p(1-p)}{p(1-p)} && (96)\\
&\leq \frac{p^2+\epsilon^2+2p\epsilon - p^3 - p\epsilon^2 - 2p^2\epsilon + p + p^3 + p\epsilon^2 - 2p^2 - 2p\epsilon + 2p^2\epsilon - p + p^2}{p(1-p)} && (97)\\
&\leq \frac{\epsilon^2}{p(1-p)}. && (98)
\end{aligned}$$

Similarly, the following calculations yield 2

$$\begin{aligned}
\mathrm{KL}(p||p+\epsilon) &\leq \frac{p^2}{p+\epsilon} + \frac{(1-p)^2}{1-p-\epsilon} - 1 && (99)\\
&\leq \frac{p^2 - p^3 - p^2\epsilon + (p+\epsilon)(1+p^2-2p) - (p+\epsilon)(1-p-\epsilon)}{(p+\epsilon)(1-p-\epsilon)} && (100)\\
&\leq \frac{p^2 - p^3 - p^2\epsilon + p + p^3 - 2p^2 + \epsilon + \epsilon p^2 - 2\epsilon p - p + p^2 + p\epsilon - \epsilon + \epsilon p + \epsilon^2}{(p+\epsilon)(1-p-\epsilon)} && (101)\\
&\leq \frac{\epsilon^2}{(p+\epsilon)(1-p-\epsilon)}. && (102)
\end{aligned}$$

□

Proof. Our proof relies on similar techniques as used in (Lattimore and Szepesvári, 2020) (Chapter 24) to derive a lower bound for linear bandit algorithms. The proof is in two steps:

1. We prove first that for $d = 3$, there is an instance where $\boldsymbol{\theta} \in \mathbb{R}^3$ and a set of contexts $\{x_i\}_{i=1,\dots,N}$, for which any algorithm would yield a regret of at least $\sqrt{NT}$. This is detailed in Lemma F.2.

2. Duplicate this instance, on $d/3$ orthogonal subspaces of $\mathbb{R}^d$, splitting the contexts among the subspaces, $\frac{3N}{d}$ each. This yields $d/3$ independent problems where each incurs a regret of $\Omega(\sqrt{\frac{N}{d}T})$ and therefore a total regret of at least $\Omega(\sqrt{dNT})$.

□

Lemma F.2. *Let $\boldsymbol{\theta} = (\frac{1}{\sqrt{d}}, \theta_2, \theta_3)$, where $(\theta_2, \theta_3) \in \{(0,\epsilon),(\epsilon,0)\}$, for $\epsilon = \frac{1}{\sqrt{NT}}$. Let $\boldsymbol{x}_{4k+r} = (k/(2N), x_{1,r}, x_{2,r})$, where $x_{1,r} = r(\textbf{mod } 2)$ and $x_{2,r} = \lfloor r/2 \rfloor$. Then, for $T > 4Nd$ any algorithm incurs a regret of at least:*

$$\Omega(\sqrt{NT}). \tag{103}$$

Proof. First, notice that the set of contexts has the following property: if $k < k' \in [0, N/4]$ then for any $r, r' \in \{0,1,2,3\}$: we have $\mu_{4k+r} \leq \mu_{4k'+r'}$. This indicates that optimal matching is, in fact, the same matching duplicated on each set of contexts $\{x_{4k+r}\}_{r=1,2,3,4}$. Suppose $4k + r$ and $4k' + r'$ are matched together, that pair provides an information of $\mathrm{KL}(p + \epsilon || p)$ or $\mathrm{KL}(p || p + \epsilon)$ depending on the value of $\boldsymbol{\theta}$, where $p = \frac{1}{2} + \frac{k-k'}{2N\sqrt{d}} \in [\frac{3}{8}, \frac{5}{8}]$. Either way, Lemma F.1 implies that the information provided is at most $\frac{16}{3}\epsilon^2$. On the other hand, the information provided by a pair where $k = k'$ is at least $\min\{\mathrm{KL}(\frac{1}{2} + \epsilon | \frac{1}{2}), \mathrm{KL}(\frac{1}{2} | \frac{1}{2} + \epsilon)\} \geq 2\epsilon^2$ (Pinsker Inequality), which implies that both pairs are similarly informative (up to a universal constant $C \leq 8/3$). However, the pair $\{4k + r, 4k' + r'\}$ incurs a regret at least $\frac{|k-k'|}{2N\sqrt{d}}$ whereas if $k = k'$, the regret incurred is at most $\epsilon = \frac{1}{\sqrt{NT}}$. Hence, it cost at least $\sqrt{\frac{T}{Nd}}$ more to sample that a pair where $k \neq k'$ compared to one where $k = k'$, for comparable informativeness. Consequently, it is not interesting for any algorithm to pick pairs where $k \neq k'$.

Going forward, we assume that the algorithm we consider $\mathbb{A}$ does not sample pairs such that $k \neq k'$, i.e. the matching (M_t) picked by $\mathbb{A}$ can be divided into $N/4$ matchings $(M_{t,k})_k$, where for each $k \in [N/4]$, $M_{t,k}$ is a matching on $\{x_{4k+r}\}_{r=1,2,3,4}$.

For the remainder of the proof, we set $k = 0$, as the same argument would translate seamlessly for other values of k. We also omit the first coordinate of the contexts in this case as it is 0. There is a unique optimal, depending on the value of $\boldsymbol{\theta}^* \in \{(0,\epsilon),(\epsilon,0)\}$:

- if $\boldsymbol{\theta}^* = (0,\epsilon) = \boldsymbol{\theta}_1$, then $M^* = M_1 := \{\{(1,1),(0,1)\},\{(0,0),(1,0)\}\}$.
- if $\boldsymbol{\theta}^* = (\epsilon,0) = \boldsymbol{\theta}_2$, then $M^* = M_2 := \{\{(1,1),(1,0)\},\{(0,0),(0,1)\}\}$.

In both cases, the cost of the optimal matching is 0. There is a third matching, $M_3 := \{\{(1,1),(0,0)\},\{(1,0),(0,1)\}\}$, never optimal. the cost of suboptimal matchings, in either cases, is 2ϵ.

For any $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$, we denote by $\mathbb{P}_{\boldsymbol{\theta}}$ the measure on outcomes induced by the instance and algorithm $\mathbb{A}$, we also denote by $M^*_{\boldsymbol{\theta}}$ the optimal matching for $\boldsymbol{\theta}$. Hence

$$\mathbb{E}_{\boldsymbol{\theta}}[R_T(\mathbb{A})] = 2\epsilon\frac{N}{4}\mathbb{E}_{\boldsymbol{\theta}}[\sum_{t=1}^{T} \mathbb{1}_{M_t \neq M^*_{\boldsymbol{\theta}}}] \tag{104}$$

$$\geq \frac{\epsilon NT}{2}\mathbb{P}_{\boldsymbol{\theta}}(\sum_{t=1}^{T} \mathbb{1}_{M_t \neq M^*_{\boldsymbol{\theta}}} \geq \frac{T}{2}) \tag{105}$$

$$= \frac{\epsilon NT}{2}\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{E}_{\boldsymbol{\theta}}). \tag{106}$$

Where we have used the Markov inequality, and $\mathcal{E}_{\boldsymbol{\theta}} = \{\sum_{t=1}^{T} \mathbb{1}_{M_t \neq M^*_{\boldsymbol{\theta}}} \geq \frac{T}{2}\}$.

Let $\boldsymbol{\theta}' = (\epsilon,\epsilon) - \boldsymbol{\theta}$ the alternative weight vector. Since $\mathcal{E}^c_{\boldsymbol{\theta}} \subset \mathcal{E}_{\boldsymbol{\theta}'}$, then $\mathbb{P}_{\boldsymbol{\theta}'}(\mathcal{E}_{\boldsymbol{\theta}'}) \geq \mathbb{P}_{\boldsymbol{\theta}'}(\mathcal{E}^c_{\boldsymbol{\theta}})$. This observation relates the regret to the Bretagnolle-Huber inequality ((Lattimore and Szepesvári, 2020) Chapter 14). It yields:

$$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{E}_{\boldsymbol{\theta}}) + \mathbb{P}_{\boldsymbol{\theta}'}(\mathcal{E}_{\boldsymbol{\theta}'}) \geq \mathbb{P}_{\boldsymbol{\theta}^*}(\mathcal{E}) + \mathbb{P}_{\boldsymbol{\theta}'}(\mathcal{E}^c) \geq \frac{1}{2}\exp -\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}^*} || \mathbb{P}_{\boldsymbol{\theta}'}) \tag{107}$$

where $\mathrm{KL}(P||Q)$ refers to the KL-divergence of two probability measures P and Q. It remains to decompose the KL appropriately to derive the lower bound. Using ((Lattimore and Szepesvári, 2020) Exercise 15.8(b)), we get that:

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}^*}||\mathbb{P}_{\boldsymbol{\theta}'}) = \sum_{t=1}^{T}\sum_{i=1}^{N/2}\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}^*}(Y_{t,i})||\mathbb{P}_{\boldsymbol{\theta}'}(Y_{t,i})). \tag{108}$$

Where $(Y_{t,i})_i$ refer to the outcomes of the pairs in the matching M_t. Given that $\mathbb{A}$ either picks M_1, M_2 or M_3, there is only three KL-divergence to compute, namely:

- $\mathrm{KL}(\frac{1}{2}+\epsilon||\frac{1}{2}) \leq \frac{16}{3}\epsilon^2$ (Lemma F.1).
- $\mathrm{KL}(\frac{1}{2}||\frac{1}{2}+\epsilon) \leq \frac{16}{3}\epsilon$ (Lemma F.1).
- $\mathrm{KL}(\frac{1}{2}-\epsilon||\frac{1}{2}+\epsilon) \leq \frac{64}{3}\epsilon^2$ (Lemma F.1 for $p=\frac{1}{2}-\epsilon$ and $\tilde{\epsilon}=2\epsilon$).

It follows that

This covers all possible pair selected by M_t, hence

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}^*}||\mathbb{P}_{\boldsymbol{\theta}'}) \leq \frac{64}{3}NT\varepsilon^2. \tag{109}$$

In conclusion:

$$\mathbb{E}_{\boldsymbol{\theta}}[R_T(\mathbb{A})] + \mathbb{E}_{\boldsymbol{\theta}'}[R_T(\mathbb{A})] \geq \frac{\epsilon NT}{2}\left(\mathbb{P}_{\boldsymbol{\theta}}\left(\mathcal{E}_{\boldsymbol{\theta}}\right) + \mathbb{P}_{\boldsymbol{\theta}'}\left(\mathcal{E}_{\boldsymbol{\theta}'}\right)\right) \tag{110}$$

$$\geq \frac{\epsilon NT}{4}\exp(-\frac{64}{3}\epsilon^2 NT). \tag{111}$$

This implies that, for either $\boldsymbol{\theta}$ or $\boldsymbol{\theta}'$, the regret is larger than $\frac{\epsilon NT}{8}\exp(\frac{64}{3}\epsilon^2 NT)$. For the choice of $\epsilon = \frac{1}{\sqrt{NT}}$, we achieve the claim. □