Enjoying Non-linearity in Multinomial Logistic Bandits

Pierre Boudart, Pierre Gaillard, Alessandro Rudi

ENJOYING NON-LINEARITY IN MULTINOMIAL LOGISTIC BANDITS: A MINIMAX-OPTIMAL ALGORITHM

A PREPRINT

Pierre Boudart
INRIA, École Normale Supérieure
CNRS, PSL Research University
Paris, France
pierre.boudart@inria.fr

Pierre Gaillard
Univ. Grenoble Alpes, Inria,
CNRS, Grenoble INP, LJK
Grenoble, France
pierre.gaillard@inria.fr

Alessandro Rudi
SDA Bocconi School of Management
Milano, Italy
alessandro.rudi@sdabocconi.it

October 7, 2025

ABSTRACT

We consider the multinomial logistic bandit problem, a variant of where a learner interacts with an environment by selecting actions to maximize expected rewards based on probabilistic feedback from multiple possible outcomes. In the binary setting, recent work has focused on understanding the impact of the non-linearity of the logistic model (Faury et al., 2020; Abeille et al., 2021). They introduced a problem-dependent constant $\kappa_* \geq 1$, that may be exponentially large in some problem parameters and which is captured by the derivative of the sigmoid function. It encapsulates the non-linearity and improves existing regret guarantees over T rounds from $O(d\sqrt{T})$ to $O(d\sqrt{T/\kappa_*})$, where d is the dimension of the parameter space. We extend their analysis to the multinomial logistic bandit framework, making it suitable for complex applications with more than two choices, such as reinforcement learning or recommender systems. To achieve this, we extend the definition of κ_* to the multinomial setting and propose an efficient algorithm that leverages the problem's non-linearity. Our method yields a problem-dependent regret bound of order $\widetilde{\mathcal{O}}(Rd\sqrt{KT/\kappa_*})$, where R is the norm of the vector of rewards and K is the number of outcomes. This improves upon the best existing guarantees of order $\widetilde{\mathcal{O}}(RdK\sqrt{T})$. Moreover, we provide a $\Omega(Rd\sqrt{KT/\kappa_*})$ lower-bound, showing that our algorithm is minimax-optimal and that our definition of κ_* is optimal.

1 Introduction

We consider the multinomial logistic (MNL) bandit problem, that unfolds as follows. At each round $t \geq 1$, a learner chooses an action $x_t \in \mathcal{X}$ from an action set $\mathcal{X} \subset \mathbb{R}^d$. Then, the environment samples an outcome $y_t \in [\![K]\!]$ from the distribution $\mu(\theta_* x_t) \in \Delta_K$, where $\theta_* \in \mathbb{R}^{K\times d}$ is an unknown parameter to be estimated and $\mu : \mathbb{R}^K \to \Delta_K$ the softmax function. At the end of the round, the learner receives the reward $r_t := \rho_{y_t}$, where $\rho \in \mathbb{R}_+^K$ is a known vector that associates a reward to each output. The goal of the learner is to minimize their expected regret defined as follows

$$\mathrm{Reg}_T := \textstyle\sum_{t=1}^T \rho^\top \big(\mu(\theta_* x_*) - \mu(\theta_* x_t)\big), \qquad \text{where} \quad x_* \in \arg\max_{x\in\mathcal{X}} \rho^\top \mu(\theta_* x)\,.$$

The MNL bandit problem falls into the umbrella of stochastic bandit frameworks (Robbins, 1952, Thompson, 1933), which studies decision-making processes with exploration-exploitation dilemma. Linear bandits (Lattimore and Szepesvári, 2020) model a linear relationship between actions $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ and rewards $r_t \in \mathbb{R}$. They have been used with success in various applications. However they fail to model complex systems with non-linear rewards. This called for the introduction of the Generalised Linear Model (GLM) framework (Filippi et al.,

2010). In GLMs the reward associated to an action $x_t \in \mathcal{X}$ is $\mu(\theta_* x_t)$ where θ_* is a parameter unknown to the learner and μ is a non linear function. The logistic bandit framework is an example of GLM obtained by choosing μ as the sigmoid function $\mu(z) = 1/(1+\exp(-z))$. It allows to model situations where evaluated by a success/failure feedback, e.g. click/no-click in add-recommendation systems.

The MNL bandit framework (Amani and Thrampoulidis, 2021) is a natural extension of it. It allows to model situations with more than two outcomes. For instance consider a recommendation system on a e-commerce website. The user has several options, he may choose 1) to buy now; 2) add to the cart; 3) add to the wish-list; 4) click on "do not recommend"; 5) do not click; 6) leave the website, etc. The probability of each outcome is modelled by the softmax function $\mu : \mathbb{R}^K \to [0,1]^K$, see Section 2 for a formal definition. In this framework each outcome is associated with a specific reward $\rho_k \geq 0$. The goal of the learner is to give recommendations that maximise the expected reward of the outcome. Note that the MNL bandit problem is not a GLM, but a multi-index model (Xia, 2008).

Related work A key aspect of the MNL bandit problem arises from the non-linearity of the reward. In the binary case, where $K = 2$ and μ is the sigmoid function, some works (Abeille et al., 2021, Faury et al., 2020, 2022, Jun et al., 2021) have focused on better understanding its impact on regret. Interestingly, this effect was shown to be captured by the constant $\kappa := 1/\min_{\|\theta\|_2 \leq S} \min_{x \in \mathcal{X}} \dot{\mu}(\theta x)$, where S is an upper-bound on $\|\theta_*\|_2$, introduced by Filippi et al. (2010), who demonstrated a regret of order $\widetilde{O}(d\kappa\sqrt{T})$. The constant κ can be understood as measuring the error incurred when making a linear approximation of the logistic model. Notably, κ may be exponentially large in S and the diameter of $\mathcal{X}$, suggesting that non-linearity significantly worsens the regret guarantees compared to the linear bandit. Consequently, subsequent work has focused on improving the dependence on κ. Faury et al. (2020) demonstrated that the non-linearity of the problem, i.e., κ, is not detrimental asymptotically, achieving a regret bound of order $\widetilde{O}(d\sqrt{T})$. Even more strikingly, Abeille et al. (2021) showed that one can leverage non-linearity to an advantage. They proved a regret bound scaling as $\widetilde{O}(d\sqrt{T/\kappa_*})$, where $\kappa_* := 1/\dot{\mu}(\theta_* x_*)$ measures the non-linearity at the optimum. This result represents a dramatic improvement, as in the most favorable cases, we have $\kappa_* \approx \kappa$. Moreover, they established that this bound is minimax optimal by deriving a $\Omega(d\sqrt{T/\kappa_*})$ problem dependent lower-bound. It is important to note that the constants κ and κ_* are indeed problem-dependent, as they are influenced by S, $\mathcal{X}$, and θ_*.

The MNL setting, which considers a reward vector $\rho \in \mathbb{R}^K$ with $K \geq 2$ outputs, whose norm is denoted by $\|\rho\|_2 = R$, and where μ is the softmax function, was introduced by Amani and Thrampoulidis (2021). They proposed a tractable algorithm that achieves a regret upper bound of order $\widetilde{O}(RdK\sqrt{\kappa T})$, where κ is a generalization of the binary setting constant defined as follows[1]

$$\kappa^{-1} := \min_{\|\theta\|_2 \leq S} \min_{x \in \mathcal{X}} \lambda_{K-1}(\nabla\mu(\theta x))\,. \tag{1}$$

Interestingly, they also provided a non-tractable algorithm with a regret scaling as $\widetilde{O}(RdK^{3/2}\sqrt{T})$. This indicates that the asymptotic dependence on κ can also be eliminated in the MNL framework, but the question of whether this can be achieved efficiently remained open. This question was recently addressed by Zhang and Sugiyama (2024), who designed an efficient algorithm that achieves a regret of order $\widetilde{O}(RdK\sqrt{T})$. An open question persists: Is it possible to extend the result of Abeille et al. (2021) in the MNL setting and demonstrate that the non-linearity indeed yields improved asymptotic regret?

Main contributions In this paper, we answer the above open question positively. To quantify the non-linearity of the problem at the optimum in the multinomial setting, we generalize the problem-dependent constant κ_* as follows:

$$\kappa_* = \frac{\|\rho\|_2^2}{\rho^\top \nabla\mu(\theta_* x_*)\rho} \text{ when } \rho \notin \mathbb{R}1_K \text{ and } \kappa_* = +\infty \text{ when } \rho \in \mathbb{R}1_K \tag{2}$$

where the definition of κ_* for $\rho \in \mathbb{R}1_K$ is given by a continuity extension. Note that this constant also depends on the reward vector ρ. As learners are expected to eventually play actions close to the optimum, κ_* quantifies the level of non-linearity of the reward signal in the long-term regime. We introduce a new algorithm (Algorithm 2) with a regret upper-bound given by (Theorem 3):

$$\mathrm{Reg}_T \leq O\left(Rd\sqrt{KT/\kappa_*}\log(T/\delta)\right) \quad \text{w.p., } 1-2\delta.$$

In some cases, κ_* can be as large as $\exp(S\max_{x\in\mathcal{X}}\|x\|_2)$, see Appendix B.2, thereby significantly improving existing asymptotic results on MNL bandits. We prove that our regret upper-bound is minimax-optimal and that our choice of κ_* is optimal (up to log factors) by deriving in Theorem 4 the following regret lower-bound $\mathrm{Reg}_T \geq \Omega\big(Rd\sqrt{KT/\kappa_*}\big)$.

We summarize existing algorithms, that focus on the dependence κ and κ_*, for binary and MNL bandits in Table 1.

[1] the constant κ is originally defined slightly differently in Amani and Thrampoulidis (2021) but the definitions are equivalent up to constant factors

Setting	Algorithm	Regret	Comput. per Iter.
Binary	GLM-UCB Filippi et al. (2010)	$d\kappa\sqrt{T}$	$O(t)$
	Logistic-UCB-1 Faury et al. (2020)	$d\sqrt{\kappa T}$	$O(t)$
	Logistic-UCB-2 Faury et al. (2020)	$d\sqrt{T} + \kappa d^2$	$O(t)$
	OFULog Abeille et al. (2021)	$d\sqrt{T/\kappa_*} + \kappa d^2$	$O(t)$
	OFU-ECOLog Faury et al. (2022)	$d\sqrt{T/\kappa_*} + \kappa d^2$	$O(\log^2(t))$
	OFUL-MLogB Zhang and Sugiyama (2024)	$d\sqrt{T/\kappa_*} + \kappa d^2$	$O(1)$
Multinomial	MNL-UCB Amani and Thrampoulidis (2021)	$RdK\sqrt{\kappa T}$	$O(t)$
	Improved MNL-UCB Amani and Thrampoulidis (2021)	$RdK^{3/2}\sqrt{T} + \kappa K^2 d$	-
	MNL-UCB+ Lee et al. (2024)	$Rd\sqrt{K\kappa T}$	$O(t)$
	Improved MNL-UCB+ Lee et al. (2024)	$Rd\sqrt{KT} + \kappa d^2 K^2$	-
	OFUL-MLogB Zhang and Sugiyama (2024)	$RdK\sqrt{T} + \kappa K^{3/2} d^2$	$O(1)$
	***REAL* (ours) - Upper Bound (Alg. 2)**	$Rd\sqrt{KT/\kappa_*} + \kappa K^2 d^2$	$O(1)$
	This work - Lower Bound (Thm 4)	$\Omega(Rd\sqrt{KT/\kappa_*})$	

Table 1: Comparison of regret bounds for logistics and multinomial bandits, with respect to $R, d, K, \kappa, \kappa_*$ and T. For simplicity we omit logarithmic terms and other constants. For the computation cost of each algorithm we only provide the dependence in t, - signifies untractable.

The algorithm we introduce (Algorithm 2) is computationally efficient, with a per round complexity of order $O(1)$. A central component of our theoretical analysis involves applying the self-concordance property without incurring exponential sub-optimal factors. To this end, our algorithm first performs an exploration phase (Algorithm 1) to design a sufficiently small high-probability confidence set Θ around θ_*, where the self-concordance property can be applied with only a constant factor penalty (see Section 3.1). Once Θ is designed, the algorithm continues to improve its estimate of θ_* by running a variant of Online Mirror Descent (OMD) constrained within Θ only. As emphasized by Zhang and Sugiyama (2024), a central difficulty in the regret analysis lies in controlling the term $\sum_t \rho^\top \nabla\mu(\theta_* x_t)\rho$. Ideally, if $x_t \to x_*$ quickly as $t \to \infty$, this term will be of the order $\sum_t \rho^\top \nabla\mu(\theta_* x_*)\rho = R^2 T/\kappa_*$, leading to the final improvement in the regret. A key technical contribution of our analysis is to address this challenge by carefully leveraging the structure of the softmax function and employing the self-concordance properties within Θ.

Multinomial *Logit* Bandits A different line of work is the Multinomial Contextual *Logit* Bandit problem (Agrawal et al., 2023, 2017, 2019, Cheung and Simchi-Levi, 2017, Dong et al., 2020), a combinatorial variant of MNL bandits that generalizes the binary logistic problem differently. At each round t, the learner is asked to choose a subset of actions $S_t \subset [\![K]\!]$ based on observed contextual vectors $x_{t,i} \in \mathcal{X}$ for $i \in [\![K]\!]$ and rewards $\rho_{t,i} \in \mathbb{R}_+$. The goal of the learner is to maximise the expected reward modeled by the multinomial *logit* model $\mathbb{E}[r_t \mid S_t] = \sum_{i\in S_t} \rho_{t,i} \exp(\theta_*^\top x_{t,i}) / \big(1 + \sum_{i\in S_t} \exp(\theta_*^\top x_{t,i})\big)$, restricted to the subset S_t of chosen actions only and where $\theta_* \in \mathbb{R}^d$ is a parameter unknown to the learner. Although it may appear similar, this framework is fundamentally different: the settings differ in their parameterisation, feedback structure, and modeling assumptions. It appears that neither framework can be easily reduced to the other. In particular, the combinatorial nature of the *Logit* framework-namely, the selection of a subset S_t-together with the normalization in the softmax function makes any such reduction highly challenging. Moreover, in our framework, every outcome has a nonzero probability of being selected. We provide further details in Appendix D. This variant also exhibits similar challenges related to the non-linearity of the rewards and the constants κ, κ_*. Agrawal et al. (2023) introduced an algorithm with $O(d\sqrt{T})$ regret bounds, for which the leading term is independent of κ, representing a significant improvement over the previous bound of $O(d\sqrt{\kappa T})$. In the case of uniform rewards, i.e., $\rho_{t,i} = 1$ for all $t \in [\![T]\!]$ and all $i \in [\![K]\!]$, Perivier and Goyal (2022) further established a bound of $\widetilde{O}(d\sqrt{T/\kappa_*})$. More recently, Lee and Oh (2025) proposed an algorithm that achieves a $\mathrm{poly}(S)$-free regret of $\widetilde{O}(d\sqrt{T/\kappa_*})$ by employing adaptive exploration to exploit self-concordance. Until now, both frameworks have been studied separately; establishing connections between them would be an interesting direction for future work.

2 Problem Formulation

In this section, we introduce our notations and assumptions and formally recall the setting of MNL bandits.

Notations Let $1_K \in \mathbb{R}^K$ be the vector of 1's and $\mathcal{H}$ be the hyperplane supported by 1_K. We denote by $\Pi : \mathbb{R}^K \to \mathbb{R}^K$ the projection on $\mathcal{H}$. We denote by Δ_K the K dimensional simplex and by $\mu : \mathbb{R}^K \to \Delta_K$ the softmax function defined by $\mu(z)_k \propto \exp(z_k)$ for all $k \in [\![K]\!]$.

Framework The MNL bandit framework is formalised as a game of $T \in \mathbb{N}$ rounds between an learner and an environment, see Framework 1 for a short summary. At each round $t \in [\![T]\!]$, the learner plays an action $x_t \in \mathcal{X}$ from an action set $\mathcal{X} \subseteq \mathbb{R}^d$. Then, the learner observes the output of the environment $y_t \in [\![K]\!]$ with $K \in \mathbb{N}$, that are generated using the softmax function. More precisely, for all $k \in [\![K]\!]$, we have $\mathbb{P}[y_t = k|x_t] := \mu(\theta_* x_t)_k$ where $\theta_* \in \Pi\mathbb{R}^{K\times d}$ is a parameter of the environment unknown to the learner such that $\|\theta_*\|_2 \leq S$. At the end of each round t, the learner receives a reward ρ_{y_t} associated with the environment output y_t, from a fixed and known beforehand reward vector $\rho \in \mathbb{R}_+^K$, $\|\rho\|_2 = R$. The goal of the learner is to maximise their expected reward which is equivalent to minimising the expected regret

$$\mathrm{Reg}_T := \sum_{t=1}^{T} \rho^\top \mu(\theta_* x_*) - \rho^\top \mu(\theta_* x_t)$$

where $x_* := \arg\max_{x\in\mathcal{X}} \rho^\top \mu(\theta_* x)$ is the action maximising the expected reward.

Note that our framework differs from the original one of Amani and Thrampoulidis (2021): instead of fixing one line of θ_* to be zero, we assume that it is such that $\sum_{k=1}^K [\theta_* x]_k = 0$ for any $x \in \mathcal{X}$. This is ensured by the fact that $\theta_* \in \Pi\mathbb{R}^{K\times d}$, which can be assumed without loss of generality since for any $\theta \in \mathbb{R}^{K\times d}$ and $x \in \mathbb{R}^d$ the probability vector of outcomes satisfies $\mu(\theta x) = \mu(\Pi\theta x)$. Hence our model is more general, as it does not assume the existence of a dedicated no-choice (NC) item; however, such an option can be naturally incorporated by assigning a reward of 0 to any item, effectively allowing users to choose nothing. Unlike existing literature, which often makes the strong and sometimes unnecessary assumption of a universally applicable NC item, our approach removes this constraint. While NC is appropriate in certain domains—such as e-commerce, online ads, or web search, where users frequently choose nothing—it is not suitable across the board. In some applications, NC isn't even feasible. For example, large language models often require explicit user preferences to proceed. Likewise, in robotics, autonomous driving, or preference-based reinforcement learning (PbRL), human feedback must indicate a choice among alternatives to guide training—NC is not an option. In Appendix C, we show that the framework of Amani and Thrampoulidis (2021) is included into ours.

Framework 1: The Multinomial Logistic (MNL) Bandit Framework.

for *Each time step* t *in* $1 \dots T$ **do**
  Play action $x_t \in \mathcal{X}$
  Observe the decision of the environment $y_t \in [\![K]\!]$ such that $\mathbb{P}[y_t = k|x_t] = \mu(\theta_* x_t)_k$
  Get reward ρ_{y_t}
end

Problem-dependent constants κ and κ_* As detailed in the introduction, a key aspect of the MNL bandit framework, compared to standard stochastic linear bandits, arises from the non-linearity of $\mu(\cdot)$, which appears both in the stochastic feedback model and in the reward definition. Earlier works Abeille et al. (2021), Amani and Thrampoulidis (2021), Filippi et al. (2010), Zhang and Sugiyama (2024) demonstrated that this non-linearity could be captured by two problem-dependent constants, κ and κ_*, respectively defined in Equations (1) and (2), where our work introduces a new formulation of κ_*. On the one hand, κ quantifies the cost of performing linear approximations within the MNL framework, with larger values of κ leading to increased regret. On the other hand, κ_* measures the curvature at the optimum, which can be exploited in the long run to improve the asymptotic regret. Note that κ is defined as the inverse of the second smallest eigenvalue of the gradient, since the smallest eigenvalue is 0 and corresponds to the eigenvector 1_K composed of ones. Our definitions of κ slightly differ from existing one due to differences in our framework notations, but they coincide with the existing definitions (see Appendix C for details) up to constant factors. In particular, the constant κ is shown in Appendix B.1 to be bounded from below and above as follows:

$$\frac{\exp(-2SX)}{K} \leq \min_{\|\theta\|_2\leq S} \min_{x\in\mathcal{X}} \lambda_{K-1}(\nabla\mu(\theta x)) \leq \frac{2\exp(-2SX)}{2\exp(-SX) + (K-2)\exp(SX)}\,. \tag{3}$$

Hence, κ is exponentially large with respect to $S \geq \|\theta_*\|$ and $X := \max_{x\in\mathcal{X}} \|x\|_2$. The nonzero eigenvalues of the gradient of μ can therefore be as small as κ^{-1}. Consequently, a naive linear approximation of the MNL framework to apply standard linear stochastic bandit analysis results in a suboptimal regret bound factor of κ, which becomes extremely large for large values of X and S.

Assumptions We use the following assumptions, which are classical in the literature (Amani and Thrampoulidis, 2021, Zhang and Sugiyama, 2024).

- The norm of each action is bounded by 1: for all $x \in \mathcal{X}$, $\|x\|_2 \leq 1$.

- The reward vector $\rho \in \mathbb{R}_+^K$ satisfies $\|\rho\|_2 = R$ and is known.
- The norm of the parameter $\theta_* \in \mathbb{R}^{K\times d}$ is bounded by S: $\|\theta_*\|_2 \leq S$. The bound S is known.
- For all $x \in \mathcal{X}$ and for all θ such that $\|\theta\|_2 \leq S$, we assume

$$\lambda_{K-1}(\nabla\mu(\theta x)) \geq \frac{1}{\kappa} > 0 \qquad \text{and} \qquad \lambda_1(\nabla\mu(\theta x)) \leq 1, \tag{4}$$

where λ_{K-1} and λ_1 denote, respectively, the second smallest and the largest eigenvalues.

Note that the assumption $\max_{x\in\mathcal{X}} \|x\|_2 \leq 1$ is made without loss of generality. Indeed, the norm of the inputs can be transferred to the norm of θ_*.

Additional Notations Given a compact set Θ, we define its diameter under an action set $\mathcal{X}$ as

$$\mathrm{diam}_{\mathcal{X}}(\Theta) := \max_{x\in\mathcal{X}} \max_{\theta_1,\theta_2\in\Theta} \|(\theta_1 - \theta_2)x\|_2 \,.$$

We denote by $\mathfrak{C}$ a universal constant, i.e., a constant independent of $S, d, K, T, R, \kappa, \kappa_*$. The notation $\lesssim$ indicates an inequality up to a universal constant. We define the filtration $\mathcal{F}_t := \{x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t\}$. Throughout the paper, the index t refers to measurability with respect to $\mathcal{F}_t$, but not with respect to $\mathcal{F}_{t-1}$. We denote by ℓ_{t+1} the logistic loss associated with the pair (x_t, y_t), defined as follows: for all $\theta \in \mathbb{R}^{K\times d}$,

$$\ell_{t+1}(\theta) := \sum_{k=1}^{K} -1[k = y_t] \log(\mu(\theta x_t)_k) \,.$$

3 Algorithm and Regret Analysis

In this section, we introduce our algorithm (see Algorithm 2) and derive a bound on its regret. The algorithm follows the explore-and-learn paradigm. Following the idea of Abeille et al. (2021) for binary logistic bandits, the first exploration phase aims to design a sufficiently small confidence set Θ around θ_*. In the second phase, the algorithm continues to improve the estimation of θ_* while choosing the action x_t optimistically.

3.1 Exploration Routine

We first introduce our exploration routine (see Algorithm 1) and discuss the main challenges associated with it. This exploration routine is then used as an initialisation phase in our main algorithm (see Algorithm 2).

Algorithm 1: EXPLORATION_ROUTINE

Input: Length of the procedure τ, regularisation parameter λ_0
Init: $V_0 = \lambda_0 I_{Kd}$
for *each round* t *in* $1 \ldots \tau$ **do**
  Choose action $x_t \in \arg\max_{x\in\mathcal{X}} \|I_K \otimes x\|_{V_{t-1}^{-1}}$
  Observe $y_t \sim \mu(\theta_* x_t)$
  Get reward ρ_{y_t}
  Update $V_t = V_{t-1} + \frac{1}{\kappa} I_K \otimes x_t x_t^\top$
end
$\widehat{\theta}_{\tau+1} = \arg\min_{\theta\in\mathbb{R}^{K\times d}} \sum_{s=1}^{\tau} \ell_s(\theta) + \frac{\lambda_0}{2}\|\theta\|^2$
Output: $\Theta := \{\theta \in \Pi\mathbb{R}^{K\times d} : \|\theta - \widehat{\theta}_{\tau+1}\|_{V_\tau}^2 \leq 84^2\lambda_0\}$

The goal of the exploration routine (see Algorithm 1) is to produce a confidence set Θ such that $\theta_* \in \Theta$ with high probability and $\mathrm{diam}_{\mathcal{X}}(\Theta) \leq 1$. This enables us to leverage the self-concordance property (Sun and Tran-Dinh, 2019, Proposition 8) of the logistic function without incurring an exponential constant. Consequently, for all $x \in \mathcal{X}$, we have w.h.p.:

$$\nabla\mu(\theta_1 x) \leq \exp(\sqrt{6}\mathrm{diam}_{\mathcal{X}}(\Theta))\nabla\mu(\theta_2 x) \leq e\nabla\mu(\theta_2 x) \quad , \forall\theta_1, \theta_2 \in \Theta \,.$$

The following lemma shows that such a set Θ can be obtained with a reasonably small exploration length τ. The proof is deferred to Appendix E.1.2.

Lemma 1. *Let $\delta \in (0, 1]$, $\lambda_0 = (S+1)Kd\log(T/\delta)$ and $\tau = 336^2\lambda_0\kappa Kd\log(T)$. Then, the set Θ returned by Algorithm 1 satisfies with probability $1-\delta$*

$$\theta_* \in \Theta \qquad \textit{and} \qquad \mathrm{diam}_{\mathcal{X}}(\Theta) \leq 1/\sqrt{6} \,.$$

As $\mathcal{X}$ and S are known to the learner, κ can, in principle, be computed (see Equation (1)). An upper-bound can also be obtained from Equation (3), which is tight up to a constant factor.

3.2 Learning Routine

We introduce the core of our algorithm, which leverages the exploration routine (see Algorithm 2). To select an action, we use the Optimism in the Face of Uncertainty (OFU) paradigm, a fundamental approach in bandit algorithms to address the exploration-exploitation trade-off. At each time step t, the learner selects an action according to the rule

$$x_t \in \arg\max_{x \in \mathcal{X}} \widetilde{r}_t(x),$$

where $\widetilde{r}_t(x)$ is an optimistic reward that upper-bounds the expected reward $\rho^\top \mu(\theta_* x)$. In the context of logistic bandits, a common approach for defining $\widetilde{r}_t(x)$ is to construct a confidence set $\mathcal{C}_t(\delta)$ at each round t around θ_* and define

$$\widetilde{r}_t(x) := \max_{\theta \in \mathcal{C}_t(\delta)} \rho^\top \mu(\theta x). \tag{5}$$

However, this formulation results in a non-concave maximization problem, which can be computationally challenging to solve. To overcome this difficulty, we adapt the optimistic reward proposed by Zhang and Sugiyama (2024) (see their Proposition 2) who, instead of directly maximizing over the confidence set, directly express $\widetilde{r}_t(x)$ in closed-form from an estimate of θ_* to which they add some bonus. We adapt their estimate by defining a new one θ_t that lies within the confidence set Θ returned by the EXPLORATION_ROUTINE procedure (see Equation (7)). Our estimate θ_t is obtained by solving the following quadratic problem:

$$\theta_t = \arg\min_{\theta \in \Theta} \langle \nabla \ell_{t+1}(\theta_t), \theta \rangle + \tfrac{1}{2\eta} \|\theta - \theta_t\|^2_{\widetilde{W}_t}, \tag{6}$$

where $\widetilde{W}_t := \sum_{s=1}^{t-1} \nabla\mu(\theta_{s+1} x_s) \otimes x_s x_s^\top + \eta \nabla\mu(\theta_t x_t) \otimes x_t x_t^\top + \lambda I_{Kd}$, with $\eta > 0$ a parameter of the algorithm. Our optimistic reward $\widetilde{r}_t(x)$ is then obtained through a Taylor expansion of μ and defined as follows. For all $t \geq T$ and $x \in \mathcal{X}$, we set

$$\widetilde{r}_t(x) := \rho^\top \mu(\theta_t x) + \varepsilon_{1,t}(x) + \varepsilon_{2,t}(x)\,, \tag{7}$$

where

$$\varepsilon_{1,t}(x) := \sigma_t(\delta) \left\| \overline{W}_t^{-1/2} (I_K \otimes x) \nabla\mu(\theta_t x) \rho \right\|_2 \quad \text{and} \quad \varepsilon_{2,t}(x) := 3R\sigma_t(\delta)^2 \left\| (I_K \otimes x^\top) \overline{W}_t^{-1/2} \right\|_2^2 .$$

Here, $\overline{W}_t = W_t + \sum_{s=1}^t 1_K 1_K^\top \otimes x_s x_s^\top$, and $\sigma_t(\delta)$ is a confidence term defined later in Lemma 5. Closely following the proof of (Zhang and Sugiyama, 2024, Proposition 1), we show the following proposition.

Proposition 2. *Let $\delta \in (0,1)$. With probability $1-\delta$, for all $t \geq 1$ and $x \in \mathcal{X}$, we have*

$$\widetilde{r}_t(x) \geq \rho^\top \mu(\theta_* x) \quad \textit{and} \quad |\rho^\top \mu(\theta_* x) - \rho^\top \mu(\theta_t x)| \leq \varepsilon_{1,t}(x) + \varepsilon_{2,t}(x)\,.$$

The key advantage of this definition of $\widetilde{r}_t(x)$ compared to the one in (5) is that it can be computed efficiently for any x and does not require solving any optimization problem.

We summarize our complete procedure in Algorithm 2 below.

Algorithm 2: *REAL: Recommendation with Exploration And Learning*

Input: Exploration length τ, regularisation parameters λ_0 and λ, step size η
Init: Run $\Theta \leftarrow$ EXPLORATION_ROUTINE(τ, λ_0)
Set $W_{\tau+1} = \lambda I_{Kd}, \overline{W}_{\tau+1} = \lambda I_{Kd}$
for *each round t in $\tau+1 \dots T$* **do**
Choose action $x_t \in \arg\max_{x \in \mathcal{X}} \widetilde{r}_t(x)$ with $\widetilde{r}_t(x)$ defined in Eq. (7)
Observe $y_t \sim \mu(\theta_* x_t)$ with $y_t \in [\![K]\!]$
Get reward ρ_{y_t}
Compute $W_t = W_t + \eta \nabla\mu(\theta_t x_t) \otimes x_t x_t^\top$
Compute $\theta_{t+1} = \arg\min_{\theta \in \Theta} \langle \nabla \ell_{t+1}(\theta_t), \theta \rangle + \frac{1}{2\eta} \|\theta - \theta_t\|^2_{\widetilde{W}_t}$
Update $W_{t+1} = W_t + \nabla\mu(\theta_{t+1} x_t) \otimes x_t x_t^\top$
Update $\overline{W}_{t+1} = \overline{W}_t + \nabla\mu(\theta_{t+1} x_t) \otimes x_t x_t^\top + 1_K 1_K^\top \otimes x_t x_t^\top$
end

3.3 Regret analysis

We now introduce our regret bound for Algorithm 2. The complete proof is deferred to Appendix E.3.

Theorem 3. *Let $\delta \in (0,1]$. Set τ, λ_0 as in Lemma 1, $\eta = 1$ and $\lambda = 144Kd$. Then, the regret of Algorithm 2 satisfies, with probability at least $1-2\delta$,*

$$\mathrm{Reg}_T \leq \mathfrak{C} Rd\sqrt{KT/\kappa_*}\log(T/\delta) + \mathfrak{C}\kappa K^2 d^2 \log^2(T/\delta)$$

where $\mathfrak{C} > 0$ is a universal constant.

A consequence for the long-term regret is that, since the dominating term scales as $Rd\sqrt{KT/\kappa_*}$, the non-linearity inherent to the problem positively influences the regret bound. This contrasts with previous results from the MNL bandit literature Amani and Thrampoulidis (2021), Lee et al. (2024), Zhang and Sugiyama (2024), where the best known rate was $O(RdK\sqrt{T})$. Our approach represents a significant improvement, as in some cases κ_* can be exponentially large in S (similarly to κ), as illustrated in the example in Appendix B.2. It is worth point out that under uniform rewards, i.e., $\rho \in \mathbb{R}\mathbf{1}_K$, any algorithm incurs zero regret. In this case, the first-order term in our regret bound vanishes, since by our definition we have $\kappa_* = +\infty$. Our result is the only one in the literature that exhibits this behavior.

The following lower-bound shows that for any number of decisions K and any dimension d, there exists a problem instance where the learner incurs a regret penalty proportional to $1/\sqrt{\kappa_*}$.

Theorem 4. *For all $K \geq 2$, $d \geq 2$ and any algorithm, there exist $\theta_* \in \Pi\mathbb{R}^{K\times d}$ and $\rho \in \mathbb{R}_+^K$ with $\rho \notin \mathbb{R}1_K$ such that for $\mathcal{X} = \mathcal{S}_1(\mathbb{R}^d)$ and for any $T \geq d^2\kappa_*$, the cumulative regret satisfies $\mathrm{Reg}_T \geq \Omega\big(Rd\sqrt{KT/\kappa_*}\big)$.*

Note that our probabilistic model with $K \geq 3$ differs from the binary one, thus our lower-bound is not a direct consequence of the binary case and requires a specific analysis, which is deferred to Appendix F. We specifically consider a non-uniform reward, i.e. $\rho \notin \mathbb{R}1_K$. For a uniform reward the regret of any algorithm is $\mathrm{Reg}_T = 0$ and $1/\kappa_* = 0$, which would render our lower-bound trivial. Our result demonstrates that the proposed algorithm is minimax-optimal and that our choice of the non-linearity constant κ_* is itself optimal.

3.3.1 Confidence Set

Before presenting the key ideas of the analysis of Theorem 3, we first establish that the confidence levels $\sigma_t(\delta)$, which appear in the definitions of the bonuses added to the reward (see Equation (7)), are sufficiently small. These levels are intrinsically linked to the size of the confidence set constructed around θ_* at each round. For each time step $t \geq \tau+1$, the pair $(\theta_{t+1}, \overline{W}_{t+1})$ is associated with the confidence set

$$\mathcal{C}_t(\delta) := \Big\{\theta : \|\theta - \theta_{t+1}\|_{\overline{W}_{t+1}} \leq \sigma_t(\delta)\Big\}$$

where $\overline{W}_{t+1} = W_{t+1} + \sum_{s=1}^t 1_K 1_K^\top \otimes x_s x_s^\top$. Leveraging the fixed diameter set we build in exploration phase and using (Lee and Oh, 2025, Theorem 4.2), we provide a $\mathrm{poly}(S)$-free confidence set. In the following lemma, we show that $\theta_* \in \mathcal{C}_t(\delta)$ with high probability. The proof is deferred to Appendix E.2.

Lemma 5. *Let $\delta \in (0,1]$. Set $\eta = 1$ and $\lambda = 144Kd$. Let us assume Lemma 1 holds. Let us define $\sigma_t(\delta) = \frac{2}{\sqrt{6}}\sqrt{Kd\log(t/\delta)} + 2S\sqrt{\lambda}$. Then we have with probability $1-\delta$, for all $t \geq 1$,*

$$\|\theta_* - \theta_{t+1}\|_{\overline{W}_{t+1}} \leq \sigma_t(\delta)\,.$$

3.3.2 Proof Sketch of Theorem 3

We start by using a classical OFU argument. Using Proposition 2 together with the definition of $x_t \in \arg\max_x \widetilde{r}_t(x)$, we bound the regret as

$$\mathrm{Reg}_T \leq \tau + \sum_{t=\tau+1}^T \rho^\top(\mu(\theta_* x_*) - \mu(\theta_* x_t)) \leq \tau + 2\sum_{t=1}^T \varepsilon_{1,t}(x_t) + 2\sum_{t=1}^T \varepsilon_{2,t}(x_t) \tag{8}$$

where $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are the bonuses defined below Equation (7). The first term τ corresponds to the exploration cost and yields the logarithmic term in T in the regret upper-bound. The sum $\sum_t \varepsilon_{2,t}$ is bounded with standard linear algebra. Defining $U_t := \frac{1}{\kappa}\sum_{s=1}^t I_K \otimes x_s x_s^\top + \frac{\lambda}{2} I_{Kd}$, we have $U_t \preccurlyeq \overline{W}_t$ (which justifies the choice of $\overline{W}_t$ instead of

W_t in the analysis), which entails

$$\begin{aligned}\sum_{t=1}^{T}\varepsilon_{2,t}(x_t) &= 3R\sum_{t=1}^{T}\sigma_t(\delta)\left\|(I_K\otimes x_t^\top)\overline{W}_t^{-1/2}\right\|_2^2 \lesssim R\kappa\sigma_T(\delta)\sum_{t=1}^{T}\mathrm{Tr}\left(\left(\tfrac{1}{\kappa}I_K\otimes x_t x_t^\top\right)\overline{W}_t^{-1}\right)\\ &\le R\kappa\sigma_T(\delta)\sum_{t=1}^{T}\mathrm{Tr}((U_t-U_{t-1})U_t^{-1}) \le R\kappa\sigma_T(\delta)\sum_{t=1}^{T}\log\frac{|U_t|}{|U_{t-1}|}\\ &\lesssim R\kappa K^2d^2\log^2(T/\delta)\,. \end{aligned} \tag{9}$$

Controlling the other sum $\sum_t \varepsilon_{1,t}$ is more challenging. Careful derivations followed by Cauchy-Schwarz inequality lead to

$$\sum_{t=1}^{T}\varepsilon_{1,t}(x_t) \lesssim \sqrt{\sigma_T(\delta)}\sqrt{\sum_{t=1}^{T}\|\overline{W}_t^{-1/2}(I_K\otimes x_t)\nabla\mu(\theta_t' x_t)^{1/2}\|_2^2}\sqrt{\sum_{t=1}^{T}\rho^\top\nabla\mu(\theta_* x_t)\rho}\,. \tag{10}$$

The first sum in the square root may again be controlled in $O(d\log T)$, i.e. K-free, through a careful linear algebra analysis of the eigenvalues and a Trace-Determinant argument. The second sum is a standard term that appears in earlier work. Indeed, a key step in achieving minimax optimal rates in the binary setting (Abeille et al., 2021, Faury et al., 2022) involves proving that

$$\sum_{t=1}^{T}\mu'(\theta_*^\top x_t) \le T/\kappa_* + \mathrm{Reg}_T\,.$$

In the MNL setting, Zhang and Sugiyama (2024, Appendix C.5) also showed that

$$\sum_{t=1}^{T}\rho^\top\nabla\mu(\theta_* x_t)\rho \le R^2T/\kappa_* + 2\mathrm{Reg}_T\,, \tag{11}$$

was sufficient to obtain a regret with a $1/\kappa_*$ dependence. However, as they admit, such a relationship is unclear in general and challenging to establish. Indeed, in the binary setting, the analysis by Abeille et al. (2021) heavily relies on specific properties of the one-dimensional sigmoid function μ, which satisfies $|\mu''| \le \mu'$. These properties do not carry over to the multi-dimensional setting when μ is the softmax function. Moreover, in the binary setting, since the sigmoid function is increasing, the optimal decision $x_* \in \arg\max_{x\in\mathcal{X}}\{\mu(\theta_*^\top x)\}$ can be easily expressed as the solution to the linear optimization problem $\arg\max_{x\in\mathcal{X}}\{\theta_*^\top x\}$. This no longer holds because μ is multi-dimensional and because x_* also depends on the reward vector ρ. Due to this difficulty, instead of (11), Zhang and Sugiyama (2024) show that

$$\sum_{t=1}^{T}\rho^\top\nabla\mu(\theta_* x_t)\rho \le R^2T/\kappa_* + 2R\mathrm{Reg}_T + \sum_{t=1}^{T}\sum_{k=1}^{K}\rho_k^2(\mu(\theta_* x_t)_k - \mu(\theta_* x_*)_k)\,.$$

The difficulty, as pointed out in Zhang and Sugiyama (2024), is that the last term may be non-negative and significantly higher than the regret. To circumvent this problem, we derive a slightly different upper-bound that replaces Reg_T in Equation (11) with an upper-bound obtained from the reward bonuses $\varepsilon_{1,t}(x_t)$ and $\varepsilon_{2,t}(x_t)$. We add and subtract $\rho^\top\nabla\mu(\theta_* x_*)\rho$. Carefully controlling the difference term we establish:

$$\begin{aligned}\sum_{t=1}^{T}\rho^\top\nabla\mu(\theta_* x_t)\rho &= \sum_{t=1}^{T}\langle\rho,\nabla\mu(\theta_* x_*)\rho\rangle + \langle\rho,(\nabla\mu(\theta_* x_t)-\nabla\mu(\theta_* x_*))\rho\rangle\\ &\le R^2T/\kappa_* + (2\sqrt{K}+4)\sum_{t=1}^{T}\left(\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t)\right)\,.\end{aligned}$$

The proof concludes by combining this with equations (9) and (10), solving a second-order equation of the form

$$\sum_{t=1}^{T}(\varepsilon_{1,t}+\varepsilon_{2,t}) \le C_1 + C_2\sqrt{R^2T/\kappa_* + \sqrt{K}\sum_{t=1}^{T}(\varepsilon_{1,t}+\varepsilon_{2,t})},$$

and substituting the solution into the initial regret bound (8).

3.4 Adaptive exploration and changing action sets

The initial exploration phase of our algorithm might be concerning from a practical viewpoint. It enforces κ rounds of exploration which given the nature of κ might be costly. In Appendix G.1, we present a variant of our algorithm (see Algorithm 3) that employs adaptive rather than hardcoded exploration based on Lee and Oh (2025) work. This adaptive approach enables the extension of our framework to non-stationary action sets $\mathcal{X}_t \subseteq \mathcal{X}$. We adapt our definition of the non-linearity constant κ_* to match the actions sets $\mathcal{X}_t$:

$$\kappa_{*,t} = \frac{\|\rho\|_2^2}{\rho^\top \nabla\mu(\theta_* x_{*,t})\rho} \text{ when } \rho \notin \mathbb{R}1_K \text{ and } \kappa_{*,t} = +\infty \text{ when } \rho \in \mathbb{R}1_K$$

where $x_{*,t} := \arg\max_{x\in\mathcal{X}_t} \rho^\top \mu(\theta_* x)$. We also modify the regret definition to take $\mathcal{X}_t$ into account:

$$\mathrm{Reg}_T := \sum_{t=1}^{T} \rho^\top \mu(\theta_* x_{*,t}) - \rho^\top \mu(\theta_* x_t)\,.$$

The algorithm is based on a trigger condition. Let $T^w \subseteq [T]$ denote the set of exploration steps of the algorithm. At any time step t, the algorithm performs an exploration step if the following condition is satisfied:

$$\max_{x\in\mathcal{X}_t} \|I_K \otimes x\|^2_{(H^w_{t-1})^{-1}} \geq \frac{1}{\tau_t^2} \qquad \text{where} \quad H^w_{t-1} = \sum_{s=1}^{t-1} \tfrac{1}{\kappa} I_K \otimes x_s x_s^\top 1\{s \in T^w\}\,.$$

Each time the algorithm explores, it refines its estimate of θ_* and updates the corresponding confidence set. Otherwise, it follows the learning procedure described in Algorithm 2.

We now introduce our regret bound for Algorithm 3. The proof is deferred to Appendix G.1.

Theorem 6. *Let $\delta \in (0,1]$. Set $\lambda^w = 72(1+\sqrt{6}S)Kd, \eta^w = (1+\sqrt{6}S)/2$ and $\lambda = 144Kd$. Then, the regret of Algorithm 3 satisfies with probability at least $1-2\delta$,*

$$\mathrm{Reg}_T \leq \widetilde{O}\left(Rd\sqrt{K\sum_{t\notin T^w}\frac{1}{\kappa_{*,t}}}\right)$$

where T^w is the set of time steps when the algorithm explores.

In the case of constant arm-sets $\mathcal{X}_t = \mathcal{X}$, we recover the regret guarantee of Theorem 3, obtaining a regret upper-bound of $\widetilde{O}(Rd\sqrt{KT/\kappa_*})$. In the non-stationary case, we obtain $\sqrt{T}\sqrt{\frac{1}{T}\sum_{t\notin T^w}\frac{1}{\kappa_{*,t}}}$, replacing the non-linearity constant in the optimum by its on-trajectory average version.

4 Conclusion

This work establishes that non-linearity in multinomial logistic bandits can be leveraged to improve asymptotic regret guarantees, extending results previously known only for the binary setting. We introduce a new problem-dependent constant κ_* and design an algorithm that achieves minimax-optimal regret bounds of order $\widetilde{O}(Rd\sqrt{KT/\kappa_*})$, while preserving computational efficiency. Crucially, we also prove a matching lower-bound of $\Omega(Rd\sqrt{KT/\kappa_*})$, thereby demonstrating that both our algorithm and our definition of κ_* are optimal up to logarithmic factors. Our analysis relies on a tailored exploration strategy and exploits the self-concordance property of the softmax function, enabling tighter control of curvature effects at the optimum. These findings demonstrate that non-linearity, rather than being a limitation, can serve as a structural advantage in sequential decision-making.

Acknowledgements. We thank Francis Bach for his precious knowledge. A.R. acknowledges the support of the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and the support of the European Research Council (grant REAL 947908).

References

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

M. Abeille, L. Faury, and C. Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

P. Agrawal, T. Tulabandhula, and V. Avadhanula. A tractable online learning algorithm for the multinomial logit contextual bandit. *European Journal of Operational Research*, 310(2):737–750, 2023.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Thompson sampling for the mnl-bandit. In *Conference on learning theory*, pages 76–78. PMLR, 2017.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

S. Amani and C. Thrampoulidis. Ucb-based algorithms for multinomial logistic regression bandits. *Advances in Neural Information Processing Systems*, 34:2913–2924, 2021.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

W. C. Cheung and D. Simchi-Levi. Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Available at SSRN 3075658*, 2017.

K. Dong, Y. Li, Q. Zhang, and Y. Zhou. Multinomial logit bandit with low switching cost. In *International Conference on Machine Learning*, pages 2607–2615. PMLR, 2020.

L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

L. Faury, M. Abeille, K.-S. Jun, and C. Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 546–580. PMLR, 2022.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

E. Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

K.-S. Jun, L. Jain, B. Mason, and H. Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pages 5148–5157. PMLR, 2021.

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

J. Lee and M.-h. Oh. Improved online confidence bounds for multinomial logistic bandits. *arXiv preprint arXiv:2502.10020*, 2025.

J. Lee, S.-Y. Yun, and K.-S. Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4482. PMLR, 2024.

N. Perivier and V. Goyal. Dynamic pricing and assortment under a contextual mnl demand. *Advances in Neural Information Processing Systems*, 35:3461–3474, 2022.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 178(1):145–213, 2019.

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

H. Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Y. Xia. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484): 1631–1640, 2008.

Y.-J. Zhang and M. Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36, 2024.

APPENDIX

This appendix is organised as follows:

- Appendix A: Notations
- Appendix B: Bounds on the Constants κ and κ_*
- Appendix C: Comparison with the Framework of Amani and Thrampoulidis (2021)
- Appendix D: Discussion of the Multinomial *Logit* Bandits
- Appendix E: Analysis of Algorithm 2
- Appendix F: Proof of Theorem 4 - Lower bound
- Appendix G: Removing the Exploration
- Appendix H: Auxiliary Results

A Notations

We detail below useful notations and basic properties used throughout the appendix.

- $[\![T]\!] := \{1, 2, \dots, T\} \quad , \forall T \in \mathbb{N}^*$
- $\mathfrak{C}$: Universal constant, i.e. independent of $S, d, K, T, \kappa, \kappa_*$
- $\kappa_*^{-1} = \frac{\rho^\top \nabla\mu(\theta_* x_*)\rho}{\|\rho\|_2^2}$
- $\kappa := \max_{\|\theta\|\leq S} \max_{x\in\mathcal{X}} \frac{1}{\lambda_{K-1}(\nabla\mu(\theta x))}$
- $\ell_{t+1}(\theta) := \sum_{k=1}^K -1[k = y_t] \log(\mu(\theta x_t)_k)$
- $\text{diam}_{\mathcal{X}}(\Theta) = \max_{x\in\mathcal{X}} \max_{\theta_1,\theta_2\in\Theta} \|(\theta_1 - \theta_2)x\|_2$
- $H_t(\theta) := \sum_{s=1}^t \nabla\mu(\theta x_s) \otimes x_s x_s^\top + \lambda_0 I_{Kd}$
- $\overline{H}_t(\theta) := \sum_{s=1}^t \nabla\mu(\theta x_s) \otimes x_s x_s^\top + \sum_{s=\tau+1}^{t-1} 1_K 1_K^\top \otimes x_s x_s^\top + \lambda_0 I_{Kd}$
- $g_t(\theta) := \sum_{s=1}^t \mu(\theta x_s) \otimes x_s + \lambda_0 \theta$
- $G_t(\theta_1, \theta_2) := \sum_{s=1}^t \int_0^1 \nabla\mu((v\theta_1 + (1-v)\theta_2)x_s)dv \otimes x_s x_s^\top + \lambda_0 I_{Kd}$
- $g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2)$ (Mean-value Theorem)
- $W_t := \sum_{s=\tau+1}^{t-1} \nabla\mu(\theta_{s+1} x_s) \otimes x_s x_s^\top + \lambda I_{Kd}$
- $\overline{W}_t := \sum_{s=\tau+1}^{t-1} \nabla\mu(\theta_{s+1} x_s) \otimes x_s x_s^\top + \sum_{s=\tau+1}^{t-1} 1_K 1_K^\top \otimes x_s x_s^\top + \lambda I_{Kd}$
- $\alpha_s(\theta_1, \theta_2) := \int_0^1 \nabla\mu(((1-v)\theta_1 + v\theta_2)x_s)dv \otimes x_s x_s^\top$
- $\alpha_s(\theta_1, \theta_2) = \alpha_s(\theta_2, \theta_1)$ (change of variable)
- $\widetilde{\alpha}_s(\theta_1, \theta_2) := \int_0^1 (1-v)\nabla\mu(((1-v)\theta_1 + v\theta_2)x_s)dv \otimes x_s x_s^\top$
- $\alpha(\theta_1 x_1, \theta_2 x_2) := \int_0^1 \nabla\mu((1-v)\theta_1 x_1 + v\theta_2 x_2)dv$

B Bounds on the Constants κ and κ_*

B.1 Upper and lower bounds on the constant κ

In this appendix, we show the following lemma that bounds κ by above and by below. In particular, we recover up to constant factors the bounds proved by Amani and Thrampoulidis (2021) for earlier definitions of κ (see Appendix C thereafter).

Lemma. *For any even $K \in \mathbb{N}$ and $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq X\}$, we have*

$$\frac{K}{4} + \frac{K}{4} e^{2SX} \leq \kappa \leq K e^{2SX} \,.$$

for κ as defined in Equation (1)*.*

Proof. We first prove the upper-bound. Fix any $z \in \mathbb{R}^K$. We bound the second smallest eigenvalue λ_{K-1} of $\nabla\mu(z)$. using Weyl's inequality (Weyl, 1912) and the definition of the gradient of the softmax $\nabla\mu(z) = \mathrm{diag}(\mu(z)) - \mu(z)\mu(z)^\top$. A direct application of Weyl's inequality gives

$$\lambda_{K-1}(\mathrm{diag}(\mu(z)) - \mu(z)\mu(z)^\top) \geq \lambda_K(\mathrm{diag}(\mu(z))) + \lambda_{K-1}(-\mu(z)\mu(z)^\top) = \min_{i \in [\![K]\!]} \mu(z)_i \,.$$

Thus we have

$$\lambda_{K-1}(\mathrm{diag}(\mu(z)) - \mu(z)\mu(z)^\top) \geq \frac{\exp(-SX)}{K\exp(SX)} = \frac{1}{K}\exp(-2SX)$$

where $X := \max_{x\in\mathcal{X}} \|x\|_2$ and S is assumed such that $\|\theta\|_2 \leq S$. Hence,

$$\kappa := \frac{1}{\min_{\|\theta\|_2 \leq S} \min_{x\in\mathcal{X}} \lambda_{K-1}(\nabla\mu(\theta x))} \leq K e^{2SX} \,.$$

We now prove the lower-bound. For simplicity, we assumed that $\mathcal{X}$ is a ball of radius X and that K is even. A direct application of the Schur-Horn Theorem gives for all $z \in \mathbb{R}^K$

$$\min_{i,j,i\neq j} \nabla\mu(z)_{ii} + \nabla\mu(z)_{jj} \geq \lambda_K(\nabla\mu(z)) + \lambda_{K-1}(\nabla\mu(z)) = \lambda_{K-1}(\nabla\mu(z)) \,.$$

We choose θ such that $\|\theta\|_2 = S$ and with the first $K/2$ rows equal to each other, i.e. $[\theta]_1 = [\theta]_i$ for $i \in [\![K/2]\!]$ and with the others rows collinear in the opposite direction, i.e. $[\theta]_i = -[\theta]_1$ for all $i \geq 2$. We choose x such that $x = -\frac{[\theta]_1}{S} X$. Thus we obtain

$$\begin{aligned}\frac{4}{K\big(1+\exp(2SX)\big)} \geq \frac{2\exp(-SX)}{\frac{K}{2}\exp(-SX) + \frac{K}{2}\exp(SX)} &\geq \min_{x\in\mathcal{X}} \min_{\|\theta\|_2\leq S} \min_{i,j,i\neq j} \mu(\theta x)_{ii} + \mu(\theta x)_{jj} \\ &\geq \min_{x\in\mathcal{X}} \min_{\|\theta\|_2\leq S} \min_{i,j,i\neq j} \nabla\mu(\theta x)_{ii} + \nabla\mu(\theta x)_{jj} \geq \min_{x\in\mathcal{X}} \min_{\|\theta\|_2\leq S} \lambda_{K-1}(\nabla\mu(\theta x)) =: \kappa^{-1} \,,\end{aligned}$$

which concludes the proof. □

B.2 Example of large κ_*

Let $K \geq 2$ be even and $d \geq 1$. Let us consider the following problem, we define $\theta_* = \Pi M_* \in \Pi\mathbb{R}^{K\times d}$ with M_* equal to

$$M_* := \begin{bmatrix} m & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^d$$

where $m > 0$, moreover M_* is such that $\|\theta_*\|_2 = S$. We define $\rho \in \mathbb{R}^K$ such that

$$\rho := \frac{1}{\sqrt{K+3}} \begin{bmatrix} 2 \\ 1 \\ \vdots \\ 1 \end{bmatrix} .$$

Note that $\|\rho\|_2 = 1$. We choose $\mathcal{X} = \mathcal{S}_1(\mathbb{R}^d)$. We have that $x_* = [M_*]_1/\|[M_*]_1\|_2$. Note that $\|x_*\|_2 = 1 = X$. Let us compute κ_*:

$$\kappa_*^{-1} = \rho^\top \nabla\mu(\theta_* x_*)\rho = \rho^\top \left(\mathrm{diag}(\mu(\theta_* x_*)) - \mu(\theta_* x_*)\mu(\theta_* x_*)^\top\right)\rho$$

where the second equality is due to the definition of $\nabla\mu(\cdot)$. This can be developed into:

$$\begin{aligned}
\kappa_*^{-1} &= \sum_{k=1}^K \rho_k \mu(\theta_* x_*)_k \left[\sum_{i=1}^K \rho_i \left(\delta_{ik} - \mu(\theta_* x_*)_i\right)\right] \\
&= \rho_1^2 \mu(\theta_* x_*)_1 (1 - \mu(\theta_* x_*)_1) - 2\rho_1 \mu(\theta_* x_*)_1 \sum_{k=2}^K \rho_k \mu(\theta_* x_*)_k \\
&\quad + \sum_{k=2}^K \rho_k \mu(\theta_* x_*)_k \left[\sum_{i=2}^K \rho_i \left(\delta_{ik} - \mu(\theta_* x_*)_i\right)\right] . \qquad (12)
\end{aligned}$$

Let us prove that the first two terms cancel each other.

$$\begin{aligned}
\rho_1^2 \mu(\theta_* x_*)_1 (1 - \mu(\theta_* x_*)_1) &= \rho_1 2\rho_2 \mu(\theta_* x_*)_1 (1 - \mu(\theta_* x_*)_1) && (\rho_1 = 2\rho_2) \\
&= 2\rho_1\rho_2 \mu(\theta_* x_*)_1 \sum_{k=2}^K \mu(\theta_* x_*)_k && (\mu \text{ is a probability}) \\
&= 2\rho_1 \mu(\theta_* x_*)_1 \sum_{k=2}^K \rho_k \mu(\theta_* x_*)_k && (\rho_2 = \rho_k, \forall k \in [\![2, K]\!])
\end{aligned}$$

Consequently, Equation (12) becomes

$$\begin{aligned}
\kappa_*^{-1} &= \sum_{k=2}^K \rho_k \mu(\theta_* x_*)_k \left[\sum_{i=2}^K \rho_i \left(\delta_{ik} - \mu(\theta_* x_*)_i\right)\right] \\
&= \frac{2}{K+3} \sum_{k=2}^K \mu(\theta_* x_*)_k \left[\sum_{i=2}^K \left(\delta_{ik} - \mu(\theta_* x_*)_i\right)\right] && (\text{Def of } \rho) \\
&= \frac{2}{K+3} \sum_{k=2}^K \mu(\theta_* x_*)_k \left[1 - \sum_{i=2}^K \mu(\theta_* x_*)_i\right] \\
&= \frac{2}{K+3} \sum_{k=2}^K \mu(\theta_* x_*)_k \mu(\theta_* x_*)_1 \\
&\leq 2\mu(\theta_* x_*)_2 \mu(\theta_* x_*)_1 .
\end{aligned}$$

We now use the definition of the softmax to upper-bound the probabilities.

$$\begin{aligned}
\kappa_*^{-1} &\leq 2 \frac{1}{K - 1 + \exp([M_*]_1^\top x_*)} \cdot \frac{\exp([M_*]_1^\top x_*)}{K - 1 + \exp([M_*]_1^\top x_*)} \\
&= 2 \frac{\exp(-[M_*]_1^\top x_*)}{(K-1)\exp(-[M_*]_1^\top x_*) + 1} \cdot \frac{1}{(K-1)\exp(-[M_*]_1^\top x_*) + 1} \\
&= 2 \frac{\exp(-S)}{(K-1)\exp(-S) + 1} \cdot \frac{1}{(K-1)\exp(-S) + 1} \\
&\leq 2\exp(-S)
\end{aligned}$$

We exhibit a case where κ_* is exponentially small in $S = \|\theta_*\|_2$. In this case, by Theorem 3, the asymptotic regret is thus of order

$$\mathrm{Reg}_T \leq \widetilde{O}\Big(Rd\exp(-S/2)\sqrt{KT}\Big) .$$

C Comparison with the Framework of Amani and Thrampoulidis (2021)

Amani and Thrampoulidis (2021) also consider a MNL bandit framework, which is equivalent but defined slightly differently from ours. In their framework, the environment parameter $\widetilde{\theta}_* \in \mathbb{R}^{K \times d}$ is defined with its last row equal to

zero $[\widetilde{\theta}_*]_K = 0_d$. Therefore the probability of a decision $i \in [\![K]\!]$ becomes

$$\mathbb{P}[y_t = i|x_t] = \begin{cases} \dfrac{1}{1 + \sum_{k=1}^{K-1} \exp([\widetilde{\theta}_*]_k x_t)} & \text{if } i = K \\ \dfrac{\exp([\widetilde{\theta}_*]_i x_t)}{1 + \sum_{k=1}^{K-1} \exp([\widetilde{\theta}_*]_k x_t)} & \text{if } i < K \end{cases} .$$

The reward vector is also defined $\widetilde{\rho} \in \mathbb{R}_+^K$ but with its last element equal to zero $\rho_K = 0$. The regret is defined as

$$\widetilde{\mathrm{Reg}}_T = \sum_{t=1}^{T} \sum_{k=1}^{K-1} \widetilde{\rho}_k \left(\mathbb{P}[y_t = k|x_*] - \mathbb{P}[y_t = k|x_t] \right) .$$

Thus the last element of the probability vector is not needed and we define the vector $\widetilde{\mu}(\theta x) \in \mathbb{R}^{K-1}$ as the truncated probability vector $[\widetilde{\mu}(\theta x)]_k = \mathbb{P}[y_t = k|x_t]$. Contrary to our case, the fact that $\widetilde{\mu}$ is not a probability ensures that its minimum eigenvalue is well-defined (Amani and Thrampoulidis, 2021, Lemma 5). The problem-dependent constant measuring the non-linearity is defined as:

$$\widetilde{\kappa} := \frac{1}{\min_{x \in \mathcal{X}} \min_{\|\theta\|_2 \leq S} \lambda_{\min}(\mathrm{diag}(\widetilde{\mu}(\theta x)) - \widetilde{\mu}(\theta x)\widetilde{\mu}(\theta x)^\top)} .$$

As shown by (Amani and Thrampoulidis, 2021, Eq. (20)) the constant $\widetilde{\kappa}$ is exponentially large with respect to S and X. These lower and upper bounds on $\widetilde{\kappa}$ show that our constant κ is comparable, see Appendix B.1.

Now note that in our framework, by choosing without loss of generality $\min_k \rho_k = \rho_K$ we have

$$\begin{aligned} \mathrm{Reg}_T &:= \sum_{t=1}^{T} \rho^\top (\mu(\theta_* x_*) - \mu(\theta_* x_t)) \\ &= \sum_{t=1}^{T} (\rho - \rho_K 1_K)^\top (\mu(\theta_* x_*) - \mu(\theta_* x_t)) \\ &+ \sum_{t=1}^{T} \rho_K 1_K^\top (\mu(\theta_* x_*) - \mu(\theta_* x_t)) \\ &= \sum_{t=1}^{T} (\rho - \rho_K 1_K)^\top (\mu(\theta_* x_*) - \mu(\theta_* x_t)) \\ &= \sum_{t=1}^{T} \sum_{k=1}^{K-1} (\rho_k - \rho_K)(\mu(\theta_* x_*)_k - \mu(\theta_* x_t)_k) \end{aligned}$$

We could then choose an arbitrary value for $[\theta_*]_K$. For $[\theta_*]_K = 0_d$ we recover the framework of Amani and Thrampoulidis (2021). Thus their framework is included in ours.

D Discussion of the Multinomial *Logit* Bandits

In this section we discuss the differences between Multinomial Logistic Bandits, our framework, and the Multinomial *Logit* Bandit framework. In our setting, the environment may have multiple reactions to a single action. On the other hand, in the *Logit* setting the agent selects a set of items to which to environment responds with either a click or no click. In our problem setting, the probability of observing decision $y_t = k$ given context x_t is:

$$\mathbb{P}[y_t = k|x_t] = \frac{\exp((\theta_*)_k^\top x_t)}{\sum_{i=1}^{K} \exp((\theta_*)_i^\top x_t)}$$

where each possible decision $i \in [\![K]\!]$ has its own parameter $(\theta_*)_i \in \mathbb{R}^d$. Thus, we are estimating a different parameter vector for each decision, and the variation in decision probabilities comes from these parameter differences.

In contrast, in the *Logit* setting, the agent chooses a subset $S_t \subseteq [\![K]\!]$, and the environment responds with a choice over that subset. The probability of observing decision $k \in S_t$ is:

$$\mathbb{P}[y_t = k|S_t] = \frac{\exp(\theta_*^\top x_{t,k})}{\sum_{i \in S_t} \exp(\theta_*^\top x_{t,i})}$$

where the agent observes the context vectors $x_{t,i}$ for all $i \in [\![K]\!]$, and there is a single share parameter $\theta_* \in \mathbb{R}^d$. In this setting, all variations in decision probabilities arise from the context vectors, not from the parameter θ_*.

In summary, the settings differ fundamentally in their parameterisation, feedback structure, and modeling assumptions. Ours involves learning distinct models per action; the *Logit* setting uses a shared parameter across all items and focuses on contextual differences.

E Analysis of Algorithm 2

E.1 Exploration Routine

E.1.1 Confidence Set

We build our confidence set over this proposition from Zhang and Sugiyama (2024, Theorem 1), which is itself an improvement of Amani and Thrampoulidis (2021, Theorem 1). As demonstrated by Abeille et al. (2021, Section 6), the confidence set presented in the following proposition is not convex. To address this, we construct a convex relaxation, see Proposition 8.

Proposition 7. *Set the parameter $\lambda_0 = (S+1)Kd\log(T/\delta)$ with a certain $\delta \in (0,1]$. Let the event E_δ be defined by*

$$E_\delta : \{\forall t \geq 1, \|g_t(\theta_*) - g_t(\widehat{\theta}_{t+1})\|^2_{H_t^{-1}(\theta_*)} \leq \gamma_t(\delta)\}$$

where $\gamma_t(\delta) := 16\lambda_0$. We have that

$$\mathbb{P}(E_\delta) \geq 1 - \delta\,.$$

Note that $V_t \preccurlyeq \overline{H}_t(\theta_*)$ for $\overline{H}_t(\theta) := H_t(\theta) + \sum_{s=1}^t 1_K 1_K^\top \otimes x_s x_s^\top$, therefore proving the following lemma is sufficient to prove that

$$\mathbb{P}(\theta_* \in \Theta) \geq 1 - \delta\,.$$

Proposition 8. *Let $\delta \in (0,1]$ and $\widehat{\theta}_{t+1}$ be defined as in Algorithm 1. We have that*

$$\mathbb{P}\left(\forall t \geq 1, \|\widehat{\theta}_{t+1} - \theta_*\|^2_{\overline{H}_t(\theta_*)} \leq \beta_t(\delta)\right) \geq 1 - \delta$$

where $\overline{H}_t(\theta) := H_t(\theta) + \sum_{s=1}^t 1_K 1_K^\top \otimes x_s x_s^\top$ and $\beta_t(\delta) := \left(1 + \frac{\gamma_t(\delta)}{\lambda_0} + \sqrt{\frac{\gamma_t(\delta)}{\lambda_0}}\right)^2 \gamma_t(\delta)$ with $\gamma_t(\delta)$ and λ_0 defined in Proposition 7.

Proof. We follow the proof of Lemma 1 in Faury et al. (2022).

Step 1: Sub-Exponential Self-Concordance.

We first show that for all time step $t \geq 1$, if the event E_δ holds, we have that

$$H_t(\theta_*) \preccurlyeq \left(1 + \frac{\gamma_t(\delta)}{\lambda_0} + \sqrt{\frac{\gamma_t(\delta)}{\lambda_0}}\right) G_t(\theta_*, \widehat{\theta}_{t+1})$$

where λ_0 and $\gamma_t(\delta)$ are defined in Proposition 7. From the proof of Lemma 13 in Amani and Thrampoulidis (2021) we have that

$$H_t(\theta_*) \preccurlyeq \sum_{s=1}^t (1 + d(x_s, \widehat{\theta}_{t+1}, \theta_*))\alpha_s(\widehat{\theta}_{t+1}, \theta_*) + \lambda_0 I_{Kd}$$

where $d(x_s, \widehat{\theta}_{t+1}, \theta_*) := \|(\widehat{\theta}_{t+1} - \theta_*)x_s\|_2$. From now on the proof of Lemma 2 of Abeille et al. (2021) also holds in the multiclass setting to conclude this proof step. We provide it for the sake of completeness. We apply Cauchy-Schwarz inequality and obtain

$$\begin{aligned} d(x_s, \widehat{\theta}_{t+1}, \theta_*) &\leq \|x_s\|_{G_t^{-1}(\widehat{\theta}_{t+1}, \theta_*)} \|\widehat{\theta}_{t+1} - \theta_*\|_{G_t(\widehat{\theta}_{t+1}, \theta_*)} \\ &\leq \|x_s\|_{G_t^{-1}(\widehat{\theta}_{t+1}, \theta_*)} \|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|_{G_t^{-1}(\widehat{\theta}_{t+1}, \theta_*)} \leq \lambda_0^{-1/2} \|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|_{G_t^{-1}(\widehat{\theta}_{t+1}, \theta_*)}\,. \end{aligned}$$

Putting it back we get

$$H_t(\theta_*) \preccurlyeq \left(1 + \lambda_0^{-1/2} \|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|_{G_t^{-1}(\widehat{\theta}_{t+1},\theta_*)}\right) G_t(\theta_*, \widehat{\theta}_{t+1}) \,. \tag{13}$$

Therefore using this matrix inequality and event E_δ we get

$$\begin{aligned}
&\|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|^2_{G_t^{-1}(\theta_*,\widehat{\theta}_{t+1})} \\
&\quad\leq \left(1 + \lambda_0^{-1/2} \|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|_{G_t^{-1}(\widehat{\theta}_{t+1},\theta_*)}\right) \|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|^2_{H_t^{-1}(\theta_*)} \\
&\quad\leq \gamma_t(\delta)\lambda_0^{-1/2} \|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|_{G_t^{-1}(\widehat{\theta}_{t+1},\theta_*)} + \gamma_t(\delta) \,.
\end{aligned}$$

Solving for $\|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|_{G_t^{-1}(\widehat{\theta}_{t+1},\theta_*)}$ we get

$$\|g_t(\widehat{\theta}_{t+1}) - g_t(\theta_*)\|_{G_t^{-1}(\widehat{\theta}_{t+1},\theta_*)} \leq \gamma_t(\delta)\lambda_0^{-1/2} + \sqrt{\gamma_t(\delta)} \,.$$

We now put this back in Equation(13) to conclude and obtain:

$$H_t(\theta_*) \preccurlyeq \left(1 + \frac{\gamma_t(\delta)}{\lambda_0} + \sqrt{\frac{\gamma_t(\delta)}{\lambda_0}}\right) G_t(\theta_*, \widehat{\theta}_{t+1}) \,.$$

Step 2: Applying Self-concordance.

We apply twice the self-concordance property to get

$$\begin{aligned}
\|\theta_* - \widehat{\theta}_{t+1}\|^2_{H_t(\theta_*)} &\leq \left(1 + \frac{\gamma_t(\delta)}{\lambda_0} + \sqrt{\frac{\gamma_t(\delta)}{\lambda_0}}\right) \|\theta_* - \widehat{\theta}_{t+1}\|^2_{G_t(\theta_*,\widehat{\theta}_{t+1})} \\
&\leq \left(1 + \frac{\gamma_t(\delta)}{\lambda_0} + \sqrt{\frac{\gamma_t(\delta)}{\lambda_0}}\right) \|g_t(\theta_*) - g_t(\widehat{\theta}_{t+1})\|^2_{G_t^{-1}(\theta_*,\widehat{\theta}_{t+1})} \\
&\leq \left(1 + \frac{\gamma_t(\delta)}{\lambda_0} + \sqrt{\frac{\gamma_t(\delta)}{\lambda_0}}\right)^2 \|g_t(\theta_*) - g_t(\widehat{\theta}_{t+1})\|^2_{H_t^{-1}(\theta_*)} \\
&\leq \left(1 + \frac{\gamma_t(\delta)}{\lambda_0} + \sqrt{\frac{\gamma_t(\delta)}{\lambda_0}}\right)^2 \gamma_t(\delta) \\
&=: \beta_t(\delta) \,.
\end{aligned}$$

Step 3: From $H_t(\theta_*)$ to $\overline{H}_t(\theta_*)$.

We decompose $\mathbb{R}^K$ as $\mathbb{R}^K = 1_K \oplus \mathcal{H}$ where $\mathcal{H}$ is the hyperplane supported by 1_K. Recall that $\theta_* \in \Pi R^{K\times d}$ and $\widehat{\theta}_{t+1} \in \Pi\mathbb{R}^{K\times d}$, for all $x \in \mathcal{X}$, by definition of Π, $\theta_* x$ and $\widehat{\theta}_{t+1} x$ are in $\mathcal{H}$. Therefore $\sum_{s=1}^t \|(\theta_* - \widehat{\theta}_{t+1})x_s\|^2_{1_K 1_K^\top} = 0$. And we can conclude by

$$\|\theta_* - \widehat{\theta}_{t+1}\|^2_{\overline{H}_t(\theta_*)} = \|\theta_* - \widehat{\theta}_{t+1}\|^2_{H_t(\theta_*)} \leq \beta_t(\delta) \,.$$

□

E.1.2 Proof of Lemma 1

Lemma 1. *Let $\delta \in (0,1]$, $\lambda_0 = (S+1)Kd\log(T/\delta)$ and $\tau = 336^2\lambda_0\kappa Kd\log(T)$. Then, the set Θ returned by Algorithm 1 satisfies with probability $1-\delta$*

$$\theta_* \in \Theta \qquad \textit{and} \qquad \mathrm{diam}_{\mathcal{X}}(\Theta) \leq 1/\sqrt{6} \,.$$

We adapt the proof of Lemma 2 of Faury et al. (2022) to the multiclass setting.

Proof. We start by making the term I_K appear in order to match the dimension of $\overline{H}_t(\theta_*)$.

$$\begin{aligned}
\mathrm{diam}_{\mathcal{X}}(\Theta) &= \max_{x\in\mathcal{X}} \max_{\theta_1,\theta_2\in\Theta} \|(\theta_1-\theta_2)x\|_2 \\
&= \max_{x\in\mathcal{X}} \max_{\theta_1,\theta_2\in\Theta} \|(I_K\otimes x^\top)(\theta_1-\theta_2)\|_2 \\
&\le \max_{x\in\mathcal{X}} \max_{\theta_1,\theta_2\in\Theta} \|I_K\otimes x^\top\|_{V_\tau^{-1}} \|\theta_1-\theta_2\|_{V_\tau} && \text{Cauchy-Schwarz} \\
&\le 2\sqrt{\beta_\tau(\delta)} \max_{x\in\mathcal{X}} \|I_K\otimes x\|_{V_\tau^{-1}} && \text{Proposition 8 and symmetry} \\
&= 2\sqrt{\beta_\tau(\delta)} \sqrt{\max_{x\in\mathcal{X}} \|I_K\otimes x\|^2_{V_\tau^{-1}}} \\
&= 2\sqrt{\beta_\tau(\delta)}\tau^{-1/2} \sqrt{\sum_{s=1}^{\tau} \max_{x\in\mathcal{X}} \|I_K\otimes x\|^2_{V_\tau^{-1}}} \\
&\le 2\sqrt{\beta_\tau(\delta)}\tau^{-1/2} \sqrt{\sum_{s=1}^{\tau} \max_{x\in\mathcal{X}} \|I_K\otimes x\|^2_{V_{s-1}^{-1}}} && (V_\tau \succcurlyeq V_{s-1}) \\
&\le 2\sqrt{\beta_\tau(\delta)}\tau^{-1/2} \sqrt{\sum_{s=1}^{\tau} \|I_K\otimes x_s\|^2_{V_{s-1}^{-1}}} && \text{definition of } x_s \\
&= 2\sqrt{\beta_\tau(\delta)}\tau^{-1/2}\kappa^{1/2} \sqrt{\sum_{s=1}^{\tau} \|\kappa^{-1/2} I_K\otimes x_s\|^2_{V_{s-1}^{-1}}} \\
&\le 4\sqrt{\beta_\tau(\delta)}\tau^{-1/2}\kappa^{1/2} \sqrt{Kd\log\left(1+\tfrac{\tau}{Kd}\right)} && \text{Abbasi-Yadkori et al. (2011, lemma 10)}
\end{aligned}$$

Thus if we choose $\tau = 96\beta_\tau(\delta)\kappa Kd\log\left(1+\frac{T}{Kd}\right)$ we have that $\mathrm{diam}_{\mathcal{X}}(\Theta) \le 1/\sqrt{6}$. □

E.2 Proof of Lemma 5

Lemma 5. *Let $\delta \in (0,1]$. Set $\eta = 1$ and $\lambda = 144Kd$. Let us assume Lemma 1 holds. Let us define $\sigma_t(\delta) = \frac{2}{\sqrt{6}}\sqrt{Kd\log(t/\delta)} + 2S\sqrt{\lambda}$. Then we have with probability $1-\delta$, for all $t \ge 1$,*

$$\|\theta_* - \theta_{t+1}\|_{\overline{W}_{t+1}} \le \sigma_t(\delta)\,.$$

Proof. First, by Lemma 1, we can apply (Lee and Oh, 2025, Theorem 4.2) with $\alpha = 1/\sqrt{6}$ to obtain

$$\|\theta_* - \theta_{t+1}\|_{W_{t+1}} \le \sigma_t(\delta)$$

with probability $1-\delta$. We then decompose $\mathbb{R}^K$ as $\mathbb{R}^K = 1_K \oplus \mathcal{H}$ where $\mathcal{H}$ is the hyperplane supported by 1_K. Recall that $\theta_*, \theta_{t+1} \in \Pi\mathbb{R}^{K\times d}$, for all $x \in \mathcal{X}$, by definition of Π, $\theta_{t+1}x$ and $\theta_* x$ are in $\mathcal{H}$. Therefore $\sum_{s=1}^{t} \|(\theta_{t+1} - \theta_*)x_s\|_{1_K 1_K^\top} = 0$. And we conclude that with probability $1-2\delta$

$$\|\theta_{t+1} - \theta_*\|_{\overline{W}_{t+1}} = \|\theta_{t+1} - \theta_*\|_{W_{t+1}} \le \sigma_t(\delta)\,.$$

□

E.3 Proof of Theorem 3

Theorem 3. *Let $\delta \in (0,1]$. Set τ, λ_0 as in Lemma 1, $\eta = 1$ and $\lambda = 144Kd$. Then, the regret of Algorithm 2 satisfies, with probability at least $1-2\delta$,*

$$\mathrm{Reg}_T \le \mathfrak{C}Rd\sqrt{KT/\kappa_*}\log(T/\delta) + \mathfrak{C}\kappa K^2 d^2 \log^2(T/\delta)$$

where $\mathfrak{C} > 0$ is a universal constant.

Proof. Throughout the proof we assume that

$$\mathrm{diam}_{\mathcal{X}}(\Theta) \leq 1 \qquad \text{and} \qquad \forall t \geq 1, \quad \|\theta_* - \theta'_{t+1}\|_{\overline{W}_{t+1}} \leq \sigma_t(\delta)$$

which is verified with probability $1 - 2\delta$ thanks to Lemma 1 and Lemma 5, we apply a Union Bound at the end. The regret of the exploration phase is smaller than

$$\tau \leq \mathfrak{C} K^{3/2} d^{3/2} \kappa \log^{3/2}(T)\,.$$

Let us now focus on the second phase of the algorithm.

Step 1: Using optimism. Using the definition of the optimistic reward we can bound the regret twice.

$$\begin{aligned}
\mathrm{Reg}_T(\text{Learning}) &:= \sum_{t=\tau+1}^{T} \rho^\top(\mu(\theta_* x_*) - \mu(\theta_* x_t)) \\
&\leq \sum_{t=\tau+1}^{T} \rho^\top \mu(\theta'_t x_*) + \varepsilon_{1,t}(x_*) + \varepsilon_{2,t}(x_*) - \rho^\top \mu(\theta_* x_t) && \text{(Prop. 2)} \\
&\leq \sum_{t=\tau+1}^{T} \rho^\top \mu(\theta'_t x_t) + \varepsilon_{1,t}(x_t) + \varepsilon_{2,t}(x_t) - \rho^\top \mu(\theta_* x_t) && \text{(Def. of } x_t\text{)} \\
&\leq 2\sum_{t=\tau+1}^{T} \varepsilon_{1,t}(x_t) + 2\sum_{t=\tau+1}^{T} \varepsilon_{2,t}(x_t) && \text{(Prop. 2)} \\
&\leq 2\sum_{t=1}^{T} \varepsilon_{1,t}(x_t) + 2\sum_{t=1}^{T} \varepsilon_{2,t}(x_t)
\end{aligned}$$

Step 2: Bounding the sum of $\varepsilon_{2,t}(x_t)$. We start by bounding the second sum

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_{2,t}(x_t) &= 3\sum_{t=1}^{T} R\sigma_t(\delta)^2 \|(I_K \otimes x_t^\top)\overline{W}_t^{-1/2}\|_2^2 \\
&\leq 3R\sigma_T(\delta)^2 \sum_{t=1}^{T} \|(I_K \otimes x_t^\top)\overline{W}_t^{-1/2}\|_2^2 \\
&= 3R\sigma_T(\delta)^2 \sum_{t=1}^{T} \|\overline{W}_t^{-1/2}(I_K \otimes x_t)\|_2^2 \\
&\leq 3R\sigma_T(\delta)^2 \sum_{t=1}^{T} \lambda_{\max}\left((I_K \otimes x_t^\top)\overline{W}_t^{-1}(I_K \otimes x_t)\right) \\
&= 3R\sigma_T(\delta)^2 \sum_{t=1}^{T} \lambda_{\max}\left(\left(I_K \otimes x_t x_t^\top\right)\overline{W}_t^{-1}\right) \\
&\leq 3R\sigma_T(\delta)^2 \sum_{t=1}^{T} \mathrm{Tr}\left(\left(I_K \otimes x_t x_t^\top\right)\overline{W}_t^{-1}\right) \\
&= 3R\kappa\sigma_T(\delta)^2 \sum_{t=1}^{T} \mathrm{Tr}\left(\left(\tfrac{1}{\kappa} I_K \otimes x_t x_t^\top\right)\overline{W}_t^{-1}\right)\,.
\end{aligned}$$

Let us define $U_t := \sum_{s=1}^{t} \frac{1}{\kappa} I_K \otimes x_s x_s^\top + \frac{\lambda}{2} I_{Kd}$. We have that $U_t \preccurlyeq \overline{W}_t$ for $\lambda \geq 2$. We have that

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_{2,t}(x_t) &\leq 3\kappa\sigma_T(\delta)^2 \sum_{t=1}^{T} \operatorname{Tr}\left((U_t - U_{t-1})U_t^{-1}\right) \\
&\leq 3R\kappa\sigma_T(\delta)^2 \sum_{t=1}^{T} \log \frac{|U_t|}{|U_{t-1}|} && \text{(Hazan et al., 2016, Lemma 4.5)} \\
&\leq 3R\kappa\sigma_T(\delta)^2 Kd \log\left(1 + \frac{T}{Kd\lambda\kappa}\right). && \text{(Lemma 15)}
\end{aligned}$$

Step 3: Decomposing the sum of $\varepsilon_{1,t}(x_t)$. Let us now focus on the first sum.

$$\begin{aligned}
&\sum_{t=1}^{T} \varepsilon_{1,t}(x_t) \\
&\quad := \sum_{t=1}^{T} \sigma_t(\delta) \|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_t' x_t)\rho\|_2 \\
&\quad \leq \sigma_T(\delta) \sum_{t=1}^{T} \|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_t' x_t)\rho\|_2 \\
&\quad \leq e\sigma_T(\delta) \sum_{t=1}^{T} \|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_t' x_t)^{1/2}\nabla\mu(\theta_* x_t)^{1/2}\rho\|_2 && \text{(Self-concordance)} \\
&\quad \leq e\sigma_T(\delta) \sum_{t=1}^{T} \|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_t' x_t)^{1/2}\|_2 \|\nabla\mu(\theta_* x_t)^{1/2}\rho\|_2 \\
&\quad \leq e\sigma_T(\delta) \sqrt{\sum_{t=1}^{T} \|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_t' x_t)^{1/2}\|_2^2} \sqrt{\sum_{t=1}^{T} \|\nabla\mu(\theta_* x_t)^{1/2}\rho\|_2^2}. && \text{(Cauchy-Schwarz)}
\end{aligned}$$

Once again we have two separate terms to bound. We start with the left term.

Step 4: Bounding the sum of $\|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_{t+1} x_t)^{1/2}\|_2^2$. First, we lower-bound $\overline{W}_{t+1}$:

$$\overline{W}_{t+1} = \sum_{s=1}^{t-1} \nabla\mu(\theta_{s+1} x_s) \otimes x_s x_s^\top + \sum_{s=1}^{t-1} 1_K 1_K^\top \otimes x_s x_s^\top + \lambda I_{Kd} \succcurlyeq \sum_{s=1}^{t-1} 1_K 1_K^\top \otimes x_s x_s^\top + \lambda I_{Kd}\,.$$

We use the following equivalent of $\frac{1_K 1_K^\top}{K}$ in the Loewner order sense:

$$e_{11} \preccurlyeq \frac{1_K 1_K^\top}{K} \preccurlyeq e_{11} \in \mathbb{R}^{K \times K}$$

to obtain:

$$\overline{W}_{t+1} \succcurlyeq K \sum_{s=1}^{t-1} e_{11} \otimes x_s x_s^\top + \lambda I_{Kd} = K \sum_{s=1}^{t-1} e_{11}^2 \otimes x_s x_s^\top + \lambda I_{Kd} = K \sum_{s=1}^{t-1} (e_{11} \otimes x_s)(e_{11} \otimes x_s)^\top + \lambda I_{Kd}\,.$$

Which is equivalent to

$$\overline{W}_{t+1} \succcurlyeq \sum_{s=1}^{t-1} \sum_{k=1}^{K} (e_{kk} \otimes x_s)(e_{kk} \otimes x_s)^\top + \lambda I_{Kd} = \sum_{s=1}^{t-1} I_K \otimes x_s x_s^\top + \lambda I_{Kd}\,.$$

Therefore we have

$$\|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_{t+1} x_t)^{1/2}\|_2^2 \leq \left\| \left(\sum_{s=1}^{t-1} I_K \otimes x_s x_s^\top + \lambda I_{Kd} \right)^{-1/2} (I_K \otimes x_t)\nabla\mu(\theta_{t+1} x_t)^{1/2} \right\|_2^2.$$

We now use that $\nabla\mu(\theta_{t+1}x_t) \preccurlyeq I_K$ and get

$$\|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_{t+1}x_t)^{1/2}\|_2^2 \leq \left\| \left(\sum_{s=1}^{t-1} I_K \otimes x_s x_s^\top + \lambda I_{Kd} \right)^{-1/2} (I_K \otimes x_t) \right\|_2^2 .$$

We now upper-bound the sum over T using a Trace-Determinant argument:

$$\begin{aligned}
\sum_{t=1}^{T} \|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_{t+1}x_t)^{1/2}\|_2^2 &\leq \sum_{t=1}^{T} \lambda_{\max}\left((I_K \otimes x_t^\top) \left(\sum_{s=1}^{t-1} I_K \otimes x_s x_s^\top + \lambda I_{Kd} \right)^{-1} (I_K \otimes x_s) \right) \\
&= \sum_{t=1}^{T} \lambda_{\max}\left((I_K \otimes x_t x_t^\top) \left(\sum_{s=1}^{t-1} I_K \otimes x_s x_s^\top + \lambda I_{Kd} \right)^{-1} \right) \\
&\leq \sum_{t=1}^{T} \lambda_{\max}\left(x_t x_t^\top \left(\sum_{s=1}^{t-1} x_s x_s^\top + \lambda I_d \right) \right) \\
&\leq \sum_{t=1}^{T} \mathrm{Tr}\left(x_t x_t^\top \left(\sum_{s=1}^{t-1} x_s x_s^\top + \lambda I_d \right)^{-1} \right) .
\end{aligned}$$

Let us define $M_t := \sum_{s=1}^{t} x_s x_s^\top + \frac{\lambda}{2} I_d$, we have that $M_t \preccurlyeq \sum_{s=1}^{t-1} x_s x_s^\top + \lambda I_d$ when $\lambda \geq 2$. We obtain

$$\begin{aligned}
\sum_{t=1}^{T} \|\overline{W}_t^{-1/2}(I_K \otimes x_t)\nabla\mu(\theta_{t+1}x_t)^{1/2}\|_2^2 &\leq \sum_{t=1}^{T} \mathrm{Tr}\left((M_t - M_{t-1}) M_t^{-1} \right) \\
&\leq \sum_{t=1}^{T} \log \frac{|M_t|}{|M_{t-1}|} && \text{(Hazan et al., 2016, Lemma 4.5)} \\
&\leq d \log\left(1 + \frac{T}{\lambda d} \right) && \text{(Lemma 15)} .
\end{aligned}$$

Step 5: Bounding the sum of $\|\nabla\mu(\theta_* x_t)^{1/2}\rho\|_2^2$. We add and subtract a term and get

$$\begin{aligned}
&\sum_{t=1}^{T} \|\nabla\mu(\theta_* x_t)^{1/2}\rho\|_2^2 \\
&\quad = \sum_{t=1}^{T} \langle \rho, \nabla\mu(\theta_* x_*)\rho \rangle + \sum_{t=1}^{T} \langle \rho, (\nabla\mu(\theta_* x_t) - \nabla\mu(\theta_* x_*))\rho \rangle \\
&\quad \leq R^2 T/\kappa_* + \sum_{t=1}^{T} \langle \rho, (\nabla\mu(\theta_* x_t) - \nabla\mu(\theta_* x_*))\rho \rangle .
\end{aligned}$$

We use the definition of $\nabla\mu(.)$ and get

$$\begin{aligned}
&\sum_{t=1}^{T}\langle\rho,(\nabla\mu(\theta_* x_t)-\nabla\mu(\theta_* x_*))\rho\rangle \\
&\quad=\sum_{t=1}^{T}\langle\rho,\operatorname{diag}(\mu(\theta_* x_t)-\mu(\theta_* x_*))\rho\rangle+\langle\rho,(\mu(\theta_* x_*)\mu(\theta_* x_*)^\top-\mu(\theta_* x_t)\mu(\theta_* x_t)^\top)\rho\rangle \\
&\quad\le\sum_{t=1}^{T}\langle\rho,2(\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t))1_K\rangle+\langle\rho,\mu(\theta_* x_*)\rangle^2-\langle\rho,\mu(\theta_* x_t)\rangle^2 \\
&\quad\le 2\sqrt{K}\sum_{t=1}^{T}(\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t))+\sum_{t=1}^{T}\langle\rho,\mu(\theta_* x_*)-\mu(\theta_* x_t)\rangle\langle\rho,\mu(\theta_* x_*)+\mu(\theta_* x_t)\rangle \\
&\quad\le 2\sqrt{K}\sum_{t=1}^{T}(\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t))+2\sum_{t=1}^{T}(\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t))\langle\rho,\mu(\theta_* x_*)+\mu(\theta_* x_t)\rangle \\
&\quad\le 2\sqrt{K}\sum_{t=1}^{T}(\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t))+4\sum_{t=1}^{T}(\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t)) \\
&\quad=(2\sqrt{K}+4)\sum_{t=1}^{T}(\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t))
\end{aligned}$$

where the first and third inequalities are by Proposition 2, the second and fourth inequalities are due to the Cauchy-Schwarz inequality.

Step 6: Putting everything together. Combining our previous results we get

$$\begin{aligned}
\operatorname{Reg}_T(\text{Learning}) \le& 2\sum_{t=1}^{T}\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t) \\
\le& 6R\kappa\sigma_T(\delta)^2 Kd\log\left(1+\tfrac{T}{Kd\lambda\kappa}\right) \\
&+e\sigma_T(\delta)\left[d\log\left(1+\tfrac{T}{\lambda d}\right)\right]^{1/2}\left[\tfrac{R^2T}{\kappa_*}+(2\sqrt{K}+4)\sum_{t=1}^{T}\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t)\right]^{1/2} \\
\le& 6R\kappa\sigma_T(\delta)^2 Kd\log\left(1+\tfrac{T}{Kd\lambda\kappa}\right) \\
&+e\sigma_T(\delta)\left[d\log\left(1+\tfrac{T}{\lambda d}\right)\right]^{1/2} R\sqrt{\tfrac{T}{\kappa_*}} \\
&+e\sigma_T(\delta)\left[d\log\left(1+\tfrac{T}{\lambda d}\right)\right]^{1/2}\left[(2\sqrt{K}+4)\sum_{t=1}^{T}\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t)\right]^{1/2}.
\end{aligned}$$

We use the fact that $x^2-bx-c\le 0 \implies x^2\le 2b^2+2c$ with $x^2=\sum_{t=1}^{T}\varepsilon_{1,t}(x_t)+\varepsilon_{2,t}(x_t)$ and get with probability $1-2\delta$

$$\begin{aligned}
\operatorname{Reg}_T(\text{Learning}) \le& 12R\kappa\sigma_T(\delta)^2 Kd\log\left(1+\tfrac{T}{Kd\lambda\kappa}\right) \\
&+2e\sigma_T(\delta)\left[d\log\left(1+\tfrac{T}{\lambda d}\right)\right]^{1/2} R\sqrt{\tfrac{T}{\kappa_*}} \\
&+e^2\sigma_T(\delta)^2(2\sqrt{K}+4)d\log\left(1+\tfrac{T}{\lambda d}\right) \\
\le& \mathfrak{C}\sqrt{K}d\log(T/\delta)R\sqrt{T/\kappa_*}+\mathfrak{C}(1+S)R\kappa K^2d^2\log^2(T/\delta).
\end{aligned}$$

where applying the Union Bound gives the result with probability $1-2\delta$.

□

F Proof of Theorem 4

Theorem 4. *For all $K \geq 2$, $d \geq 2$ and any algorithm, there exist $\theta_* \in \Pi\mathbb{R}^{K\times d}$ and $\rho \in \mathbb{R}_+^K$ with $\rho \notin \mathbb{R}1_K$ such that for $\mathcal{X} = \mathcal{S}_1(\mathbb{R}^d)$ and for any $T \geq d^2\kappa_*$, the cumulative regret satisfies* $\mathrm{Reg}_T \geq \Omega\big(Rd\sqrt{KT/\kappa_*}\big)$.

Proof. We use the canonical bandit probability space $(\Omega_t, \mathcal{F}_t, \mathbb{P}_{\pi\theta\rho})$ of Lattimore and Szepesvári (2020, Section 4.7). To simplify let us denote $\mathbb{P}_\theta = \mathbb{P}_{\pi\theta\rho}$ the probability of the random sequence $\{x_1, y_1, \ldots, x_T, y_T\}$ obtained by having the algorithm π interact with the environment (θ, ρ). The expectation $\mathbb{E}_\theta$ is computed with respect to the probability $\mathbb{P}_\theta$.

We start by defining an instance of a MNL bandit problem. Let $\theta_0 = \Pi M_0$ with $M_0 \in \mathbb{R}^{K\times d}$ be defined as follows

$$M_0 := \frac{1}{\sqrt{K+3}} \begin{bmatrix} 2 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \ldots & 0 \end{bmatrix}$$

and $\rho \in \mathbb{R}^K$ be defined by

$$\rho := \frac{R}{\sqrt{K+3}} \begin{bmatrix} 2 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Even though this defines a binary problem, we cannot directly apply the proof of Abeille et al. (2021) as for $K \geq 3$ our κ_* will be different than in the binary setting. Indeed the probability distributions of the reward are different, see the $\ln(K-1)$ term in Equation (14).

We define the action set by the sphere $\mathcal{X} = \mathcal{S}_1(\mathbb{R}^d)$. We show that a slight variation $\widetilde{M}$ of the matrix M_0 results in a regret lower-bounded by $\Omega(Rd\sqrt{KT/\kappa_*(\theta)})$. Let us define the set of perturbed matrices $\mathcal{M}$ by

$$\mathcal{M} := \left\{ M_0 + \varepsilon \sum_{i=2}^{d} v_i e_{1i} + \frac{\varepsilon}{2} \sum_{k=2}^{K} \sum_{i=2}^{d} v_i e_{ki} \quad , v \in \{-1,1\}^d \right\}$$

where $\varepsilon > 0$ is to be defined later. For now we only assume that

$$\varepsilon \leq \|[M_0]_1\|_2/\sqrt{d-1} = 2/\sqrt{(K+3)(d-1)}\,.$$

Note that we do not modify the first column. Let $x_*(\theta) := \arg\max_{x\in\mathcal{X}} \rho^\top \mu(\theta x)$. Let $M \in \mathbb{R}^{K\times d}$, as for $\theta = \Pi M$ we have $\mu(\theta x) = \mu(Mx)$, we may abuse the notation and write $x_*(M) = x_*(\theta)$. For every problem instance $(\theta = \Pi M \in \Pi\mathcal{M}, \rho, \mathcal{S}_1(\mathbb{R}^d))$, we have that $x_*(\theta) = [M]_1/\|[M]_1\|_2$.

We introduce a second set $\widetilde{\mathcal{M}} \subseteq \mathbb{R}^{K\times d}$ of matrices, which is in bijection with $\mathcal{M}$. This alternative set simplifies the presentation of the proof, but should be regarded as equivalent to $\mathcal{M}$. It is defined as follows:

$$\widetilde{\mathcal{M}} := \left\{ \begin{bmatrix} [M]_1/\|[M]_1\|_2 & \\ 0 \quad \ldots \quad 0 \\ \vdots \quad \ddots \quad \vdots \\ 0 \quad \ldots \quad 0 \end{bmatrix} : M \in \mathcal{M} \right\}.$$

We denote by $\gamma : \mathcal{M} \to \widetilde{\mathcal{M}}$ the canonical bijection from $\mathcal{M}$ to $\widetilde{\mathcal{M}}$. For all $M \in \mathcal{M}$ we have that $\arg\max_{x\in\mathcal{X}} \rho^\top \mu(Mx) = \arg\max_{x\in\mathcal{X}} \rho^\top \mu(\gamma(M)x)$.

We assume that for all $\widetilde{M} \in \widetilde{\mathcal{M}}$, $\mathrm{Reg}_T(\Pi\widetilde{M}) \leq Rd\sqrt{KT/\kappa_*}$, which can be done without loss of generality, since otherwise the lower-bound already holds. The proof then consists in showing that there exists a matrix $M_* \in \widetilde{\mathcal{M}}$ such that, for $\theta_* = \Pi M_*$, the regret is lower-bounded as $\mathrm{Reg}_T(\theta_*) \geq \mathfrak{C}Rd\sqrt{KT/\kappa_*}$.

Step 1: Lower-bounding by the optimum regime. In this step, we follow the idea of Proposition 6 from Abeille et al. (2021). Let $\widetilde{\theta} = \Pi\widetilde{M} \in \Pi\widetilde{\mathcal{M}}$, we lower-bound the regret $\mathrm{Reg}_T(\widetilde{\theta})$ by the derivative of the sigmoid function in the

optimum $\mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right)$. We first express the regret $\mathrm{Reg}_T(\widetilde{\theta})$ in terms of Bernoulli variables.

$$\begin{aligned}
\mathrm{Reg}_T(\widetilde{\theta}) &:= \sum_{t=1}^{T} \rho^\top \mu(\widetilde{\theta} x_*(\widetilde{\theta})) - \rho^\top \mu(\widetilde{\theta} x_t) \\
&= \sum_{t=1}^{T} \rho^\top \mu(\widetilde{M} x_*(\theta)) - \rho^\top \mu(\widetilde{M} x_t) \\
&= \sum_{t=1}^{T} \rho_1[\mathbb{P}_{\widetilde{\theta}}(\rho_1|x_*(\widetilde{\theta})) - \mathbb{P}_{\widetilde{\theta}}(\rho_1|x_t)] + \rho_2[\mathbb{P}_{\widetilde{\theta}}(\rho_2|x_*(\widetilde{\theta})) - \mathbb{P}_{\widetilde{\theta}}(\rho_2|x_t)] \\
&= \sum_{t=1}^{T} \rho_1[\mathbb{P}_{\widetilde{\theta}}(\rho_1|x_*(\widetilde{\theta})) - \mathbb{P}_{\widetilde{\theta}}(\rho_1|x_t)] + \rho_2[1 - \mathbb{P}_{\widetilde{\theta}}(\rho_1|x_*(\widetilde{\theta})) - 1 + \mathbb{P}_{\widetilde{\theta}}(\rho_1|x_t)]
\end{aligned}$$

where we have $\mathbb{P}_{\widetilde{\theta}}(\rho_1|x) = [\mu(\widetilde{M}x)]_1$ and $\mathbb{P}_{\widetilde{\theta}}(\rho_2|x) = \sum_{k=2}^{K}[\mu(\widetilde{M}x)]_k$. Using the definition of ρ we get

$$\mathrm{Reg}_T(\widetilde{\theta}) = \frac{R}{\sqrt{K+3}} \sum_{t=1}^{T} [\mathbb{P}_{\widetilde{\theta}}(\rho_1|x_*(\widetilde{\theta})) - \mathbb{P}_{\widetilde{\theta}}(\rho_1|x_t)]\,.$$

Substituting,

$$\begin{aligned}
\mathbb{P}_{\widetilde{\theta}}(\rho_1|x) = [\mu(\widetilde{M}x)]_1 = \frac{\exp([\widetilde{M}]_1^\top x)}{\exp([\widetilde{M}]_1^\top x) + (K-1)} \\
= \frac{1}{1 + \exp(-[\widetilde{M}]_1^\top x + \ln(K-1))} = \mu([\widetilde{M}]_1 \top x - \ln(K-1))\,. \quad (14)
\end{aligned}$$

Thus we get

$$\mathrm{Reg}_T(\widetilde{\theta}) \overset{(14)}{=} \frac{R}{\sqrt{K+3}} \sum_{t=1}^{T} \mu\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) - \mu\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\,.$$

We now apply the Mean-value Theorem:

$$\mathrm{Reg}_T(\widetilde{\theta}) = \frac{R}{\sqrt{K+3}} \sum_{t=1}^{T} \int_0^1 \mu'\left([\widetilde{M}]_1^\top (v x_*(\widetilde{\theta}) + (1-v)x_t) - \ln(K-1)\right) \mathrm{d}v \left([\widetilde{M}]_1^\top (x_*(\widetilde{\theta}) - x_t)\right)\,. \quad (15)$$

Using the self-concordance property (Sun and Tran-Dinh, 2019, Corollary 2) on μ' between $[\widetilde{M}]_1^\top x_t$ and $[\widetilde{M}]_1^\top x_*(\widetilde{\theta})$, we get

$$\begin{aligned}
\mathrm{Reg}_T(\widetilde{\theta}) &\geq \frac{R}{\sqrt{K+3}} \frac{1}{1 + \left\|[\widetilde{M}]_1^\top (x_*(\widetilde{\theta}) - x_t)\right\|_2} \sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right)\left([\widetilde{M}]_1^\top (x_*(\widetilde{\theta}) - x_t)\right) \\
&\geq \frac{R}{\sqrt{K+3}} \frac{1}{1 + \left\|[\widetilde{M}]_1\right\|_2 \|(x_*(\widetilde{\theta}) - x_t)\|_2} \sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right)\left([\widetilde{M}]_1^\top (x_*(\widetilde{\theta}) - x_t)\right) \\
&\geq \frac{R}{\sqrt{K+3}} \frac{1}{1 + 2\|[\widetilde{M}]_1\|_2} \sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right)\left([\widetilde{M}]_1^\top (x_*(\widetilde{\theta}) - x_t)\right) \\
&\geq \frac{R}{3\sqrt{K+3}} \sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right)\left([\widetilde{M}]_1^\top (x_*(\widetilde{\theta}) - x_t)\right)
\end{aligned}$$

where the second inequality is by Cauchy-Schwarz inequality, the third inequality is because the actions are in the sphere $\mathcal{S}_1(\mathbb{R}^d)$, and the last inequality is because $\|[\widetilde{M}]_1\|_2 = 1$. Using the definition of x_* we have that

$$([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - [\widetilde{M}]_1^\top x_t) = \|[\widetilde{M}]_1\|_2 \left(1 - \tfrac{[\widetilde{M}]_1^\top}{\|[\widetilde{M}]_1\|_2} x_t\right) = \|[\widetilde{M}]_1\|_2 \tfrac{1}{2} \|x_*(\widetilde{\theta}) - x_t\|_2^2$$

where the last equality is due to $1 - x^\top y = \frac{1}{2}\|x-y\|_2^2$ for all $x, y \in \mathcal{S}_1(\mathbb{R}^d)$. Thus we obtain

$$\begin{aligned}\mathrm{Reg}_T(\widetilde{\theta}) &\geq \frac{R\|[\widetilde{M}]_1\|_2}{6\sqrt{K+3}} \sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) \|x_*(\widetilde{\theta}) - x_t\|_2^2 \\ &= \frac{R}{6\sqrt{K+3}} \sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) \|x_*(\widetilde{\theta}) - x_t\|_2^2 \qquad (16) \\ &\geq \frac{R}{6\sqrt{K+3}} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) \sum_{t=1}^{T}\sum_{i=2}^{d} [x_*(\widetilde{\theta}) - x_t]_i^2 \,.\end{aligned}$$

Let us denote $M = \gamma^{-1}(\widetilde{M}) \in \mathcal{M}$ and $\theta = \Pi M$, we have $x_*(\theta) = x_*(\widetilde{\theta})$. Thus

$$\mathrm{Reg}_T(\widetilde{\theta}) \geq \frac{R}{6\sqrt{K+3}} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) \sum_{t=1}^{T}\sum_{i=2}^{d} [x_*(\theta) - x_t]_i^2 \,. \qquad (17)$$

Let us define the event $A_i(\theta)$ for all $i \in [\![d]\!]$ and all $\theta \in \Pi\mathbb{R}^{K\times d}$ as

$$A_i(\theta) := \left\{ [x_*(\theta) - x_*(\theta_0)]_i \cdot \left[x_*(\theta_0) - \frac{1}{T}\sum_{t=1}^{T} x_t \right] \geq 0 \right\} .$$

We bound the regret of any $\widetilde{\theta} = \Pi\widetilde{M} \in \Pi\widetilde{\mathcal{M}}$ using the event $A_i(\theta)$. By applying Lemma 3 of (Abeille et al., 2021) we obtain

$$\sum_{t=1}^{T}\sum_{i=2}^{d} [x_*(\theta) - x_t]_i^2 \geq \frac{3T\varepsilon^2}{8\|[M]_1\|_2^2} \sum_{i=2}^{d} \mathbb{P}_{\widetilde{\theta}}(A_i(\theta)) \,.$$

We apply this result in Equation (17) to get

$$\begin{aligned}\mathbb{E}_{\widetilde{\theta}}\left[\mathrm{Reg}_T(\widetilde{\theta})\right] &\geq \frac{RT\varepsilon^2}{16\sqrt{K+3}\|[M]_1\|_2^2} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) \sum_{i=2}^{d} \mathbb{P}_{\widetilde{\theta}}(A_i(\theta)) \\ &= \frac{RT\varepsilon^2\sqrt{K+3}}{64} \mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) \sum_{i=2}^{d} \mathbb{P}_{\widetilde{\theta}}(A_i(\theta)) \,. \qquad (18)\end{aligned}$$

Step 2: Showing that $\mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) = (K+3)/\kappa_*(\widetilde{\theta})$.

Recall that $\kappa_*(\widetilde{\theta})$ is defined by

$$\kappa_*(\widetilde{\theta})^{-1} := \frac{\rho^\top \nabla\mu(\widetilde{\theta}x_*(\widetilde{\theta}))\rho}{\|\rho\|_2^2} = \frac{\rho^\top \mathrm{diag}(\mu(\widetilde{\theta}x_*(\widetilde{\theta})))\rho - \rho^\top \mu(\widetilde{\theta}x_*(\widetilde{\theta}))\mu(\widetilde{\theta}x_*(\widetilde{\theta}))^\top\rho}{R^2} \,.$$

This develops into

$$\begin{aligned}\frac{R^2}{\kappa_*(\widetilde{\theta})} = \rho_1^2\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1 + \sum_{k=2}^{K}\rho_k\left[\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k \sum_{i=2}^{K}\rho_i\left(\delta_{ik} - \mu(\widetilde{\theta}x_*(\widetilde{\theta}))_i\right)\right] \\ - 2\rho_1\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1 \sum_{k=2}^{K}\rho_k\mu(\widetilde{\theta}x_*(\widetilde{\theta})_k \,.\end{aligned}$$

We start by using the definition of μ':

$$\begin{aligned}\mu'([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)) &= \mathbb{P}_{\widetilde{\theta}}(\rho_2|x_*(\widetilde{\theta}))(1 - \mathbb{P}_{\widetilde{\theta}}(\rho_2|x_*(\widetilde{\theta}))) \\ &= \sum_{k=2}^{K}\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\left(1 - \sum_{i=2}^{K}\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_i\right) \\ &= \sum_{k=2}^{K}\left[\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\sum_{i=2}^{K}\delta_{ik} - \mu(\widetilde{\theta}x_*(\widetilde{\theta}))_i\right] \\ &= \sum_{k=2}^{K}\frac{\rho_k}{\rho_k}\left[\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\sum_{i=2}^{K}\frac{\rho_i}{\rho_i}\left(\delta_{ik} - \mu(\widetilde{\theta}x_*(\widetilde{\theta}))_i\right)\right] \,.\end{aligned}$$

Using the definition of ρ we get

$$\mu'([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)) = \frac{K+3}{R^2}\sum_{k=2}^{K}\rho_k\left[\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\sum_{i=2}^{K}\rho_i\left(\delta_{ik}-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_i\right)\right]. \tag{19}$$

Let us now consider $4\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1)$:

$$\begin{aligned}
4\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1) &= 4\frac{\rho_1^2}{\rho_1^2}\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1)\\
&= 4\frac{K+3}{4R^2}\rho_1^2\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1)\\
&= \frac{K+3}{R^2}\rho_1^2\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1).
\end{aligned}$$

We can write $4\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1)$ differently to obtain:

$$\begin{aligned}
4\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1) &= 4\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1\sum_{k=2}^{K}\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\\
&= 4\frac{\rho_1}{\rho_1}\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1\sum_{k=2}^{K}\frac{\rho_k}{\rho_k}\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\\
&= 2\cdot 2\frac{K+3}{2R^2}\rho_1\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1\sum_{k=2}^{K}\rho_k\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\\
&= 2\frac{K+3}{R^2}\rho_1\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1\sum_{k=2}^{K}\rho_k\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k.
\end{aligned}$$

We add and subtract $4\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1)$ in Equation (19) to obtain the desired result:

$$\begin{aligned}
&\mu'([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1))\\
&\quad= \frac{K+3}{R^2}\sum_{k=2}^{K}\rho_k\left[\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\sum_{i=2}^{K}\rho_i\left(\delta_{ik}-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_i\right)\right]\\
&\qquad+\frac{K+3}{R^2}\rho_1^2\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1(1-\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1) - 2\frac{K+3}{R^2}\rho_1\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_1\sum_{k=2}^{K}\rho_k\mu(\widetilde{\theta}x_*(\widetilde{\theta}))_k\\
&\quad= \frac{K+3}{\kappa_*(\widetilde{\theta})}.
\end{aligned}$$

By substituting into Equation (18) we obtain

$$\mathrm{Reg}_T(\widetilde{\theta}) \geq \frac{RT\varepsilon^2(K+3)^{3/2}}{64\kappa_*(\widetilde{\theta})}\sum_{i=2}^{d}\mathbb{P}_{\widetilde{\theta}}(A_i(\theta)). \tag{20}$$

Step 3: Averaging Hammer and Average Relative Entropy. Let us define $\Xi := \Pi\widetilde{\mathcal{M}}$. In order to find a $\widetilde{\theta}\in\Xi$ with a large regret lower-bound, we use the averaging hammer technique as in Lattimore and Szepesvári (2020, Section 24.1). Let us recall Lemma 4 from (Abeille et al., 2021), the following holds:

$$\frac{1}{|\Xi|}\sum_{\widetilde{\theta}\in\Xi}\sum_{i=2}^{d}\mathbb{P}_{\widetilde{\theta}}(A_i(\theta)) \geq \frac{d}{4} - \frac{\sqrt{d}}{2}\sqrt{\frac{1}{|\Xi|}\sum_{\widetilde{\theta}\in\Xi}\sum_{i=2}^{d}KL\left(\mathbb{P}_{\widetilde{\theta}},\mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right)}. \tag{21}$$

where the flipping operator Flip_i is defined by

$$[\mathrm{Flip}_i(\theta)]_i = -[\theta]_i \qquad \text{and} \qquad [\mathrm{Flip}_i(\theta)]_j = [\theta]_j \text{ for all } j\neq i.$$

We study the average relative entropy and upper-bound it by the regret. Let us denote $P_{x_t}^{\widetilde{\theta}} = \mathbb{P}_{\widetilde{\theta}}(\cdot|x)$. Using the Divergence Decomposition Lemma (Lattimore and Szepesvári, 2020, Exercise 15.8(b)) and the fact that the χ^2-divergence upper-bounds the KL divergence we get

$$KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) = \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} KL\left(P_{x_t}^{\widetilde{\theta}}, P_{x_t}^{\mathrm{Flip}_i(\widetilde{\theta})}\right)\right] \leq \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} D_{\chi^2}\left(P_{x_t}^{\widetilde{\theta}}, P_{x_t}^{\mathrm{Flip}_i(\widetilde{\theta})}\right)\right]. \tag{22}$$

Remember that we have

$$\mathbb{P}_{\widetilde{\theta}}(\rho_1|x) = \frac{1}{1+\exp(-[\widetilde{M}]_1^\top x + \ln(K-1))} = \mu\left([\widetilde{M}]_1^\top x - \ln(K-1)\right).$$

Thus the multinomial variables $P_{x_t}^{\widetilde{\theta}}$ and $P_{x_t}^{\mathrm{Flip}_i(\widetilde{\theta})}$ can be written as Bernoulli variables. Therefore by substituting in Equation (22) we have that

$$\begin{aligned}
&KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \\
&\quad\leq \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} D_{\chi^2}\left(\mathrm{Bernoulli}\left(\mu\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\right), \mathrm{Bernoulli}\left(\mu\left([\mathrm{Flip}_i(\widetilde{M})]_1^\top x_t - \ln(K-1)\right)\right)\right)\right].
\end{aligned}$$

Using the expression of the χ^2-divergence for Bernoulli random variables gives

$$KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \leq \mathbb{E}_{\theta}\left[\sum_{t=1}^{T} \frac{\left(\mu\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right) - \mu\left([\mathrm{Flip}_i(\widetilde{M})]_1^\top x_t - \ln(K-1)\right)\right)^2}{\mu'\left([\mathrm{Flip}_i(\widetilde{M})]_1^\top x_t - \ln(K-1)\right)}\right].$$

We apply the Mean-value Theorem and get

$$KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \leq \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \frac{\left(\int_0^1 \mu'\left(\left(v[\widetilde{M}]_1 + (1-v)[\mathrm{Flip}_i(\widetilde{M})]_1\right)^\top x_t - \ln(K-1)\right)\mathrm{d}v\right)^2}{\mu'\left([\mathrm{Flip}_i(\widetilde{M})]_1^\top x_t - \ln(K-1)\right)} \left(([\widetilde{M}]_1 - [\mathrm{Flip}_i(\widetilde{M})]_1)^\top x_t\right)^2\right].$$

Now applying the self-concordance property gives

$$\begin{aligned}
&\int_0^1 \mu'\left(\left(v[\widetilde{M}]_1 + (1-v)[\mathrm{Flip}_i(\widetilde{M})]_1\right)^\top x_t - \ln(K-1)\right)\mathrm{d}v \\
&\qquad\leq \mu'\left([\mathrm{Flip}_i(\widetilde{M})]_1^\top x_t - \ln(K-1)\right)\exp\left(\left|([\widetilde{M}]_1 - [\mathrm{Flip}_i(\widetilde{M})]_1)^\top x_t\right|\right)
\end{aligned}$$

and

$$\begin{aligned}
&\int_0^1 \mu'\left(\left(v[\widetilde{M}]_1 + (1-v)[\mathrm{Flip}_i(\widetilde{M})]_1\right)^\top x_t - \ln(K-1)\right)\mathrm{d}v \\
&\qquad\leq \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\exp\left(\left|([\widetilde{M}]_1 - [\mathrm{Flip}_i(\widetilde{M})]_1)^\top x_t\right|\right).
\end{aligned}$$

Thus we obtain

$$\begin{aligned}
&KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \\
&\quad\leq \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t/2 - \ln(K-1)\right)\left(([\widetilde{M}]_1 - [\mathrm{Flip}_i(\widetilde{M})]_1)^\top x_t\right)^2 \exp\left(2\left|([\widetilde{M}]_1 - [\mathrm{Flip}_i(\widetilde{M})]_1)^\top x_t\right|\right)\right] \\
&\quad\leq \exp(2\varepsilon)\mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\left(([\widetilde{M}]_1 - [\mathrm{Flip}_i(\widetilde{M})]_1)^\top x_t\right)^2\right] \\
&\quad\leq \exp(2\varepsilon)4\varepsilon^2\mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_t]_i^2\right].
\end{aligned}$$

We add and subtract $x_*(\widetilde{\theta})$ and apply Young's Inequality:

$$\begin{aligned}
&KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \\
&\quad \leq \exp(2\varepsilon) 4\varepsilon^2 \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_*(\widetilde{\theta}) - x_t - x_*(\widetilde{\theta})]_i^2\right] \\
&\quad \leq 8\exp(2\varepsilon)\varepsilon^2 \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_*(\widetilde{\theta}) - x_t]_i^2 + \sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_*(\widetilde{\theta})]_i^2\right].
\end{aligned}$$

Thus by summing over d we obtain

$$\begin{aligned}
&\sum_{i=2}^{d} KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \\
&\quad \leq 8\exp(2\varepsilon)\varepsilon^2 \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T}\sum_{i=2}^{d} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_t - x_*(\widetilde{\theta})]_i^2 + \sum_{t=1}^{T}\sum_{i=2}^{d} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_*(\widetilde{\theta})]_i^2\right] \\
&\quad \leq 8\exp(2\varepsilon)\varepsilon^2 \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_t - x_*(\widetilde{\theta})]_i^2 + \sum_{t=1}^{T}\sum_{i=2}^{d} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_*(\widetilde{\theta})]_i^2\right] \\
&\quad = 8\exp(2\varepsilon)\varepsilon^2 \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\|x_t - x_*(\widetilde{\theta})\|_2^2 + \sum_{t=1}^{T}\sum_{i=2}^{d} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)[x_*(\widetilde{\theta})]_i^2\right] \\
&\quad \leq 8\exp(2\varepsilon)\varepsilon^2 \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\|x_t - x_*(\widetilde{\theta})\|_2^2 + \frac{(d-1)\varepsilon^2}{\|[M_0]_1\|_2^2 + (d-1)\varepsilon^2}\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\right] \\
&\quad \leq 8\exp(2\varepsilon)\varepsilon^2 \mathbb{E}_{\widetilde{\theta}}\left[\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\|x_t - x_*(\widetilde{\theta})\|_2^2 + \frac{K+3}{4}(d-1)\varepsilon^2\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\right]
\end{aligned} \tag{23}$$

where for the second to last inequality we use the fact that $\|[M_0]_1\|_2 = 2/\sqrt{K+3}$.

Step 4: Bounding the First Term of the Average Entropy.
In this step, we upper-bound $\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\|x_t - x_*(\widetilde{\theta})\|_2^2$ using $\mathrm{Reg}_T(\widetilde{\theta})$. We follow the Step 1 up to Equation (15) and get

$$\mathrm{Reg}_T(\widetilde{\theta}) \geq \frac{R}{\sqrt{K+3}}\sum_{t=1}^{T}\int_0^1 \mu'\left([\widetilde{M}]_1^\top(v x_*(\widetilde{\theta}) + (1-v)x_t) - \ln(K-1)\right)\mathrm{d}v\left([\widetilde{M}]_1^\top(x_*(\widetilde{\theta}) - x_t)\right).$$

We apply the self-concordance property and obtain

$$\mathrm{Reg}_T(\widetilde{\theta}) \geq \frac{R}{\sqrt{K+3}}\frac{1}{1 + \left\|[\widetilde{M}]_1^\top(x_*(\widetilde{\theta}) - x_t)\right\|_2}\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\left([\widetilde{M}]_1^\top(x_*(\widetilde{\theta}) - x_t)\right).$$

We now follow our previous computations between Equation (15) and (16) to get

$$\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\|x_t - x_*(\widetilde{\theta})\|_2^2 \leq 6\frac{\sqrt{K+3}}{R}\mathrm{Reg}_T(\widetilde{\theta}).$$

Step 5: Bounding the Second Term of the Average Entropy. In this step, we upper-bound $\sum_{t=1}^{T} \mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)$ using $\mathrm{Reg}_T(\widetilde{\theta})$ and $\kappa_*(\widetilde{\theta})$. We apply a Taylor decomposition with integral remain-

der:

$$\begin{aligned}
&\sum_{t=1}^{T}\mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\\
&\quad= \sum_{t=1}^{T}\mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right)\\
&\qquad + \int_0^1 \mu''\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) + v[\widetilde{M}]_1^\top (x_t - x_*(\widetilde{\theta})) - \ln(K-1)\right)\mathrm{d}v\left([\widetilde{M}]_1^\top (x_t - x_*(\widetilde{\theta}))\right)\\
&\quad\le \sum_{t=1}^{T}\mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right)\\
&\qquad + \left|\int_0^1 \mu''\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) + v[\widetilde{M}]_1^\top (x_t - x_*(\widetilde{\theta})) - \ln(K-1)\right)\mathrm{d}v\right| \cdot \left|[\widetilde{M}]_1^\top (x_t - x_*(\widetilde{\theta}))\right| .
\end{aligned}$$

Using the facts that for the sigmoid $|\mu''| \le \mu'$ and $[\widetilde{M}]_1^\top x_*(\widetilde{\theta}) \ge [\widetilde{M}]_1^\top x_t$ we get

$$\begin{aligned}
&\sum_{t=1}^{T}\mu'\left([\widetilde{M}]_1^\top x_t - \ln(K-1)\right)\\
&\quad\le \sum_{t=1}^{T}\mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right)\\
&\qquad + \int_0^1 \mu'\left([\widetilde{M}]_1^\top (v x_*(\widetilde{\theta}) + (1-v)x_t) - \ln(K-1)\right)\mathrm{d}v \cdot \left([\widetilde{M}]_1^\top (x_*(\widetilde{\theta}) - x_t)\right) .
\end{aligned}$$

The first term of the sum can be rewritten using Step 3:

$$\sum_{t=1}^{T}\mu'\left([\widetilde{M}]_1^\top x_*(\widetilde{\theta}) - \ln(K-1)\right) = \frac{T(K+3)}{\kappa_*(\widetilde{\theta})} .$$

The second term already appears in Equation (15) and is therefore bounded by

$$\sum_{t=1}^{T}\int_0^1 \mu'\left([\widetilde{M}]_1^\top (v x_*(\widetilde{\theta}) + (1-v)x_t) - \ln(K-1)\right)\mathrm{d}v \cdot \left([\widetilde{M}]_1^\top (x_*(\widetilde{\theta}) - x_t)\right) \le \frac{\sqrt{K+3}}{R}\mathrm{Reg}_T(\widetilde{\theta}) .$$

Step 6: Putting Everything Together. We are now ready to carry out the final step of the proof. We apply Steps 4 and 5 and substitute them in Equation (23)

$$\sum_{i=2}^{d} KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \le 8\exp(2\varepsilon)\varepsilon^2\left[6\tfrac{\sqrt{K+3}}{R}\mathrm{Reg}_T(\widetilde{\theta}) + \tfrac{K+3}{4}(d-1)\varepsilon^2\left(\frac{T(K+3)}{\kappa_*(\widetilde{\theta})} + \tfrac{\sqrt{K+3}}{R}\mathrm{Reg}_T(\widetilde{\theta})\right)\right] .$$

Using our assumption on $\widetilde{\mathcal{M}}$, we can now upper-bound $\mathrm{Reg}_T(\widetilde{\theta})$ by $Rd\sqrt{KT/\kappa_*(\widetilde{\theta})}$. We obtain

$$\sum_{i=2}^{d} KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \le 8\exp(2\varepsilon)\varepsilon^2\left[6(K+3)d\sqrt{\frac{T}{\kappa_*(\widetilde{\theta})}} + \tfrac{K+3}{4}(d-1)\varepsilon^2\left(\frac{T(K+3)}{\kappa_*(\widetilde{\theta})} + (K+3)d\sqrt{\frac{T}{\kappa_*(\widetilde{\theta})}}\right)\right] .$$

Thus by taking the average over $\widetilde{\theta} \in \Xi$ we have:

$$\frac{1}{|\Xi|}\sum_{\widetilde{\theta}\in\Xi}\sum_{i=2}^{d} KL\left(\mathbb{P}_{\widetilde{\theta}}, \mathbb{P}_{\mathrm{Flip}_i(\widetilde{\theta})}\right) \le 8\exp(2\varepsilon)\varepsilon^2\left[6(K+3)d\sqrt{\frac{T}{\kappa_*(\widetilde{\theta})}} + \tfrac{K+3}{4}(d-1)\varepsilon^2\left(\frac{T(K+3)}{\kappa_*(\widetilde{\theta})} + (K+3)d\sqrt{\frac{T}{\kappa_*(\widetilde{\theta})}}\right)\right] .$$

Hence by substituting in Equation (21) we get

$$\begin{aligned}
&\frac{2}{|\Xi|}\sum_{\theta\in\Xi}\sum_{i=2}^{d}\mathbb{P}_{\widetilde{\theta}}[A_i(M)]\\
&\qquad\ge \tfrac{d}{2} - \sqrt{\tfrac{d-1}{2}}\sqrt{8\exp(2\varepsilon)\varepsilon^2\left[6(K+3)d\sqrt{\frac{T}{\kappa_*(\widetilde{\theta})}} + \tfrac{K+3}{4}(d-1)\varepsilon^2\left(\frac{T(K+3)}{\kappa_*(\widetilde{\theta})} + (K+3)d\sqrt{\frac{T}{\kappa_*(\widetilde{\theta})}}\right)\right]} .
\end{aligned}$$

If this is true for the average over Ξ, then there exists at least one $\theta_* = \Pi M_* \in \Xi$ such that

$$\sum_{i=2}^{d} \mathbb{P}_{\theta_*}[A_i(M_*)] \geq \frac{d}{2} - \sqrt{\frac{d-1}{2}} \sqrt{8\exp(2\varepsilon)\varepsilon^2 \left[6(K+3)d\sqrt{\frac{T}{\kappa_*(\widetilde{\theta}_*)}} + \frac{K+3}{4}(d-1)\varepsilon^2 \left(\frac{T(K+3)}{\kappa_*(\widetilde{\theta}_*)} + (K+3)d\sqrt{\frac{T}{\kappa_*(\widetilde{\theta}_*)}}\right)\right]}.$$

We substitute in Equation (18) and get

$$\begin{aligned}
&\mathbb{E}_{\theta_*}[\mathrm{Reg}_T(\theta_*)] \\
&\geq \frac{R(K+3)^{3/2}T\varepsilon^2}{64\kappa_*(\theta_*)} \left[\frac{d}{2} - \sqrt{\frac{d-1}{2}}\left[8\exp(2\varepsilon)\varepsilon^2\left[6(K+3)d\sqrt{\frac{T}{\kappa_*(\theta_*)}} + \frac{K+3}{4}(d-1)\varepsilon^2\left(\frac{T(K+3)}{\kappa_*(\theta_*)} + (K+3)d\sqrt{\frac{T}{\kappa_*(\theta_*)}}\right)\right]\right]^{1/2}\right] \\
&\geq \frac{R(K+3)^{3/2}T\varepsilon^2}{64\kappa_*(\theta_*)} \left[\frac{d}{2} - \frac{d}{2}\sqrt{8\exp(2\varepsilon)}\left[6\varepsilon^2(K+3)\sqrt{\frac{T}{\kappa_*(\theta_*)}} + \varepsilon^4\frac{(K+3)^2}{4}\frac{T}{\kappa_*(\theta_*)} + \varepsilon^4\frac{(K+3)^2}{4}d\sqrt{\frac{T}{\kappa_*(\theta_*)}}\right]^{1/2}\right].
\end{aligned}$$

We choose $\varepsilon^2 = c(K+3)^{-1}\sqrt{\kappa_*(\theta_*)/T}$ with $c = 0.01$ and get

$$\mathbb{E}_{\theta_*}[\mathrm{Reg}_T(\theta_*)] \geq \frac{cRd\sqrt{(K+3)T}}{64\sqrt{\kappa_*(\theta_*)}} \left[1 - \sqrt{8\exp(2\sqrt{c}(K+3)^{-1/2}[\kappa_*(\theta_*)/T]^{1/4})}\left[6c + \frac{1}{4}c^2 + \frac{1}{4}c^2 d\sqrt{\frac{T}{\kappa_*(\theta_*)}}\right]^{1/2}\right].$$

When $T \geq d^2\kappa_*(\theta_*)$ we have that

$$\mathbb{E}_{\theta_*}[\mathrm{Reg}_T(\theta_*)] \geq \frac{cRd\sqrt{(K+3)T}}{64\sqrt{\kappa_*(\theta_*)}} \left[1 - \sqrt{8\exp(\sqrt{c})}\left[6c + \frac{1}{4}c^2 + \frac{1}{4}c^2\right]^{1/2}\right] \geq \frac{Rd\sqrt{(K+3)T}}{25000\sqrt{\kappa_*(\theta_*)}}.$$

□

G Removing the Exploration

In this section we introduce a variant of our algorithm with an adaptive exploration and prove its regret bound.

G.1 Proof of Theorem 6

In this section we prove Theorem 6, the regret upper-bound of Algorithm 3. We start by studying the exploration part of the algorithm.

G.1.1 Analysis of the Adaptive Exploration

We start by showing that at each iteration of the algorithm, the set $\mathcal{W}_t(\delta)$ is a confidence set.

Lemma 9. *Let $\delta \in (0,1]$, $\eta^w = (1+\sqrt{6}S)/2$ and $\lambda^w = 144\eta^w Kd$. Let us define $\beta_t(\delta) = 4S\sqrt{Kd\log(t/\delta)} + 2S\sqrt{\lambda^w}$. Then we have with probability $1-\delta$, for all $t \geq 1$,*

$$\theta_* \in \mathcal{W}_t(\delta).$$

Proof. Let $t \in [\![T]\!]$. For all $x \in \mathcal{X}_t$, using Cauchy-Schwarz inequality we have

$$\max_{\theta \in \mathcal{W}} \|(\theta - \theta_*)x\|_2 \leq \max_{\theta \in \mathcal{W}} \|\theta - \theta_*\|_2 \|x\|_2 \leq 2S.$$

We apply (Lee and Oh, 2025, Theorem 4.2) with $\alpha = 2S$ and get $\beta_t(\delta) = 4S\sqrt{Kd\log(t/\delta)} + 2S\sqrt{\lambda^w}$. □

We now upper-bound the number of exploration steps to show it is negligible in the regret.

Lemma 10. *Let T^w the set of exploration steps. We have*

$$|T^w| \leq 2\tau_t^2 \kappa Kd \log\left(1 + \frac{T}{Kd\lambda^w}\right).$$

Algorithm 3: Using an adaptive exploration

Input: regularisation parameters λ^w, λ, learning rate η^w
Init: $H_0^w = \lambda^w I_{Kd}, H_1 = \lambda I_{Kd}$
for *each time step* t *in* $1 \dots T$ **do**
  Get action set $\mathcal{X}_t \subseteq \mathcal{X}$
  if $\max_{x \in \mathcal{X}_t} \|I_K \otimes x\|^2_{(H^w_{t-1})^{-1}} \geq 1/\tau_t^2$ **then**
    Play $x_t = \arg\max_{x \in \mathcal{X}_t} \|I_K \otimes x\|^2_{(H^w_{t-1})^{-1}}$
    Observe $y_t \sim \mu(\theta_* x_t)$
    Get reward ρ_{y_t}
    $\widetilde{H}^w_t \leftarrow H_{t-1} + \frac{\eta^w}{\kappa} I_K \otimes x_t x_t^\top$
    $\theta^w_{t+1} \leftarrow \arg\min_{\theta \in \mathbb{R}^{K \times d}} \langle \nabla \mu(\theta^w_t), \theta \rangle + \frac{1}{2\eta^w} \|\theta^w_t - \theta\|^2_{\widetilde{H}^w_t}$
    $H^w_t \leftarrow H^w_{t-1} + \frac{1}{\kappa} I_K \otimes x_t x_t^\top$
    $\mathcal{W}_{t+1}(\delta) \leftarrow \{\theta \in \mathbb{R}^{K \times d} : \|\theta - \theta^w_{t+1}\|_{H^w_t} \leq \beta_{t+1}(\delta)\}$
  else
    Play $x_t = \arg\max_{x \in \mathcal{X}_t} \widetilde{r}_t(x)$ with $\widetilde{r}_t(x)$ defined in Eq. (7)
    $\widetilde{H}_{t+1} \leftarrow H_t + \nabla \mu(\theta_t x_t) \otimes x_t x_t^\top$
    $\theta_{t+1} \leftarrow \arg\min_{\theta \in \mathcal{W}_t(\delta)} \langle \nabla \ell_{t+1}(\theta_t), \theta \rangle + \|\theta - \theta_t\|^2_{\widetilde{H}_{t+1}}$
    $H_{t+1} \leftarrow H_t + \nabla \mu(\theta_{t+1} x_t) \otimes x_t x_t^\top$
    $\overline{H}_{t+1} \leftarrow H_{t+1} + 1_K 1_K^\top \otimes x_t x_t^\top$
    $\mathcal{W}_{t+1}(\delta) \leftarrow \mathcal{W}_t(\delta)$
    $\theta^w_{t+1} \leftarrow \theta^w_t$
    $H^w_t \leftarrow H^w_{t-1}$
  end
end

Proof. We start with Trace-Determinant argument to upper-bound the following sum:

$$\begin{aligned}
&\sum_{t \in T^w} \max_{x \in \mathcal{X}_t} \|I_K \otimes x\|^2_{(H^w_{t-1})^{-1}} \\
&\quad = \sum_{t \in T^w} \|I_K \otimes x_t\|^2_{(H^w_{t-1})^{-1}} \\
&\quad = \kappa \sum_{t \in T^w} \frac{1}{\kappa} \|I_K \otimes x_t\|^2_{(H^w_{t-1})^{-1}} \\
&\quad = \kappa \sum_{t \in T^w} \|\kappa^{-1/2} I_K \otimes x_t\|^2_{(H^w_{t-1})^{-1}} \\
&\quad \leq 2\kappa K d \log\left(1 + \frac{T}{Kd\lambda^w}\right). \qquad \text{(Abbasi-Yadkori et al., 2011, Lemma 10)}
\end{aligned}$$

We now lower-bound this sum using the exploration rule:

$$\sum_{t \in T^w} \max_{x \in \mathcal{X}_t} \|I_K \otimes x\|^2_{(H^w_{t-1})^{-1}} \geq \sum_{t \in T^w} \frac{1}{\tau_t^2} = |T^w| \frac{1}{\tau_t^2}.$$

Therefore we have

$$|T^w| \leq 2\tau_t^2 \kappa K d \log\left(1 + \frac{T}{Kd\lambda^w}\right).$$

□

Finally we bound the diameters of the confidence sets $\mathcal{W}_t(\delta)$. It will allow us to leverage the self-concordance property for a constant cost.

Lemma 11. *Let us define* $\tau_t = 2\sqrt{6}\beta_t(\delta)$. *Let* $\delta \in (0, 1]$, *with probability* $1 - \delta$, *for all* $t \geq 1$ *we have*

$$\max_{x \in \mathcal{X}_t} \max_{\theta_1, \theta_2 \in \mathcal{W}_t(\delta)} \|(\theta_1 - \theta_2)x\|_2 \leq \frac{1}{\sqrt{6}}.$$

Proof. Let $t \in [\![T]\!]$. For all $x \in \mathcal{X}_t$, using the Cauchy-Schwarz inequality and the Triangle inequality we have

$$\begin{aligned}
&\max_{\theta_1,\theta_2 \in \mathcal{W}_t(\delta)} \|(\theta_1 - \theta_2)x\|_2 \\
&\quad = \max_{\theta_1,\theta_2 \in \mathcal{W}_t(\delta)} \|(I_K \otimes x)(\theta_1 - \theta_2)\|_2 \\
&\quad \leq \|I_K \otimes x\|_{(H_{t-1}^w)^{-1}} \max_{\theta_1,\theta_2 \in \mathcal{W}_t(\delta)} \|\theta_1 - \theta_2\|_{H_{t-1}^w} && \text{(CS)} \\
&\quad \leq \max_{x \in \mathcal{X}_t} \|I_K \otimes x\|_{(H_{t-1}^w)^{-1}} \left(\max_{\theta_1,\theta_2 \in \mathcal{W}_t(\delta)} \|\theta_1 - \theta_t^w\|_{H_{t-1}^w} + \|\theta_t^w - \theta_2\|_{H_{t-1}^w} \right) \\
&\quad \leq \frac{1}{\tau_t} 2\beta_t(\delta) && \text{(Lemma 9)} \\
&\quad = \frac{1}{\sqrt{6}}
\end{aligned}$$

where the last equality is by definition of τ_t. □

G.1.2 Regret Upper-bound

We now focus on the learning part of the algorithm. We start by showing that at each iteration of the algorithm $\sigma_t(\delta)$ defines a confidence set.

Lemma 12. *Let $\delta \in (0,1], \eta = 1$ and $\lambda = 144Kd$. Let us define $\beta_t(\delta) = 4S\sqrt{Kd\log(t/\delta)} + 2S\sqrt{\lambda^w}, \sigma_t(\delta) = 2\sqrt{Kd\log(t/\delta)} + 24S\sqrt{Kd}$ and $\tau_t = 2\sqrt{6}\beta_t(\delta)$. Then we have with probability $1 - 2\delta$, for all $t \geq 1$,*

$$\|\theta_* - \theta_{t+1}\|_{\overline{H}_{t+1}} \leq \sigma_t(\delta)\,.$$

Proof. First, by Lemma 1, we can apply (Lee and Oh, 2025, Theorem 4.2) with $\alpha = 1/\sqrt{6}$ to obtain

$$\|\theta_* - \theta_{t+1}\|_{H_{t+1}} \leq \sigma_t(\delta)\,.$$

Then, we decompose $\mathbb{R}^K$ as $\mathbb{R}^K = 1_K \oplus \mathcal{H}$ where $\mathcal{H}$ is the hyperplane supported by 1_K. Recall that $\theta_*, \theta_{t+1} \in \Pi\mathbb{R}^{K\times d}$, for all $x \in \mathcal{X}$, by definition of Π, $\theta_{t+1}x$ and $\theta_* x$ are in $\mathcal{H}$. Therefore $\sum_{s=1}^t \|(\theta_{t+1} - \theta_*)x_s\|_{1_K 1_K^\top} = 0$. And we conclude that with probability $1 - 2\delta$

$$\|\theta_{t+1} - \theta_*\|_{\overline{H}_{t+1}} = \|\theta_{t+1} - \theta_*\|_{H_{t+1}} \leq \sigma_t(\delta)\,.$$

□

We can now recall and prove our regret upper-bound for Algorithm 3.

Theorem 6. *Let $\delta \in (0,1]$. Set $\lambda^w = 72(1+\sqrt{6}S)Kd, \eta^w = (1+\sqrt{6}S)/2$ and $\lambda = 144Kd$. Then, the regret of Algorithm 3 satisfies with probability at least $1 - 2\delta$,*

$$\mathrm{Reg}_T \leq \widetilde{O}\left(Rd\sqrt{K \sum_{t \notin T^w} \frac{1}{\kappa_{*,t}}} \right)$$

where T^w is the set of time steps when the algorithm explores.

Proof. **Step 1: Tackling the exploration part.** We separate the regret from the exploration and the regret from the learning part:

$$\begin{aligned}
\mathrm{Reg}_T &:= \sum_{t=1}^T \rho^\top \mu(\theta_* x_{*,t}) - \rho^\top \mu(\theta_* x_t) \\
&= \sum_{t \in T^w} \rho^\top \mu(\theta_* x_{*,t}) - \rho^\top \mu(\theta_* x_t) + \sum_{t \notin T^w} \rho^\top \mu(\theta_* x_{*,t}) - \rho^\top \mu(\theta_* x_t) \\
&\leq R|T^w| + \sum_{t \notin T^w} \rho^\top \mu(\theta_* x_{*,t}) - \rho^\top \mu(\theta_* x_t) \\
&\leq 2R\tau_t^2 \kappa K d \log\left(1 + \frac{T}{Kd\lambda^w}\right) + \sum_{t \notin T^w} \rho^\top \mu(\theta_* x_{*,t}) - \rho^\top \mu(\theta_* x_t)
\end{aligned}$$

where the last inequality is due to Lemma 10. Let us now focus on the learning phase of the algorithm.

Step 2: Using optimism. Using the definition of the optimistic reward we can bound the regret twice

$$\begin{aligned}
\mathrm{Reg}_T(\text{Learning}) &:= \sum_{t\notin T^w} \rho^\top(\mu(\theta_* x_{*,t}) - \mu(\theta_* x_t)) \\
&\le \sum_{t\notin T^w} \rho^\top \mu(\theta_t' x_{*,t}) + \varepsilon_{1,t}(x_{*,t}) + \varepsilon_{2,t}(x_{*,t}) - \rho^\top \mu(\theta_* x_t) && \text{(Prop. 2)} \\
&\le \sum_{t\notin T^w} \rho^\top \mu(\theta_t' x_t) + \varepsilon_{1,t}(x_t) + \varepsilon_{2,t}(x_t) - \rho^\top \mu(\theta_* x_t) && \text{(Def. of } x_t\text{)} \\
&\le 2\sum_{t\notin T^w} \varepsilon_{1,t}(x_t) + 2\sum_{t\notin T^w} \varepsilon_{2,t}(x_t) && \text{(Prop. 2)}\,.
\end{aligned}$$

Step 3: Concluding. We may now follow our proof of Theorem 3 to obtain with probability $1-2\delta$

$$\begin{aligned}
\mathrm{Reg}_T(\text{Learning}) &\le \frac{24\sigma_T(\delta)\kappa K d R}{\lambda}\log\left(1+\frac{T}{K\lambda}\right) + 12\kappa\sigma_T(\delta)^2 K d\log\left(1+\frac{T}{Kd\lambda\kappa}\right) \\
&\quad + 4\sqrt{2}e\sigma_T(\delta)\sqrt{d\log\left(1+\frac{T}{\lambda d}\right)}R\sqrt{\sum_{t\notin T^w}\frac{1}{\kappa_{*,t}}} \\
&\quad + 16e^2\sigma_T(\delta)^2 d\log\left(1+\frac{T}{\lambda d}\right)(2\sqrt{K}+4) \\
&\lesssim \mathfrak{C}\kappa K^2 d^2 + \mathfrak{C}Rd\sqrt{K\sum_{t\notin T^w}\frac{1}{\kappa_{*,t}}}\,.
\end{aligned}$$

□

H Auxiliary Results

Lemma 13. *[Boyd and Vandenberghe (2004, Section 4.2.3)] Let $f:\mathbb{R}^d\to\mathbb{R}$ be a convex and differentiable function and $\mathcal{C}\subseteq\mathbb{R}^d$ a convex set. Further, denote:*

$$x_0 := \operatorname*{arg\,min}_{x_0\in\mathcal{C}} f(x)\,.$$

Then for any $y\in\mathcal{C}$:

$$\nabla f(x_0)^\top(y-x_0)\ge 0\,.$$

Lemma 14. *[Modified Freedman's Inequality, Lee et al. (2024, Lemma 3)] Let $X_1,\dots,X_t$ be a martingale difference sequence satisfying $\max_s|X_s|\le D$ a.s., and let $\mathcal{F}_s$ be the σ-field generated by $(X_1,\dots,X_s)$. Then for any $\delta\in(0,1]$ and any $\eta\in[0,1/D]$ the following holds with probability $1-\delta$*

$$\sum_{s=1}^t X_s \le (e-2)\eta\sum_{s=1}^t\mathbb{E}[X_s^2|\mathcal{F}_{s-1}] + \frac{1}{\eta}\log\delta^{-1}\qquad\forall t\ge 1\,.$$

Lemma 15. *[Determinant-Trace Inequality, Abbasi-Yadkori et al. (2011, Lemma 10)] Let $\{x_s\}_{s=1}^\infty$ be a sequence in $\mathbb{R}^d$ such that $\|x_s\|_2\le X$ for all $s\ge 1$, and let $\lambda\ge 0$. For $t\ge 1$ define $V_t := \sum_{s=1}^t x_s x_s^\top + \lambda I_d$. The following inequality holds:*

$$\det(V_t)\le(\lambda+tX^2/d)^d\,.$$