Improved Regret Bounds of (Multinomial) Logistic Bandits via Regret-to-Confidence-Set Conversion

Junghyun Lee[1] **Se-Young Yun**[1] **Kwang-Sung Jun**[2]
[1]Kim Jaechul Graduate School of AI, KAIST, Seoul, Republic of Korea
[2]Department of Computer Science, University of Arizona, Tucson, USA
{jh_lee00, yunseyoung}@kaist.ac.kr kjun@cs.arizona.edu

Abstract

Logistic bandit is a ubiquitous framework of modeling users' choices, e.g., click vs. no click for advertisement recommender system. We observe that the prior works overlook or neglect dependencies in $S \geq \|\boldsymbol{\theta}_\star\|_2$, where $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ is the unknown parameter vector, which is particularly problematic when S is large, e.g., $S \geq d$. In this work, we improve the dependency on S via a novel approach called *regret-to-confidence set conversion (R2CS)*, which allows us to construct a convex confidence set based on only the *existence* of an online learning algorithm with a regret guarantee. Using R2CS, we obtain a strict improvement in the regret bound w.r.t. S in logistic bandits while retaining computational feasibility and the dependence on other factors such as d and T. We apply our new confidence set to the regret analyses of logistic bandits with a new martingale concentration step that circumvents an additional factor of S. We then extend this analysis to multinomial logistic bandits and obtain similar improvements in the regret, showing the efficacy of R2CS. While we applied R2CS to the (multinomial) logistic model, R2CS is a generic approach for developing confidence sets that can be used for various models, which can be of independent interest.

1 INTRODUCTION

The bandit problem (Robbins, 1952; Thompson, 1933) provides a ubiquitous framework to model the exploration-exploitation dilemma, with various variants depending on the application domain. Out of them, (multinomial) logistic bandits (Amani and Thrampoulidis, 2021; Faury et al., 2020; Filippi et al., 2010) has recently received much attention due to its power in modeling binary-valued (discrete-valued) rewards with observed covariates and contexts (respectively). Their applications are abundant in interactive machine learning tasks including news recommendation (Li et al., 2010) where the rewards are ('click', 'no click') or online ad placements where the rewards are one of the multiple outcomes ('click', 'show me later', 'never show again', 'no click').

In logistic bandits, at every time step t, the learner observes a potentially infinite arm-set $\mathcal{X}_t \subset \mathbb{R}^d$ that can vary over time, then plays an action $\boldsymbol{x}_t \in \mathcal{X}_t$. She then receives a reward $r_t \sim \text{Bernoulli}(\mu(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star \rangle))$ for some unknown $\boldsymbol{\theta}_\star \in \mathbb{R}^d$, where $\mu(z) = (1+e^{-z})^{-1}$ is the logistic function. The goal of the learner is to maximize the cumulative reward, and the performance is typically measured by the (pseudo-) regret:

$$\text{Reg}^B(T) := \sum_{t=1}^{T} \left\{ \mu(\langle \boldsymbol{x}_{t,\star}, \boldsymbol{\theta}_\star \rangle) - \mu(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star \rangle) \right\}, \quad (1)$$

where $\boldsymbol{x}_{t,\star} := \arg\max_{\boldsymbol{x} \in \mathcal{X}_t} \mu(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle)$ is the optimal action at time t. The multinomial problem is defined in Section 5.

One popular bandit strategy is the optimistic approach (also known as "optimism in the face of uncertainty"), which selects the next arm with the largest upper confidence bound (UCB). In generalized linear models, the UCB of an arm $\boldsymbol{x} \in \mathbb{R}^d$ is typically constructed by constructing a confidence set $\mathcal{C}_t$ for the unknown parameter $\boldsymbol{\theta}_\star$ and then computing $\max_{\boldsymbol{\theta} \in \mathcal{C}_t} \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$ (Abbasi-Yadkori et al., 2011; Dani et al., 2008; Faury et al., 2022). For this, it is important to ensure that $\mathcal{C}_t$ is a convex set since otherwise the maximization above is computationally intractable in general, and one often needs to resort to using a significantly loosened UCB (e.g., Faury et al. (2020)), which hurts the performance.

Preprint.

	Algorithm	Regret Upper Bound	Tractable?
Logistic Bandits	`SupLogistic` (Jun et al., 2021)	$\sqrt{dT} + d^3\kappa(T)^2$	✓
	`OFULog` (Abeille et al., 2021)	$dS^{\frac{3}{2}}\sqrt{\frac{T}{\kappa_\star(T)}} + \min\left\{d^2S^3\kappa_{\mathcal{X}}(T), R_{\mathcal{X}}(T)\right\}$	✗
	`OFULog-r` (Abeille et al., 2021)	$dS^{\frac{5}{2}}\sqrt{\frac{T}{\kappa_\star(T)}} + \min\left\{d^2S^4\kappa_{\mathcal{X}}(T), R_{\mathcal{X}}(T)\right\}$	✓
	`ada-OFU-ECOLog` (Faury et al., 2022)	$dS\sqrt{\frac{T}{\kappa_\star(T)}} + d^2S^6\kappa(T)$	✓
	`OFULog+` (**ours**)	$dS\sqrt{\frac{T}{\kappa_\star(T)}} + \min\left\{d^2S^2\kappa_{\mathcal{X}}(T), R_{\mathcal{X}}(T)\right\}$	✓
MNL Bandits	`MNL-UCB` (Amani and Thrampoulidis, 2021)	$dK^{\frac{3}{4}}S\sqrt{\kappa(T)T}$	✓
	`Improved MNL-UCB` (Amani and Thrampoulidis, 2021)	$dK^{\frac{5}{4}}S^{\frac{3}{2}}\left(\sqrt{T} + dK^{\frac{5}{4}}S\kappa(T)\right)$	✗
	`MNL-UCB+` (**ours**)	$d\sqrt{KS\kappa(T)T}$	✓
	`Improved MNL-UCB+` (**ours**)	$dKS\left(\sqrt{T} + dK^{\frac{3}{4}}S^{\frac{1}{2}}\kappa(T)\right)$	✗

Table 1: Comparison of regret upper bounds for contextual logistic and multinomial logistic (MNL) bandits, w.r.t. $\kappa_\star(T)$, $\kappa_{\mathcal{X}}(T)$, $\kappa(T)$, d, T, K, and S (see Section 2 and 5 for definitions). For simplicity, we omit logarithmic factors. For logistic bandits, $R_{\mathcal{X}}(T)$ is an arm-set-dependent term that may be much smaller than $\kappa_{\mathcal{X}}(T)$.

One way to construct a convex confidence set is to leverage the loss function, which first appeared in Abeille et al. (2021):

$$\mathcal{C}_t = \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \le S, \bar{\mathcal{L}}_t(\boldsymbol{\theta}) - \bar{\mathcal{L}}_t(\widehat{\boldsymbol{\theta}}_t) \le \beta_t^2\right\}$$

where $\bar{\mathcal{L}}_t$ is the regularized negative log-likelihood, $\widehat{\boldsymbol{\theta}}_t$ is the regularized MLE at time t, and β_t is slowly growing in t. This set $\mathcal{C}_t$ is convex due to the convexity of $\bar{\mathcal{L}}_t$. Such a confidence set is natural as it is based on the log-likelihood ratio and leads to the state-of-the-art regret bound and numerical performance (Abeille et al., 2021; Faury et al., 2022). However, the tightness of the set above, specifically the radius $\beta_t^2 = \mathcal{O}(dS^3\log(t))$, is not clear, which is important given that the tightness directly affects the performance of the algorithm, both in the analysis and the numerical performance.

Contributions. In this paper, we make a number of contributions in (multinomial) logistic bandits that are enabled by a tightened loss-based confidence set.

Firstly, we propose a novel and generic confidence set construction method that we call regret-to-confidence-set conversion (R2CS). Specifically, R2CS constructs a loss-based confidence set via an achievable regret bound in the online learning problem with the matching loss *without* ever having to run the online algorithm. Using R2CS, we provide new confidence sets for logistic loss (Theorem 1) and MNL loss (Theorem 4) that are tighter than prior art Abeille et al. (2021); Amani and Thrampoulidis (2021). Specifically, for the logistic model, our radius is $\beta_t^2 = \mathcal{O}(d\log(t) + S)$ which is a significant improvement upon $\mathcal{O}(dS\log(t))$ from Abeille et al. (2021) when S is large.

R2CS depends on regret bounds of online learning algorithms just like similar approaches of online-to-confidence-set conversion (O2CS) (Abbasi-Yadkori et al., 2012) or online Newton step-based confidence set (Dekel et al., 2010). However, R2CS is fundamentally different from them as R2CS does *not* run the online learning algorithm itself, which allows us to leverage the tight regret guarantees that are currently only available via computationally intractable algorithms (Foster et al., 2018; Mayo et al., 2022); see Appendix A.1 for a detailed comparison.

Secondly, we obtained improved regret bounds of contextual (multinomial) logistic bandits with our new confidence sets as outlined in Table 1. For logistic bandits, we improve by a factor of $\sqrt{S}$ in the leading term and S for lower-order term compared to Abeille et al. (2021), and we improve by a factor of S^4 and possibly κ in the lower-order term compared to Faury et al. (2022). For MNL bandits, we improve by $K^{\frac{1}{4}}\sqrt{S}$ for the leading terms and $\sqrt{KS}$ for the lower-order term. This is discussed in detail in the last paragraphs of Section 4.1 and 5.1.

Outline. Section 2 provides the preliminaries of logistic bandits. Section 3 describes in detail the core ideas of R2CS for logistic bandits, and based on the new confidence set, Section 4 discusses the resulting

improved regret bound of logistic bandits. Lastly, in Section 5, we address how R2CS's applicability extends to multinomial logistic bandits.

Notations. $A \lesssim B$ is when we have $A \le cB$ for some *universal* constant c independent of any quantities we explicitly mention, up to any logarithmic factors. For an integer n, let $[n] := \{1, 2, \cdots, n\}$. $\Delta^K_{>0}$ is the interior of $(K-1)$-dimensional probability simplex. $\mathcal{B}^d(S)$ is the Euclidean d-ball of radius S, and $\mathcal{B}^{K\times d}(S)$ is the ball of radius S in $\mathbb{R}^{K\times d}$ endowed with the Frobenius metric. For a square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ is the minimum and maximum eigenvalue of $\boldsymbol{A}$, respectively. Also, we define the *Loewner ordering* $\succeq$ as $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite. Let Categorical($\boldsymbol{\mu}$) be the $(K+1)$-categorical distribution over $\{0, 1, \ldots, K\}$ with $\boldsymbol{\mu} := [\mu_i]_{i\in[K]} \in [0,1]^K$ where $\mu_i \in \mathbb{R}$ is the mean parameter for category $i \in [K]$ and $\mu_0 = 1 - \sum_i \mu_i$. Denote by $\mathrm{KL}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ the KL-divergence from Categorical($\boldsymbol{\mu}_1$) to Categorical($\boldsymbol{\mu}_2$).

2 PROBLEM SETTING

We first consider stochastic contextual logistic bandit setting that proceeds as described in Section 1. For $s \ge 1$, let $\mathcal{F}_s := \sigma\left(\{\boldsymbol{x}_1, r_1, \cdots, \boldsymbol{x}_s, r_s, \boldsymbol{x}_{s+1}\}\right)$, which constitutes the so-called canonical bandit model; also see Chapter 4.6 of Lattimore and Szepesvári (2020).

We consider the following standard assumptions (Faury et al., 2020):

Assumption 1. $\mathcal{X}_t \subseteq \mathcal{B}^d(1)$ *for all* $t \ge 1$.

Assumption 2. $\boldsymbol{\theta}_\star \in \mathcal{B}^d(S)$ *with known* $S > 0$.

We define the following problem-dependent quantities:

$$\kappa_\star(T) := \frac{1}{\frac{1}{T}\sum_{t=1}^T \dot\mu(\boldsymbol{x}_{t,\star}^\intercal \boldsymbol{\theta}_\star)}, \quad \kappa_{\mathcal{X}}(T) := \max_{t\in[T]} \max_{\boldsymbol{x}\in\mathcal{X}_t} \frac{1}{\dot\mu(\boldsymbol{x}^\intercal\boldsymbol{\theta}_\star)},$$

$$\text{and} \quad \kappa(T) := \max_{t\in[T]} \max_{\boldsymbol{x}\in\mathcal{X}_t} \max_{\boldsymbol{\theta}\in\mathcal{B}^d(S)} \frac{1}{\dot\mu(\boldsymbol{x}^\intercal\boldsymbol{\theta})}.$$

These quantities can scale exponentially in S in the worst-case (Faury et al., 2020).

3 IMPROVED CONFIDENCE SET

Overview and Main Theorem. Our R2CS approach starts by directly constructing a *loss-based* confidence set that contains the true parameter $\boldsymbol{\theta}_\star$ with probability at least $1-\delta$. This confidence set is centered around the maximum likelihood estimator (MLE), $\widehat{\boldsymbol{\theta}}_t$, defined as

$$\widehat{\boldsymbol{\theta}}_t := \operatorname*{arg\,min}_{\|\boldsymbol{\theta}\|_2 \le S} \left[\mathcal{L}_t(\boldsymbol{\theta}) \triangleq \sum_{s=1}^{t-1} \ell_s(\boldsymbol{\theta}) \right], \tag{2}$$

where ℓ_s is the logistic loss at time s, defined as $\ell_s(\boldsymbol{\theta}) := -r_s \log \mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta}\rangle) - (1-r_s)\log(1-\mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta}\rangle))$.

Our loss-based confidence set is then of the form $\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \le \beta_t(\delta)^2$; note that as $\mathcal{L}_t$ is convex, so is the resulting confidence set. Ultimately, we want its radius $\beta_t(\delta)$ to be as small as possible while retaining the high-probability guarantee.

Remark 1. *The existence of* $\widehat{\boldsymbol{\theta}}_t$ *is guaranteed as* $\mathcal{B}^d(S)$ *is compact. Also, as the domain and the objectives are both convex, one can use standard convex optimization algorithms, e.g., Frank-Wolfe method (Frank and Wolfe, 1956) or interior point method (Boyd and Vandenberghe, 2004), to tractably compute* $\widehat{\boldsymbol{\theta}}_t$.

We now present the first main theorem characterizing our new, improved confidence set:

Theorem 1 (Improved Confidence Set for Logistic Loss)**.** *We have*

$$\mathbb{P}\left[\forall t \ge 1,\ \boldsymbol{\theta}_\star \in \mathcal{C}_t(\delta)\right] \ge 1-\delta,$$

where we define

$$\mathcal{C}_t(\delta) := \left\{ \boldsymbol{\theta} \in \mathcal{B}^d(S) : \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \le \beta_t(\delta)^2 \right\},$$

$$\beta_t(\delta) := \sqrt{10d\log\left(\frac{St}{4d} + e\right) + 2((e-2)+S)\log\frac{1}{\delta}}.$$

Roughly speaking, the confidence set of Abeille et al. (2021) resulted in the radius of $\beta_t(\delta) = \mathcal{O}(\sqrt{dS^3\log t})$, while ours result in $\mathcal{O}(\sqrt{(d+S)\log t})$. This separation of d and S leads to an overall improvement in factors of S. Another important observation is that for any $\boldsymbol{\theta}'$, $\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\boldsymbol{\theta}') \le \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \le \beta_t(\delta)^2$, and thus, even when one could find only an approximate estimate of $\mathcal{L}_t(\boldsymbol{\theta})$, the high-probability guarantee of $\boldsymbol{\theta}_\star \in \mathcal{C}_t(\delta)$ still holds! This is in contrast to the prior confidence set (Abeille et al., 2021, Section 3.1), which is geometrically centered around $\widehat{\boldsymbol{\theta}}_t$ and thus a biased estimate shifts the confidence set, breaking the high-probability guarantee.

We now present the proof of Theorem 1, which is the essence of our R2CS approach.

Proof Sketch of Theorem 1. The proof has three main technical novelties, which constitute the crux of our R2CS approach and may be of independent interest to other applications. The first novelty is the two novel decomposition lemmas for the logistic loss (Lemma 1, 2) that express $\beta_t(\delta)^2$ as the sum of the regret of *any* online learning algorithm of our choice, a sum of martingales, and a sum of KL-divergences. The second novelty is when bounding the sum of martingales, we derive and utilize an anytime variant of the Freedman's inequality for martingales (Lemma 3).

The third novelty is when bounding the sum of KL-divergences, we combine the self-concordant result of Abeille et al. (2021) and the information geometric interpretation of the KL-divergence (Lemma 4).

We then use the state-of-the-art online logistic regression regret guarantee of Foster et al. (2018) to obtain the final confidence set (Theorem 1). To use the result of Foster et al. (2018), we use the norm-constrained, unregularized MLE (Eqn. (2)) instead of a regularized MLE used in Abeille et al. (2021). We emphasize here that we do not need to explicitly run the online learning algorithm of Foster et al. (2018), which is quite costly; otherwise, we would have to consider its efficient variant (Jézéquel et al., 2020), which gives an online regret bound scaling with S that gives us no improvement.

3.1 Complete Proof of Theorem 1

To utilize martingale concentrations later, we start by writing

$$r_s = \mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star \rangle) + \xi_s, \tag{3}$$

where ξ_s is a real-valued martingale difference noise.

The following is the first decomposition lemma:

Lemma 1. *For the logistic loss ℓ_s, the following holds for any $\boldsymbol{\theta}$:*

$$\ell_s(\boldsymbol{\theta}_\star) = \ell_s(\boldsymbol{\theta}) + \xi_s \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \boldsymbol{\theta} \rangle - \mathrm{KL}(\mu_s(\boldsymbol{\theta}_\star), \mu_s(\boldsymbol{\theta})).$$

Proof. The proof follows from the first-order Taylor expansion with integral remainder and some careful rearranging of the terms (which is nontrivial); see Appendix C.4.1 for the full proof. □

We can then replace $\boldsymbol{\theta}$ in the above lemma with a sequence of parameters, $\{\tilde{\boldsymbol{\theta}}_s\}$, "outputted" from an online learning algorithm of our choice. This is formalized in the second decomposition lemma:

Lemma 2. *For the logistic loss ℓ_s, the following holds:*

$$\sum_{s=1}^{t} \ell_s(\boldsymbol{\theta}_\star) - \ell_s(\widehat{\boldsymbol{\theta}}_t) \le \mathrm{Reg}^O(t) + \zeta_1(t) - \zeta_2(t), \tag{4}$$

where $\mathrm{Reg}^O(t) := \sum_{s=1}^t \ell_s(\tilde{\boldsymbol{\theta}}_s) - \sum_{s=1}^t \ell_s(\widehat{\boldsymbol{\theta}}_t)$ is the regret incurred by the online learning algorithm of our choice up to time t, and

$$\zeta_1(t) := \sum_{s=1}^{t} \xi_s \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s \rangle,$$

$$\zeta_2(t) := \sum_{s=1}^{t} \mathrm{KL}(\mu_s(\boldsymbol{\theta}_\star), \mu_s(\tilde{\boldsymbol{\theta}}_s)).$$

Proof. The proof follows from Lemma 1 and some rearranging; see Appendix C.4.2 for the full proof. □

For $\mathrm{Reg}^O(t)$, we use the following regret bound for online logistic regression scaling *logarithmically* in S:

Theorem 2 (Theorem 3 of Foster et al. (2018))**.** *There exists an (improper learning) algorithm for online logistic regression with the following regret:*

$$\mathrm{Reg}^O(t) \le 10d \log\left(\frac{St}{4d} + e\right). \tag{5}$$

Remark 2. *The output of Algorithm 1 of Foster et al. (2018) is a sequence of $\hat{\boldsymbol{z}}_s = (\hat{z}_0, \hat{z}_1)$, corresponding to $\boldsymbol{x}_s$ at each time s. For our purpose, we need to designate a vector $\tilde{\boldsymbol{\theta}}_t \in \mathcal{B}^d(S)$ such that $\boldsymbol{\sigma}(\hat{\boldsymbol{z}}_s) = \boldsymbol{\sigma}\left((\langle \boldsymbol{x}_s, \tilde{\boldsymbol{\theta}}_t \rangle)\right)$, where $\boldsymbol{\sigma} : \mathbb{R}^1 \to \Delta^2_{>0}$ is the softmax function defined as $\boldsymbol{\sigma}(z_1) = \left(\frac{1}{1+e^{z_1}}, \frac{e^{z_1}}{1+e^{z_1}}\right)$; see Proposition 1 in Appendix B.2 for a generalization of this for $(K+1)$-classification. Furthermore, the analysis shows that for our purpose, it suffices to use $B = \frac{S}{2}$ in the notation of Foster et al. (2018); see footnote 7 of Appendix B.2 for an explanation.*

Upper Bounding $\zeta_1(t)$: Martingale Concentrations. Recall that $\mathcal{F}_s = \sigma\left(\{\boldsymbol{x}_1, r_1, \cdots, \boldsymbol{x}_s, r_s, \boldsymbol{x}_{s+1}\}\right)$ is the filtration for the canonical bandit model. We start by observing that $\boldsymbol{x}_s$ and $\tilde{\boldsymbol{\theta}}_s$ are $\mathcal{F}_{s-1}$-measurable, and ξ_s is a martingale difference sequence w.r.t. $\mathcal{F}_{s-1}$. We also have that

$$|\xi_s \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s \rangle| \le 2S,$$

$$\mathbb{E}[\xi_s \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s \rangle | \mathcal{F}_{s-1}] = 0,$$

and

$$\mathbb{E}[\xi_s^2 \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s \rangle^2 | \mathcal{F}_{s-1}] = \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star) \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s \rangle^2.$$

We now use a variant[1] of Freedman's inequality for martingales, combined with Ville's inequality to make the concentration hold for any $t \ge 1$.

Lemma 3 (Modification of Theorem 1 of Beygelzimer et al. (2011))**.** *Let $X_1, \cdots, X_t$ be martingale difference sequence satisfying $\max_s |X_s| \le R$ a.s, and let $\mathcal{F}_s$ be the σ-field generated by $(X_1, \cdots, X_s)$. Then for any $\delta \in (0,1)$ and any $\eta \in [0, 1/R]$, the following holds with probability at least $1-\delta$:*

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

Proof. Define $Z_0 = 1$ and $Z_t = Z_{t-1} \cdot \exp(\lambda X_t - (e-2)\lambda^2 \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}]), \forall t \ge 1$. The proof of Theorem 1 of Beygelzimer et al. (2011) shows that $(Z_t)_{t=0}^\infty$ is supermartingale and then applies Markov's inequality. In our proof, we apply Ville's inequality (Lemma 7 in Appendix B.1), to conclude the proof. □

[1] This is a slight variant from the original inequality (Freedman, 1975, Theorem 1.6) in that this uses any fixed estimate of the variance rather than an upper bound.

Thus, for $\eta \in \left[0, \frac{1}{2S}\right]$ to be chosen later, the following holds with probability at least $1-\delta$: for all $t \geq 1$,

$$\zeta_1(t) \leq (e-2)\eta \sum_{s=1}^{t} \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star)\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s\rangle^2 + \frac{1}{\eta}\log\frac{1}{\delta}. \quad (6)$$

Lower Bounding $\zeta_2(t)$: Second-order Expansion of KL Divergence. We first recall the definition of Bregman divergence:

Definition 1. *For a given $m : \mathcal{Z} \to \mathbb{R}$, the* **Bregman divergence** *$D_m(\cdot,\cdot)$ is defined as follows:*

$$D_m(\boldsymbol{z}_1, \boldsymbol{z}_2) = m(\boldsymbol{z}_1) - m(\boldsymbol{z}_2) - \nabla m(\boldsymbol{z}_2)^\intercal(\boldsymbol{z}_1 - \boldsymbol{z}_2)$$

In our case, $\mathcal{Z} = \mathbb{R}$, and thus, from the first-order Taylor's expansion with integral remainder, we have that

$$D_m(z_1, z_2) = \int_{z_2}^{z_1} m''(z)(z_1 - z)dz. \quad (7)$$

The following lemma, which is a standard result in information geometry (Amari, 2016; Brekelmans et al., 2020; Nielsen, 2020), relates Bernoulli KL divergence to a specific Bregman divergence; we provide the proof in Appendix C.4.3 for completeness.

Lemma 4. *Let $m(z) := \log(1+e^z)$ be the log-partition function for Bernoulli distribution and $\mu(z) = \frac{1}{1+e^{-z}}$. Then, we have that $\mathrm{KL}(\mu(z_2), \mu(z_1)) = D_m(z_1, z_2)$.*

Combining all of the above and the fact that $m''(z) = \dot{\mu}(z)$, we have that

$$\begin{aligned}
&\mathrm{KL}(\mu_t(\boldsymbol{x}_s^\intercal\boldsymbol{\theta}_\star)), \mu(\boldsymbol{x}_s^\intercal\tilde{\boldsymbol{\theta}}_s)) \\
&= D_m(\boldsymbol{x}_s^\intercal\tilde{\boldsymbol{\theta}}_s, \boldsymbol{x}_s^\intercal\boldsymbol{\theta}_\star) && \text{(Lemma 4)} \\
&= \int_{\boldsymbol{x}_s^\intercal\boldsymbol{\theta}_\star}^{\boldsymbol{x}_s^\intercal\tilde{\boldsymbol{\theta}}_s} \dot{\mu}(z)(\boldsymbol{x}_s^\intercal\tilde{\boldsymbol{\theta}}_s - z)dz && \text{(Eqn. (7))} \\
&= \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s\rangle^2 \int_0^1 (1-v)\dot{\mu}(\boldsymbol{x}_s^\intercal(\tilde{\boldsymbol{\theta}}_s + (1-v)\boldsymbol{\theta}_\star))dv \\
& && \text{(change-of-variable)} \\
&\overset{(*)}{\geq} \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s\rangle^2 \frac{\dot{\mu}(\boldsymbol{x}_s^\intercal\boldsymbol{\theta}_\star)}{2 + |\boldsymbol{x}_s^\intercal(\boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s)|} \\
&\geq \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s\rangle^2 \frac{\dot{\mu}(\boldsymbol{x}_s^\intercal\boldsymbol{\theta}_\star)}{2+2S}, \\
& && \text{(Assumption 1, 2 and triangle inequality)}
\end{aligned}$$

where $(*)$ is due to the following self-concordant result:

Lemma 5 (Lemma 8 of Abeille et al. (2021))**.** *Let f be any strictly increasing self-concordant function, i.e., $|\ddot{\mu}| \leq \dot{\mu}$, and let $\mathcal{Z} \subset \mathbb{R}$ be bounded. Then, the following holds for any $z_1, z_2 \in \mathcal{Z}$:*

$$\int_0^1 (1-v)\dot{f}(z_1 + v(z_2 - z_1))dv \geq \frac{\dot{f}(z_1)}{2 + |z_1 - z_2|}.$$

Algorithm 1: `OFU-Log+`

1 **for** $t = 1, \ldots, T$ **do**
2 $\quad \widehat{\boldsymbol{\theta}}_t \leftarrow \arg\min_{\|\boldsymbol{\theta}\|_2 \leq S} \mathcal{L}_t(\boldsymbol{\theta})$;
3 $\quad (\boldsymbol{x}_t, \boldsymbol{\theta}_t) \leftarrow \arg\max_{\boldsymbol{x}\in\mathcal{X}_t, \boldsymbol{\theta}\in\mathcal{C}_t(\delta)} \mu(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle)$, with $\mathcal{C}_t(\delta)$ as defined in Theorem 1;
4 $\quad$ Play $\boldsymbol{x}_t$ and observe reward r_t;
5 **end**

All in all, we have that

$$\zeta_2(t) \geq \frac{1}{2+2S}\sum_{s=1}^{t} \dot{\mu}(\boldsymbol{x}_s^\intercal\boldsymbol{\theta}_\star)\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s\rangle^2. \quad (8)$$

Wrapping up the proof. Combining Eqn. (4), (5), (6), (8) with $\eta = \frac{1}{2(e-2)+2S} < \frac{1}{2S}$ and the fact that $-\frac{1}{2+2S} + \frac{e-2}{2(e-2)+2S} < 0$, we are done.

4 IMPROVED REGRET

4.1 `OFULog+` and Improved Regret

Our new loss-based confidence set (Theorem 1) leads to an OFUL-type algorithm (Abbasi-Yadkori et al., 2011), which we refer to as `OFULog+`; its pseudocode is shown in Algorithm 1.

Note that the optimization in line 2 is tractable because $\mathcal{C}_t(\delta)$ is always convex (as $\mathcal{L}_t$ is convex, and the level set of any convex function is convex), and $\mu(\cdot)$ is an increasing function, meaning that line 2 can be equivalently rewritten as

$$(\boldsymbol{x}_t, \boldsymbol{\theta}_t) \in \underset{\boldsymbol{x}\in\mathcal{X}_t, \boldsymbol{\theta}\in\mathcal{C}_t(\delta)}{\arg\max} \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle.$$

The existing confidence-set-based approach to logistic bandit was due to Abeille et al. (2021), in which they first proposed a nonconvex confidence set, from which a loss-based confidence set was derived via convex relaxation. As our R2CS directly constructs the loss-based confidence set, this can be elegantly "plugged-in" to the algorithm and proof of Abeille et al. (2021) with minimal change. This is in contrast to Faury et al. (2022), which requires major algorithmic innovations.

We now present the regret bound of `OFULog+` (See Theorem 6 in Appendix C.2 for the full statement, including the omitted logarithmic factors.):

Theorem 3 (Simplified)**.** *`OFULog+` attains the following regret bound with probability at least $1-\delta$:*

$$\mathrm{Reg}^B(T) \lesssim dS\sqrt{\frac{T}{\kappa_\star(T)}} + \min\left\{d^2S^2\kappa_{\mathcal{X}}(T), R_{\mathcal{X}}(T)\right\},$$

where $R_{\mathcal{X}}(T) := S\sum_{t=1}^{T}\mu(\boldsymbol{x}_{t,\star}^\intercal\boldsymbol{\theta}_\star)\mathbb{1}[\boldsymbol{x}_t \in \mathcal{X}_-(t)]$ and the RHS hides the dependency on $\log\frac{1}{\delta}$. Here, $\mathcal{X}_-(t)$ is the set of detrimental arms at time t as defined in

Abeille et al. (2021) with $\mathcal{X}$ replaced by $\mathcal{X}_t$.

Extending upon Table 1, below, we discuss in detail how our bound compares to existing works[2]:

Comparison to Prior Arts. Contextual logistic bandits, with time-varying arm-set, were first studied by Faury et al. (2020), in which the authors derived the regret bounds of $\widetilde{\mathcal{O}}(\sqrt{\kappa(T)T})$ and $\widetilde{\mathcal{O}}(\sqrt{T}+\kappa(T))$ (corresponding to their two algorithms) based on self-concordant analyses of logistic regression (Bach, 2010). Although not tight, their analyses laid a stepping stone for the subsequent works on logistic bandits. Abeille et al. (2021) provided the first algorithm that attains[3] a regret bound of $\widetilde{\mathcal{O}}\left(dS^{\frac{3}{2}}\sqrt{\frac{T}{\kappa_\star(T)}}+\min\left\{d^2S^3\kappa_{\mathcal{X}}(T),R_{\mathcal{X}}(T)\right\}\right)$ along with near-matching minimax lower bound via an intricate local analysis. Abeille et al. (2021) also proposed a tractable variant of the algorithm, `OFULog-r`, via a convex relaxation, but it incurs an extra dependency on S as shown in Table 1. Faury et al. (2022) provided a jointly efficient and optimal algorithm with $\widetilde{\mathcal{O}}\left(d\sqrt{\frac{T}{\kappa_\star(T)}}+d^2S^6\kappa(T)\right)$ regret that takes $\Omega(1)$ time complexity. Our regret bound's leading term, $dS\sqrt{\frac{T}{\kappa_\star(T)}}$, improves upon Abeille et al. (2021) by a factor of $\sqrt{S}$ and matches that of Faury et al. (2022), and our lower-order term, $\min\{d^2S^2\kappa_{\mathcal{X}}(T),R_{\mathcal{X}}(T)\}$, improves upon Abeille et al. (2021) by a factor of S and improves upon Faury et al. (2022) by a factor of S^4 and possibly $\kappa(T)$.

In Appendix E, we provide numerical results for logistic bandits, showing that our `OFULog+` obtains the state-of-the-art performance in regret over prior arts and results in a tighter confidence set.

On a slightly different approach, Mason et al. (2022) proposed an experimental design-based algorithm. However, the algorithm and its guarantee require the arm-set to be *not* time-varying, making them incomparable to ours. Moreover, the current arm-elimination approach like Mason et al. (2022) is impractical as it needs a long warmup length of order at least $\mathcal{O}(\kappa d^2)$. This is in contrast to the optimism-based approach, which incurs a lower-order algorithm adaptive to the arm-set geometry in that the lower-order term may scale independently of $\kappa_{\mathcal{X}}(T)$, given that the arm-set is sufficiently benign, e.g., unit ball (Abeille et al., 2021, Theorem 3). `SupLogistic` of Jun et al. (2021) assumes that the context vectors follow a distribution and further assumes the minimum eigenvalue condition on the context covariance matrix, which is rather limiting.

Remark 3. *Note that Mason et al. (2022) completely removes the factor of S from the leading term in the regret bound in the fixed arm set setting. We speculate that it is possible to construct an optimism-based algorithm that does not scale with S in the leading term of the regret (up to logarithmic factors), at least for the fixed arm set setting. A related question is whether it is possible to improve further the radius of the confidence set from $\mathcal{O}(\sqrt{(d+S)\log t})$ to $\mathcal{O}(\sqrt{d\log t})$. We leave this as a future work.*

4.2 Proof Sketch of Theorem 3

The proof of Abeille et al. (2021) heavily relies on an upper bound on the Hessian-induced distance between $\boldsymbol{\theta}\in\mathcal{C}_t(\delta)$ and $\boldsymbol{\theta}_\star$, $\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)}$. Here, we define a regularized Hessian $\boldsymbol{H}_t(\boldsymbol{\theta}_\star)$ centered at $\boldsymbol{\theta}_\star$ as

$$\boldsymbol{H}_t(\boldsymbol{\theta}_\star):=\sum_{s=1}^{t-1}\dot{\mu}(\boldsymbol{x}_s^\intercal\boldsymbol{\theta}_\star)\boldsymbol{x}_s\boldsymbol{x}_s^\intercal+\lambda_t\boldsymbol{I}_d,$$

where the regularization coefficient $\lambda_t>0$ is to be chosen later. Note that although our MLE is not regularized (Eqn. 2), The regularization ensures that $\boldsymbol{H}_t$ is positive definite, allowing us to use the elliptical potential lemma argument w.r.t. $\boldsymbol{H}_t^{-1}$-induced norm in the later proof. We remark here that unlike Abeille et al. (2021) where λ_t directly impacts the algorithm design, in our case, λ_t is solely for the proof and does not impact our algorithm in any way.

The key difference between our proof and Abeille et al. (2021) is that we derive a new (high-probability) upper bound on $\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)}$ (Lemma 6). Naïvely using Cauchy-Schwartz inequality and self-concordant controls (as done in the proof of Lemma 1 of Abeille et al. (2021)) gives us an extra factor of S. To circumvent this, we instead use the martingale decomposition of the logistic bandit reward (Eqn. (3)) and Freedman's inequality (Lemma 3) with an ε-net argument, leading to extra factors of S shaved off at the end. With this and our new confidence set guarantee (Theorem 1), we follow through the line of computations of Abeille et al. (2021) to arrive at our new regret bound.

4.3 Complete Proof of Theorem 3

We start with the following crucial lemma bounding the Hessian-induced distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_\star$:

Lemma 6. *For any $\boldsymbol{\theta}\in\mathcal{C}_t(\delta)$, the following holds with probability at least $1-\delta$:*

$$\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|^2_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)}\lesssim\gamma_t(\delta)^2\triangleq S^2\left(d\log\frac{St}{d}+\log\frac{t}{\delta}\right).$$

[2] see Appendix C.2 for the omitted full statements of prior regret bounds.

[3] In the original paper, the authors considered $\lambda_t=d\log t$, which incurred additional factors in S. Here, for a fair comparison, we re-tracked the S-dependencies with the "optimal" choice of $\lambda_t=\frac{d}{S}\log\frac{St}{d\delta}$.

Proof. By Theorem 1, we have that with probability at least $1-\delta$, $\mathcal{L}_t(\boldsymbol{\theta}_\star) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \le \beta_t(\delta)^2$; throughout the proof let us assume that this event is true. Also, let $\boldsymbol{\theta} \in \mathcal{C}_t(\delta)$. Then, by second-order Taylor expansion of $\mathcal{L}_t(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_\star$,

$$\mathcal{L}_t(\boldsymbol{\theta}) = \mathcal{L}_t(\boldsymbol{\theta}_\star) + \nabla\mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal(\boldsymbol{\theta}-\boldsymbol{\theta}_\star) + \|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|^2_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta})-\lambda_t\boldsymbol{I}}$$

where we define the following quantities:

$$\widetilde{\alpha}(\boldsymbol{x},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) := \int_0^1 (1-v)\dot{\mu}\left(\boldsymbol{x}^\intercal(\boldsymbol{\theta}_1 + v(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1))\right) dv$$

$$\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) := \sum_{s=1}^{t-1} \widetilde{\alpha}(\boldsymbol{x}_s,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\boldsymbol{x}_s\boldsymbol{x}_s^\intercal + \lambda_t \boldsymbol{I}_d,$$

where again, $\lambda_t > 0$ is to be determined.

Lemma 5 implies that $\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \succeq \frac{1}{2+2S}\boldsymbol{H}_t(\boldsymbol{\theta}_1)$. Thus, we have that

$$\begin{aligned}
&\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|^2_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)} \\
&\lesssim S\left(\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\boldsymbol{\theta}_\star) + \nabla\mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal(\boldsymbol{\theta}_\star-\boldsymbol{\theta}) + \lambda_t\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_2^2\right) \\
&\lesssim S\left(\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) + \nabla\mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal(\boldsymbol{\theta}_\star-\boldsymbol{\theta})\right) \\
&\qquad\qquad (\mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \le \mathcal{L}_t(\boldsymbol{\theta}_\star),\ \lambda_t = \tfrac{1}{4S^2(2+2S)}) \\
&\lesssim S\beta_t(\delta)^2 + S\nabla\mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal(\boldsymbol{\theta}_\star-\boldsymbol{\theta}), \qquad (\boldsymbol{\theta}\in\mathcal{C}_t(\delta))
\end{aligned}$$

where the last inequality holds with probability at least $1-\delta$. Note that we do not need λ_t to vary over t.

We then observe that $\nabla\mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal(\boldsymbol{\theta}_\star-\boldsymbol{\theta})$ can be written as a sum of martingale difference sequences and that $\boldsymbol{\theta}_\star-\boldsymbol{\theta} \in \mathcal{B}^d(2S)$. The proof then concludes via a time-dependent ε-net argument on $\mathcal{B}^d(2S)$ with Freedman's inequality; see Appendix C.4.4 for the missing details. □

The proof of Theorem 3 finally concludes by tracking the regret analysis of Appendix C of Abeille et al. (2021); see Appendix C.3 for the remaining argument.

5 EXTENSION TO MNL BANDITS

Problem Setting. We now consider a natural extension of logistic bandits, namely, multinomial logistic (MNL) bandits (Amani and Thrampoulidis, 2021). At every round t, the learner observes a potentially infinite arm-set $\mathcal{X}_t$, which can also be time-varying, and plays an action $\boldsymbol{x}_t \in \mathcal{X}$. She then receives a reward of $r_t = \boldsymbol{\rho}^\intercal\boldsymbol{y}_t$, where $\boldsymbol{\rho} \in \mathbb{R}^K$ is a known reward vector, and $\boldsymbol{y}_t = (y_{t,1},\cdots,y_{t,K}) \in \{0,1\}^K$ satisfies $\|\boldsymbol{y}_t\|_1 \le 1$. $y_{s,k} = 1$ when k-th item is chosen at time s, and for simplicity we denote $y_{t,0} := 1 - \|\boldsymbol{y}_t\|_1$. Then, $(y_0,\boldsymbol{y}_t)$ follows the multinomial logit choice model:

$$\mathbb{P}[\boldsymbol{y}_t = \boldsymbol{\delta}_k|\boldsymbol{x}_t] = \begin{cases} \mu_k(\boldsymbol{x}_t,\boldsymbol{\Theta}_\star) & k>0, \\ 1-\sum_{j=1}^K \mu_j(\boldsymbol{x}_t,\boldsymbol{\Theta}_\star) & k=0, \end{cases} \tag{9}$$

where $\boldsymbol{\delta}_k$ is the K-dimensional one-hot encoding for the index k and $\boldsymbol{\delta}_0 := \boldsymbol{0}$. Intuitively, $\boldsymbol{y}_t = \boldsymbol{\delta}_0$ corresponds to the scenario where the user has not chosen any of the K possible choices. Here, we denote

$$\mu_k(\boldsymbol{x}_t,\boldsymbol{\Theta}_\star) := \frac{\exp\left(\langle\boldsymbol{x}_t,(\boldsymbol{\theta}_\star^{(k)})\rangle\right)}{1+\sum_{j=1}^K \exp\left(\langle\boldsymbol{x}_t,(\boldsymbol{\theta}_\star^{(j)})\rangle\right)}. \tag{10}$$

for some unknown $\left\{\boldsymbol{\theta}_\star^{(j)}\right\}_{j=1}^K \subset \mathbb{R}^d$. Here, we use $K\times d$ matrix to denote the unknown parameter, namely, $\boldsymbol{\Theta}_\star := [\boldsymbol{\theta}_\star^{(1)},\cdots,\boldsymbol{\theta}_\star^{(K)}]^\intercal \in \mathbb{R}^{K\times d}$ and $\boldsymbol{\mu}(\boldsymbol{x}_t,\boldsymbol{\Theta}_\star) := [\mu_t(\boldsymbol{\theta}_\star^{(1)}),\cdots,\mu_t(\boldsymbol{\theta}_\star^{(K)})]^\intercal$. This simplifies some parts of the analysis (e.g., avoid using Kronecker products).

The regret of MNL bandits is defined as follows:

$$\mathrm{Reg}^B(T) := \sum_{t=1}^T \boldsymbol{\rho}^\intercal\left(\boldsymbol{\mu}(\boldsymbol{x}_{t,\star},\boldsymbol{\Theta}_\star) - \boldsymbol{\mu}(\boldsymbol{x}_t,\boldsymbol{\Theta})\right), \tag{11}$$

where $\boldsymbol{x}_{t,\star} := \arg\max_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{\rho}^\intercal\boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{\Theta}_\star)$.

We define the following quantity, which will be crucial in our overall analysis:

$$\boldsymbol{A}(\boldsymbol{x},\boldsymbol{\Theta}) := \mathrm{diag}(\boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{\Theta})) - \boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{\Theta})\boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{\Theta})^\intercal. \tag{12}$$

We also have the following assumptions with problem-dependent quantities: denoting $\mathcal{X}_T := \bigcup_{t=1}^T \mathcal{X}_t$,

Assumption 3. *$\mathcal{X}_t \subseteq \mathcal{B}^d(1)$ for all $t \ge 1$.*

Assumption 4. *There exist known constants $S, R > 0$ such that $\boldsymbol{\Theta}_\star \in \mathcal{B}^{K\times d}(S)$ and $\boldsymbol{\rho} \in \mathcal{B}^d(R)$.*

We consider the following problem-dependent quantity (Amani and Thrampoulidis, 2021):

$$\kappa(T) := \max_{\boldsymbol{x}\in\mathcal{X}_T} \max_{\boldsymbol{\Theta}\in\mathcal{B}^{K\times d}(S)} \frac{1}{\lambda_{\min}\left(\boldsymbol{A}(\boldsymbol{x},\boldsymbol{\Theta})\right)}.$$

Improved Confidence Set. We proceed similarly to how we applied R2CS to logistic bandits; to make the correspondence explicit, we overload the notations used in previous sections. We first define the norm-constrained, unregularized MLE for multiclass logistic regression as

$$\widehat{\boldsymbol{\Theta}}_t := \operatorname*{arg\,min}_{\boldsymbol{\Theta}\in\mathcal{B}^{K\times d}(S)} \mathcal{L}_t(\boldsymbol{\Theta}) \triangleq \sum_{s=1}^{t-1} \ell_s(\boldsymbol{\Theta}), \tag{13}$$

where ℓ_s is the multiclass logistic (or *softmax-cross-entropy*) loss at time s, defined as

$$\ell_s(\boldsymbol{\Theta}) := -\sum_{k=0}^K y_{s,k}\log\mu_k(\boldsymbol{x}_s,\boldsymbol{\Theta}),$$

where we denote $\mu_0(\boldsymbol{x}_s,\boldsymbol{\Theta}) := 1 - \sum_{j=1}^K \mu_j(\boldsymbol{x}_0,\boldsymbol{\Theta})$.

Via similar (but with different details) analysis, we obtain the following new confidence set:

Theorem 4 (Improved Confidence Set for Logistic

Algorithm 2: MNL-UCB+

for $t = 1, \dots, T$ **do**
  $\widehat{\boldsymbol{\Theta}}_t \leftarrow \arg\min_{\|\boldsymbol{\Theta}\|_2 \le S} \mathcal{L}_t(\boldsymbol{\Theta})$;
  $\boldsymbol{x}_t \leftarrow \arg\max_{\boldsymbol{x} \in \mathcal{X}_t} \boldsymbol{\rho}^\top \boldsymbol{\mu}(\boldsymbol{x}, \widehat{\boldsymbol{\Theta}}_t) + \epsilon_t(\boldsymbol{x})$, with $\epsilon_t(\boldsymbol{x}) = \sqrt{2\kappa} R L \gamma_t(\delta) \|\boldsymbol{x}\|_{\boldsymbol{V}_t^{-1}}$;
  Play $\boldsymbol{x}_t$ and observe reward r_t;
end

Loss). *We have*

$$\mathbb{P}\left[\forall t \ge 1,\ \boldsymbol{\Theta}_\star \in \mathcal{C}_t(\delta)\right] \ge 1 - \delta,$$

where we define

$$\mathcal{C}_t(\delta) := \left\{\boldsymbol{\Theta} \in \mathcal{B}^{K\times d}(S) : \mathcal{L}_t(\boldsymbol{\Theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\Theta}}_t) \le \beta_t(\delta)^2\right\},$$

$$\beta_t(\delta) := \sqrt{5dK \log \frac{St}{2dK} + 2((e-2) + \sqrt{6KS}) \log \frac{t}{\delta}}.$$

Proof. We extend our previous proof of Theorem 1 to the multinomial scenario. We again use the algorithm of Foster et al. (2018) for our choice of the online learning algorithm. The rest of the proof is quite similar, except we have to use generalized self-concordant control (Sun and Tran-Dinh, 2019; Tran-Dinh et al., 2015); see Appendix D.2 for the full proof. □

5.1 MNL-UCB+ and Improved Regret

Following Amani and Thrampoulidis (2021), our new confidence set leads to our algorithm with an improved bonus term, MNL-UCB+; its pseudocode is shown in Algorithm 2. We can improve further with a tighter bonus term and constrained $\mathcal{C}_t(\delta)$; see Algorithm 3.

For the below theorem statements, we ignore any logarithmic factors, and we also assume that $\kappa(T)$ is very large, as it scales exponentially in S; see Section 3 of Amani and Thrampoulidis (2021).

Theorem 5 (Simplified). *MNL-UCB+ and its improved version attain the following regret bounds, respectively, w.p.* $1 - \delta$*:*

$$\mathrm{Reg}^B(T) \lesssim L_T R_T d\sqrt{KS}\sqrt{\kappa(T)T},$$

$$\mathrm{Reg}^B_{imp}(T) \lesssim R_T dKS\left(\sqrt{T} + dK\sqrt{S}\kappa(T)\right).$$

Proof. See Theorem 12 in Appendix D.3 for the full statement, including the omitted logarithmic factors and other problem-dependent quantities. The full proof is presented in Appendix D.4. □

Again, extending upon Table 1, below, we discuss in detail how our bound compares to existing works:

Comparison to Prior Arts. To the best of our knowledge, the only work with the same setting as ours is Amani and Thrampoulidis (2021); see Appendix A.2 for a review of works on the combinatorial variant of MNL bandits. There, the authors provide two bonus-based algorithms inspired by Faury et al. (2020), each leading[4] to the regret bound of $\widetilde{\mathcal{O}}\left(dK^{\frac{3}{4}}S\sqrt{\kappa(T)T}\right)$ and $\widetilde{\mathcal{O}}\left(dK^{\frac{5}{4}}S^{\frac{3}{2}}\left(\sqrt{T} + dK^{\frac{5}{4}}S\kappa(T)\right)\right)$, respectively. Our bound's leading terms, $d\sqrt{KS}$ and dKS, improve by a factor of $K^{\frac{1}{4}}\sqrt{S}$, and our lower-order term, $d^2K^2S^{\frac{3}{2}}$, improve by a factor of $\sqrt{KS}$. The improvement in K comes from a new martingale concentration argument to bound $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star, \boldsymbol{\theta})}$ (Lemma 14).

Algorithm 3: Improved MNL-UCB+

$\mathcal{M}_1(\boldsymbol{\Theta}) \leftarrow \mathcal{B}^{K\times d}(S)$;
for $t = 1, \dots, T$ **do**
  $\widehat{\boldsymbol{\Theta}}_t \leftarrow \arg\min_{\boldsymbol{\Theta} \in \mathcal{M}_t} \mathcal{L}_t(\boldsymbol{\Theta})$;
  $\boldsymbol{x}_t \leftarrow \arg\max_{\boldsymbol{x} \in \mathcal{X}_t} \boldsymbol{\rho}^\top \boldsymbol{\mu}(\boldsymbol{x}, \widehat{\boldsymbol{\Theta}}_t) + \bar{\epsilon}_t(\boldsymbol{x})$, with $\bar{\epsilon}_t(\boldsymbol{x})$ defined in Eqn. (47) (Appendix D.4.2);
  Play $\boldsymbol{x}_t$ and observe reward r_t;
  $\mathcal{M}_{t+1} \leftarrow \mathcal{M}_t \cap \left\{\boldsymbol{\Theta} : \exists \boldsymbol{\Theta}'_t \in \min(\mathcal{C}_t(\delta)) \text{ s.t. } \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\Theta}) \succeq \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\Theta}'_t)\right\}$[5];
end

Remark 4. *With a closer look at the assumption, a realistic scenario is when* $R = \sqrt{K}R'$ *and* $S = \sqrt{K}S'$*, where* $\|\boldsymbol{\theta}^{(k)}\|_2 \le S'$*. In this scenario, note that contrary to the claim in Amani and Thrampoulidis (2021), their bound results in* $\mathcal{O}(dK^{\frac{7}{4}})$ *(ignoring all the other factors), while ours result in* $\mathcal{O}(dK^{\frac{5}{4}})$*, which is closer to the conjectured optimal dependency of* $\mathcal{O}(dK)$*.*

6 CONCLUSION

In this paper, we propose regret-to-confidence-set conversion (R2CS) that converts an online learning regret guarantee to a new confidence set, without the need to run the online algorithm explicitly. Using a novel combination of self-concordant control and information-geometric interpretation of KL-divergence as well as new martingale concentration arguments, we proved new confidence sets for logistic and MNL bandits, leading to the state-of-the-art regret bounds with improved dependencies on S and K.

One crucial and exciting future direction is to extend our R2CS to various other settings such as generalized linear bandits (Filippi et al., 2010), norm-agnostic scenario (Gales et al., 2022), and even multinomial logistic MDP (Hwang and Oh, 2023).

[4] See Appendix C.2 for the full statement. Also, similarly to logistic bandits, we re-tracked the S-dependency with the "optimal" choice of $\lambda = \frac{K^{\frac{3}{2}}d}{S} \log \frac{ST}{dK\delta}$.

[5] $\min(\mathcal{C}_t(\delta))$ is the set of all minimal elements of the poset $\mathcal{C}_t(\delta)$, endowed with the Loewner ordering w.r.t. $\boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\Theta})$.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean government(MSIT) (No.2022-0-00311, Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects; No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1–9, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL https://proceedings.mlr.press/v22/abbasi-yadkori12.html.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3691–3699. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/abeille21a.html.

Naman Agarwal, Satyen Kale, and Julian Zimmert. Efficient Methods for Online Multiclass Logistic Regression. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 3–33. PMLR, 29 Mar–01 Apr 2022. URL https://proceedings.mlr.press/v167/agarwal22a.html.

Priyank Agrawal, Theja Tulabandhula, and Vashist Avadhanula. A tractable online learning algorithm for the multinomial logit contextual bandit. *European Journal of Operational Research*, 310(2):737–750, 2023. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2023.02.036. URL https://www.sciencedirect.com/science/article/pii/S0377221723001832.

Sanae Amani and Christos Thrampoulidis. UCB-based Algorithms for Multinomial Logistic Regression Bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 2913–2924. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/16f852a6d01b6065c8ff5cc11caae9c6-Paper.pdf.

Shun'ichi Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Tokyo, 2016.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(none): 384 – 414, 2010. doi: 10.1214/09-EJS521. URL https://doi.org/10.1214/09-EJS521.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual Bandit Algorithms with Supervised Learning Guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 19–26, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/beygelzimer11a.html.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Rob Brekelmans, Vaden Masrani, Frank Wood, Greg Ver Steeg, and Aram Galstyan. All in the Exponential Family: Bregman Duality in Thermodynamic Variational Inference. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1111–1122. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/brekelmans20a.html.

Koby Crammer and Claudio Gentile. Multiclass classification with bandit feedback using adaptive regularization. *Machine Learning*, 90(3):347–383, Mar 2013. ISSN 1573-0565. doi: 10.1007/s10994-012-5321-8. URL https://doi.org/10.1007/s10994-012-5321-8.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of the 2008 Conference on Learning Theory*, pages 355–366, 2008.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Robust selective sampling from single and multiple teachers. In *Proceedings of the 2010 Conference on Learning Theory*, 2010.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective Sampling and Active Learning from Single and Multiple Teachers. *Journal of Machine Learning Research*, 13(86):2655–2697, 2012. URL http://jmlr.org/papers/v13/dekel12b.html.

J. L. Doob. Review: Jean Ville, Étude Critique de la Notion de Collectif. *Bulletin of the American Mathematical Society*, 45(11):824, 1939.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/faury20a.html.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly Efficient and Optimal Algorithms for Logistic Bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 546–580. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/faury22a.html.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric Bandits: The Generalized Linear Case. In *Advances in Neural Information Processing Systems*, volume 23, pages 586–594. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic Regression: The Importance of Being Improper. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 167–208. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/foster18a.html.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956. doi: https://doi.org/10.1002/nav.3800030109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800030109.

David A. Freedman. On Tail Probabilities for Martingales. *The Annals of Probability*, 3(1):100 – 118, 1975. doi: 10.1214/aop/1176996452. URL https://doi.org/10.1214/aop/1176996452.

Spencer B. Gales, Sunder Sethuraman, and Kwang-Sung Jun. Norm-Agnostic Linear Bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 73–91. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/gales22a.html.

Claudio Gentile and Francesco Orabona. On Multilabel Classification and Ranking with Bandit Feedback. *Journal of Machine Learning Research*, 15(70):2451–2487, 2014. URL http://jmlr.org/papers/v15/gentile14a.html.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, Dec 2007. ISSN 1573-0565. doi: 10.1007/s10994-007-5016-8. URL https://doi.org/10.1007/s10994-007-5016-8.

Taehyun Hwang and Min-hwan Oh. Model-Based Reinforcement Learning with Multinomial Logistic Function Approximation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):7971–7979, Jun. 2023. doi: 10.1609/aaai.v37i7.25964. URL https://ojs.aaai.org/index.php/AAAI/article/view/25964.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2085–2108. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jezequel20a.html.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Mixability made efficient: Fast online multiclass logistic regression. In *Advances in Neural Information Processing Systems*, volume 34, pages 23692–23702. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c74214a3877c4d8297ac96217d5189b7-Paper.pdf.

Kwang-Sung Jun and Francesco Orabona. Parameter-Free Online Convex Optimization with Sub-Exponential Noise. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1802–1823. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/jun19a.html.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems*, volume 30, pages 98–108. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/28dd2c7955ce926456240b2ff0100bde-Paper.pdf.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5148–5157. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jun21a.html.

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

Peter D. Lax. *Functional Analysis*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley-Interscience, 2002.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

Blake Mason, Kwang-Sung Jun, and Lalit Jain. An Experimental Design Approach for Regret Minimization in Logistic Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7): 7736–7743, Jun. 2022. doi: 10.1609/aaai.v36i7.20741. URL https://ojs.aaai.org/index.php/AAAI/article/view/20741.

Jack J Mayo, Hédi Hadiji, and Tim van Erven. Scale-free unconstrained online learning for curved losses. In *Conference on Learning Theory*, 2022.

Frank Nielsen. An Elementary Introduction to Information Geometry. *Entropy*, 22(10), Sep 2020. ISSN 1099-4300. doi: 10.3390/e22101100. URL http://dx.doi.org/10.3390/e22101100.

Min-hwan Oh and Garud Iyengar. Multinomial Logit Contextual Bandits: Provable Optimality and Practicality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9205–9213, May 2021. doi: 10.1609/aaai.v35i10.17111. URL https://ojs.aaai.org/index.php/AAAI/article/view/17111.

Francesco Orabona and Kwang-Sung Jun. Tight Concentrations and Confidence Sequences from the Regret of Universal Portfolio. *arXiv preprint arXiv:2110.14099*, 2021.

Alexander Rakhlin and Karthik Sridharan. On Equivalence of Martingale Tail Bounds and Deterministic Regret Inequalities. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1704–1722. PMLR, 07–10 Jul 2017. URL https://proceedings.mlr.press/v65/rakhlin17a.html.

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

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for Newton-type methods. *Mathematical Programming*, 178(1):145–213, 2019. doi: 10.1007/s10107-018-1282-4. URL https://doi.org/10.1007/s10107-018-1282-4.

William 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. ISSN 00063444. URL http://www.jstor.org/stable/2332286.

Quoc Tran-Dinh, Yen-Huan Li, and Volkan Cevher. Composite Convex Minimization Involving Self-concordant-Like Cost Functions. In *Modelling, Computation and Optimization in Information Systems and Management Sciences*, pages 155–168, Cham, 2015. Springer International Publishing.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Jean Ville. *Étude critique de la notion de collectif*. Monographies des Probabilités. Paris: Gauthier-Villars, 1939. URL http://eudml.org/doc/192893.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-hua Zhou. Online Stochastic Linear Optimization under One-bit Feedback. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 392–401, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/zhangb16.html.

Contents

A FURTHER RELATED WORK

A.1 Online-to-Confidence Set Conversion

Recently, many results have connected online learning to the concentration of measure, starting from Rakhlin and Sridharan (2017), followed by Jun and Orabona (2019); Orabona and Jun (2021), which is also closely related to the "reduction" framework championed by John Landford[6].

For linear models, there are two main categories of techniques for building confidence sets based on online learning algorithms. The first is to leverage the negative term $-\|\widehat{\boldsymbol{\theta}}_{T+1} - \boldsymbol{\theta}^\star\|^2_{\boldsymbol{V}_T}$ from the regret bound of online Newton step (ONS) (Hazan et al., 2007) where $\boldsymbol{V}_T := \lambda \boldsymbol{I} + \sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t^\intercal$ and $\widehat{\boldsymbol{\theta}}_{T+1}$ is the parameter predicted at the time step $T+1$. This way, one can construct a confidence set centered at $\widehat{\boldsymbol{\theta}}_{T+1}$ with a confidence radius that depends on the rest of the terms in the regret bound (Crammer and Gentile, 2013; Dekel et al., 2010, 2012; Gentile and Orabona, 2014; Zhang et al., 2016). The second one, which is dubbed as *online-to-confidence-set conversion (O2CS)*, is to start from the regret bound $\sum_{t=1}^T \ell_t(\boldsymbol{\theta}_t) - \ell_t(\boldsymbol{\theta}^\star) \leq B_T$ where ℓ_t is a properly defined loss function (e.g., squared loss), $\boldsymbol{\theta}_t$ is the parameter predicted at time t, and B_T is the regret bound of the algorithm. We then lower bound its left-hand side with a standard concentration inequality, which results in a quadratic constraint on $\boldsymbol{\theta}^\star$ (Abbasi-Yadkori et al., 2012; Jun et al., 2017). While this itself defines a confidence set for $\boldsymbol{\theta}^\star$, one can further manipulate the quadratic constraint into a confidence set centered at a new estimator that regresses on the prediction $\hat{y}_t$'s from the online learning algorithm rather than the actual label y_t's. The benefit of O2CS over the ONS-based one is that we are not married to the particular algorithm of ONS but are open to using any online learning algorithm, and thus "progress in constructing better algorithms for online prediction problems directly translates into tighter confidence sets" (Abbasi-Yadkori et al., 2012); see also Chapter 23.3 of Lattimore and Szepesvári (2020).

However, these two techniques have one fundamental difference from our proposed R2CS: they require running the online learning algorithm directly, whereas R2CS relies only on knowing an achievable regret bound without actually running it. This means that our R2CS establishes a third category of techniques for building confidence sets based on online learning algorithms.

A.2 Multinomial Logistic (MNL) Bandits

There are two lines of work in multinomial logistic (MNL) bandits. One line of work, closely related to ours and which we have discussed extensively in the main, considers $K+1$ outcomes modeled by the multinomial logit model. To the best of our knowledge, the only work in this line is Amani and Thrampoulidis (2021), which considers a multinomial extension of Faury et al. (2020). Another line of work considers a combinatorial bandit-type extension for assortment selection problem from choice model theory (Agrawal et al., 2023; Oh and Iyengar, 2021). Here, the considered setting is fundamentally different from ours as in their setting: the learner chooses an assortment (a subset of indices) $\mathcal{Q}_t$, from which the reward follows the multinomial logit distribution over $\mathcal{Q}_t$.

[6] https://hunch.net/~jl/projects/reductions/reductions.html

B MISSING RESULTS

In this section, we provide two missing results from the main text.

B.1 Ville's Inequality

We used a martingale version of Markov's inequality in the proof of Lemma 3, known as Ville's inequality. Here's the full statement:

Lemma 7 (Théorème 1 of pg. 84 of Ville (1939))**.** *Let X_n be a nonnegative supermartingale. Then, for any $\lambda > 0$, $\mathbb{P}\left[\sup_{n\geq 0} X_n \geq \lambda\right] \leq \frac{\mathbb{E}[X_0]}{\lambda}$.*

A fun historical note: this is also commonly known as the *Doob's maximal inequality*, but historically, Jean Ville was the first to report this in literature in his 1939 thesis (Ville, 1939). Interestingly, Joseph L. Doob has an article reviewing (and criticizing) Ville's book (Doob, 1939).

B.2 "Outputs" from Algorithm 1 of Foster et al. (2018)

The following proposition justifies using the improper learning algorithm of Foster et al. (2018) for our purpose (specifically, the existence of $\tilde{\boldsymbol{\theta}}_s$ for logistic bandits and $\widetilde{\boldsymbol{\Theta}}_s$ for multinomial logistic bandits; see Remark 2):

Proposition 1. *Consider a softmax function $\boldsymbol{\sigma} : \mathbb{R}^K \to \Delta^{K+1}_{>0}$ defined as $\boldsymbol{\sigma}(\boldsymbol{z})_k = \frac{e^{z_k}}{1+\sum_{k'\in[K]} e^{z_{k'}}}$ for $k \in [K]$ and $\boldsymbol{\sigma}(\boldsymbol{z})_0 = \frac{1}{1+\sum_{k'\in[K]} e^{z_{k'}}}$. Then, for any $\boldsymbol{x} \in \mathcal{B}^d(1)$ and $\hat{\boldsymbol{z}} \in \mathbb{R}^{K+1}$ outputted from Algorithm 1 of Foster et al. (2018) (see their line 4), there exists $\boldsymbol{\Theta} = [\boldsymbol{\theta}^{(1)}|\cdots|\boldsymbol{\theta}^{(K)}]^\intercal \in \mathcal{B}^{K\times d}(\sqrt{K}S)$ s.t. $\boldsymbol{\sigma}(\hat{\boldsymbol{z}}) = \boldsymbol{\sigma}\left((\langle \boldsymbol{x}, \boldsymbol{\theta}^{(1)}\rangle, \cdots, \langle \boldsymbol{x}, \boldsymbol{\theta}^{(K)}\rangle)\right)$.*

Proof. From line 4 of Algorithm 1 of Foster et al. (2018) with $\mu = 0$, we have that for some distribution P_t whose support is $\mathcal{S} := \left(\mathcal{B}^d(S)\right)^{\otimes K}$ (set of $K \times d$ matrices where the norm of each row is bounded by S),[7]

$$\boldsymbol{\sigma}(\hat{\boldsymbol{z}}) = \mathbb{E}_{\boldsymbol{\Theta}\sim P_t}\left[\boldsymbol{\sigma}(\boldsymbol{\Theta}\boldsymbol{x})\right].$$

Define $F : \mathcal{S} \to \Delta^{K+1}_{>0}$ to be $F(\boldsymbol{\Theta}) = \boldsymbol{\sigma}(\boldsymbol{\Theta}\boldsymbol{x})$, which is continuous. We have the following two lemmas:

Lemma 8. *Let $(\mathcal{X}, P)$ be a probability space with the usual Borel σ-algebra, $Y \subset \mathcal{H}$ be a compact, convex subset of a separable, Hilbert space $\mathcal{H}$, and $F : \mathcal{X} \to Y$ be (Bochner) measurable. Then, for any random variable X on $\mathcal{X}$, we have that $\mathbb{E}[F(X)] \in Y$.*

Lemma 9. $\operatorname{conv}\left(F(\mathcal{S})\right) \subseteq F(\mathcal{B}^{K\times d}(\sqrt{K}S))$*, where* $\operatorname{conv}(\cdot)$ *is the convex hull operator.*

The proof then concludes as the following: by the above two lemmas, we have that $\boldsymbol{\sigma}(\hat{\boldsymbol{z}}) = \mathbb{E}[F(\boldsymbol{\Theta})] \in F(\mathcal{B}^{K\times d}(\sqrt{K}S))$, i.e., there exists $\boldsymbol{\Theta} \in \mathcal{B}^{K\times d}(\sqrt{K}S)$ such that $\boldsymbol{\sigma}(\hat{\boldsymbol{z}}) = F(\boldsymbol{\Theta})$. □

B.2.1 Proof of Lemma 8

(The proof here is inspired by an old StackExchange post. Also, see e.g., Lax (2002) for the necessary background on functional analysis.)

It is clear that $\mathbb{E}[F(X)]$ exists. The proof now proceeds via *reductio ad absurdum*, i.e., suppose that $e \triangleq \mathbb{E}[F(X)] \notin Y$. Then, as $\{e\}$ and Y are disjoint, compact, and convex sets in a separable Hilbert space, by the Hahn-Banach Separation Theorem and Riesz Representation Theorem, there exists a $v \in \mathcal{H}$ such that

$$\langle v, F(x)\rangle < \langle v, e\rangle, \quad \forall x \in \mathcal{X}.$$

[7] The softmax considered in Foster et al. (2018) is actually of the form $\boldsymbol{\sigma}(\boldsymbol{z})_{k'} = \frac{e^{z_{k'}}}{\sum_{l\in\{0\}\cup[K]} e^{z_l}}$ for $k' \in \{0\} \cup [K+1]$. By dividing the denominator and numerator by e^{z_0} and recalling that $z_k = \langle \boldsymbol{x}, \boldsymbol{\theta}^{(k)}\rangle$, by triangle inequality, it can be seen that our parameter space, $\mathcal{S}$, and the parameter space of Foster et al. (2018) with $B = S/2$, $\left(\mathcal{B}^d(S/2)\right)^{\otimes(K+1)}$, are equivalent. In the notation of Foster et al. (2018), we set $B = S/2$.

Then, we have that

$$\int_{\mathcal{X}} \langle v, F(x)\rangle dP(x) = \left\langle v, \int_{\mathcal{X}} F(x) dP(x) \right\rangle = \langle v, e\rangle < \langle v, e\rangle,$$

a contradiction.

B.2.2 Proof of Lemma 9

Let $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathcal{S}$ and $\lambda \in [0,1]$. We will show that $\lambda F(\boldsymbol{\Theta}_1) + (1-\lambda) F(\boldsymbol{\Theta}_2) \in F(\mathcal{B}^{K\times d}(\sqrt{K}S))$.

First, for some given $\boldsymbol{p} = (p_1, \cdots, p_K)^\intercal$, we show that there exists $\boldsymbol{\Theta} = [\boldsymbol{\theta}^{(1)}| \cdots |\boldsymbol{\theta}^{(K)}]^\intercal$ that satisfies the following system of equations: for each $k \in [K]$,

$$\frac{\exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}^{(k)}\rangle\right)}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}^{(k')}\rangle\right)} = p_k.$$

Denoting $\alpha_k := \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}^{(k)}\rangle\right)$, above can be rearranged to the following system of equations:

$$\underbrace{\begin{bmatrix} 1-p_1 & -p_1 & \cdots & -p_1 \\ -p_2 & 1-p_2 & \cdots & -p_2 \\ \vdots & \vdots & \vdots & \vdots \\ -p_K & -p_K & \cdots & 1-p_K \end{bmatrix}}_{\triangleq \boldsymbol{C}_K} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_K \end{bmatrix}.$$

From simple computation, one can easily see that

$$\boldsymbol{C}_K^{-1} = \frac{1}{p_0} \boldsymbol{p}\boldsymbol{1}^\intercal + \boldsymbol{I}_K,$$

where we recall that $p_0 = 1 - \sum_{k=1}^K p_k$. This gives a unique solution

$$\alpha_k^* = \frac{p_k}{p_0} > 0.$$

Then, we arrive at another system of linear equations: $\boldsymbol{x}^\intercal \boldsymbol{\theta}^{(k)} = \log \alpha_k^*$ for each $k \in [K]$. One can easily see that $\boldsymbol{\theta}^{(k)} = \frac{\log \alpha_k^*}{\|\boldsymbol{x}\|_2} \boldsymbol{x}$ satisfies the system.

All in all, we showed that there exists a $\boldsymbol{\Theta}$ such that $\lambda F(\boldsymbol{\Theta}_1) + (1-\lambda) F(\boldsymbol{\Theta}_2) = F(\boldsymbol{\Theta})$ and

$$\|\boldsymbol{\Theta}\|_F^2 = \sum_{k \in [K]} \left(\log \alpha_k^*\right)^2,$$

where in our case,

$$p_k = \lambda \frac{\exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_1^{(k)}\rangle\right)}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_1^{(k')}\rangle\right)} + (1-\lambda) \frac{\exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_2^{(k)}\rangle\right)}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_2^{(k')}\rangle\right)}.$$

Then,

$$\begin{aligned}
\frac{p_k}{p_0} &= \frac{\lambda \frac{\exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_1^{(k)}\rangle\right)}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_1^{(k')}\rangle\right)} + (1-\lambda) \frac{\exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_2^{(k)}\rangle\right)}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_2^{(k')}\rangle\right)}}{\lambda \frac{1}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_1^{(k')}\rangle\right)} + (1-\lambda) \frac{1}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_2^{(k')}\rangle\right)}} \\
&\le \frac{\lambda \frac{e^S}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_1^{(k')}\rangle\right)} + (1-\lambda) \frac{e^S}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_2^{(k')}\rangle\right)}}{\lambda \frac{1}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_1^{(k')}\rangle\right)} + (1-\lambda) \frac{1}{1 + \sum_{k' \in [K]} \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_2^{(k')}\rangle\right)}} \qquad (\boldsymbol{\Theta}_i \in \mathcal{S}, \text{ i.e., } \left\|\boldsymbol{\theta}_i^{(k)}\right\|_2 \le S \text{ for each } k \in [K]) \\
&= e^S,
\end{aligned}$$

and thus,

$$\|\boldsymbol{\Theta}\|_F^2 \le KS^2.$$

C PROOFS - LOGISTIC BANDITS

C.1 Notations

Recall from the main text that $\mathcal{L}_t(\boldsymbol{\theta}) := \sum_{s=1}^t \ell_s(\boldsymbol{\theta})$ is the cumulative *unregularized* logistic loss up to time t. Recall that we also consider the following quantities (Abeille et al., 2021):

$$\widetilde{\alpha}(\boldsymbol{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \int_0^1 (1-v)\dot{\mu}\left(\boldsymbol{x}^\intercal(\boldsymbol{\theta}_1 + v(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1))\right) dv \tag{14}$$

$$\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \sum_{s=1}^{t-1} \widetilde{\alpha}(\boldsymbol{x}_s, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\boldsymbol{x}_s\boldsymbol{x}_s^\intercal + \lambda_t \boldsymbol{I}_d \tag{15}$$

$$\boldsymbol{H}_t(\boldsymbol{\theta}) := \sum_{s=1}^{t-1} \dot{\mu}(\boldsymbol{x}_s^\intercal\boldsymbol{\theta})\boldsymbol{x}_s\boldsymbol{x}_s^\intercal + \lambda_t \boldsymbol{I}_d, \tag{16}$$

where $\lambda_t > 0$ is to be determined, and the following problem-dependent constants:

$$\kappa_\star(T) := \frac{1}{\frac{1}{T}\sum_{t=1}^T \dot{\mu}(\boldsymbol{x}_{t,\star}^\intercal\boldsymbol{\theta}_\star)}, \ \kappa_{\mathcal{X}}(T) := \max_{t\in[T]}\max_{\boldsymbol{x}\in\mathcal{X}_t} \frac{1}{\dot{\mu}(\boldsymbol{x}^\intercal\boldsymbol{\theta}_\star)}, \ \kappa(T) := \max_{t\in[T]}\max_{\boldsymbol{x}\in\mathcal{X}_t}\max_{\boldsymbol{\theta}\in\mathcal{B}^d(S)} \frac{1}{\dot{\mu}(\boldsymbol{x}^\intercal\boldsymbol{\theta})}. \tag{17}$$

C.2 Full Theorem Statements for Regret Bounds

We here provide full theorem statements for our regret analyses and prior arts for logistic bandits.

We start by providing the regret bound of our `OFULog+`:

Theorem 6. *`OFULog+` attains the following regret bound:*

$$\mathrm{Reg}^B(T) \le R_{leading}(T) + R_{log}(T) + R_{detr}(T),$$

where w.p. at least $1-\delta$,

$$R_{leading}(T) \lesssim \left(dS\log\frac{ST}{d} + \sqrt{d}S\log\frac{T}{\delta}\right)\sqrt{\frac{T}{\kappa_\star(T)}},$$

$$R_{log}(T) \lesssim d^2S^2\left(\log\frac{ST}{d}\right)^2 + dS^2\left(\log\frac{T}{\delta}\right)^2,$$

$$R_{detr}(T) \lesssim \min\left\{\kappa_{\mathcal{X}}(T)R_{log}(T), S\sum_{t=1}^T \mu(\boldsymbol{x}_{t,\star}^\intercal\boldsymbol{\theta}_\star)\mathbb{1}[\boldsymbol{x}_t \in \mathcal{X}_-(t)]\right\},$$

where $\mathcal{X}_-(t)$ *is the set of detrimental arms at time* t *as defined in Abeille et al. (2021).*

We now provide the prior state-of-the-art regret bounds that we compare ourselves to:

Theorem 7 (Theorem 1 of Abeille et al. (2021))**.** *`OFULog` with* $\lambda_t = \frac{d}{S}\log\frac{St}{d\delta}$ *attains the following regret bound:*

$$\mathrm{Reg}^B(T) \le R_{leading}(T) + R_{log}(T) + R_{detr}(T),$$

where w.p. at least $1-\delta$,

$$R_{leading}(T) \lesssim dS^{\frac{3}{2}}(\log T)\left(\log\frac{ST}{d} + \log\frac{T}{\delta}\right)\sqrt{\frac{T}{\kappa_\star(T)}},$$

$$R_{log}(T) \lesssim d^2S^3(\log T)^2\left(\log\frac{ST}{d} + \log\frac{T}{\delta}\right)^2,$$

$$R_{detr}(T) \lesssim \min\left\{\kappa_{\mathcal{X}}(T)R_{log}(T), S\sum_{t=1}^T \mu(\boldsymbol{x}_{t,\star}^\intercal\boldsymbol{\theta}_\star)\mathbb{1}[\boldsymbol{x}_t \in \mathcal{X}_-(t)]\right\}.$$

Theorem 8 (Theorem 2 of Abeille et al. (2021))**.** *`OFULog-r` with* $\lambda_t = \frac{d}{S}\log\frac{St}{d\delta}$ *attains the following regret bound:*

$$\mathrm{Reg}^B(T) \le R_{leading}(T) + R_{log}(T) + R_{detr}(T),$$

where w.p. at least $1-\delta$,

$$R_{leading}(T) \lesssim dS^{\frac{5}{2}}(\log T)\left(\log\frac{ST}{d}+\log\frac{T}{\delta}\right)\sqrt{\frac{T}{\kappa_\star(T)}},$$

$$R_{log}(T) \lesssim d^2S^4(\log T)^2\left(\log\frac{ST}{d}+\log\frac{T}{\delta}\right)^2,$$

$$R_{detr}(T) \lesssim \min\left\{\kappa_{\mathcal{X}}(T)R_{log}(T), S\sum_{t=1}^{T}\mu(\boldsymbol{x}_{t,\star}^\intercal\boldsymbol{\theta}_\star)\mathbb{1}[\boldsymbol{x}_t\in\mathcal{X}_-(t)]\right\}.$$

Theorem 9 (Theorem 2 of Faury et al. (2022))**.** *`ada-OFU-ECOLog` attains the following w.p.* $1-\delta$*:*

$$\mathrm{Reg}^B(T) \lesssim dS\log\frac{T}{\delta}\sqrt{\frac{T}{\kappa_\star(T)}}+d^2S^6\kappa\left(\log\frac{T}{\delta}\right)^2.$$

Lastly, although incomparable to our setting, for completeness, we provide the regret bound as provided in Mason et al. (2022) for fixed arm-set setting:

Theorem 10 (Theorem 2 and Corollary 3 of Mason et al. (2022))**.** *`HOMER` with the naive warmup attains the following w.p.* $1-\delta$*:*

$$\mathrm{Reg}^B(T) \lesssim \min\left\{\sqrt{d\frac{T}{\kappa_\star}\log\frac{|\mathcal{X}|}{\delta}}, \frac{d}{\kappa_\star\Delta}\log\frac{|\mathcal{X}|}{\delta}\right\}+d^2\kappa\log\frac{|\mathcal{X}|}{\delta},$$

where $\Delta := \min_{\boldsymbol{x}\in\mathcal{X}\setminus\{\boldsymbol{x}_\star\}}\mu(\boldsymbol{x}_\star^\intercal\boldsymbol{\theta}_\star)-\mu(\boldsymbol{x}^\intercal\boldsymbol{\theta}_\star)$ *is the instance-dependent reward gap. Here, doubly logarithmic terms are omitted.*

C.3 Proof of Theorem 6 – Regret Bound of `OFULog+`

Let us first recall the crucial lemma:

Lemma 6. *For any* $\boldsymbol{\theta}\in\mathcal{C}_t(\delta)$*, the following holds with probability at least* $1-\delta$*:*

$$\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|^2_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)} \lesssim \gamma_t(\delta)^2 \triangleq S^2\left(d\log\frac{St}{d}+\log\frac{t}{\delta}\right).$$

Denote $X=\mathrm{Reg}^B$ as the final bandit regret that we want to bound. We consider the same decomposition $X\le R_1+R_2$ as in Appendix C.1 of Abeille et al. (2021). We first have that

$$R_1 \lesssim \left(dS\log\frac{ST}{d}+\sqrt{d}S\log\frac{T}{\delta}\right)\left(\sqrt{\frac{T}{\kappa_\star}}+\sqrt{X}\right).$$

Next, via two different proof processes, we obtain

$$R_2 \lesssim d^2S^2\kappa_{\mathcal{X}}\left(\log\frac{ST}{d}\right)^2+dS^2\kappa_{\mathcal{X}}\left(\log\frac{T}{\delta}\right)^2$$

and

$$R_2 \lesssim S\sum_{t=1}^{T}\mu(\boldsymbol{x}_{t,\star}^\intercal\boldsymbol{\theta}_\star)\mathbb{1}[\boldsymbol{x}_t\in\mathcal{X}_-(t)]+d^2S^2\left(\log\frac{ST}{d}\right)^2+dS^2\left(\log\frac{T}{\delta}\right)^2.$$

Finally, we obtain the desired result by solving for X via several elementary inequalities (e.g., Proposition 7 of Abeille et al. (2021)).

Remark 5. *For the computations, crude approximations were used for the logarithmic factors, namely,*

$$\sqrt{\log\frac{T}{d}\log\frac{ST}{d}} \le \log\frac{ST}{d},$$

$$\sqrt{\log\frac{T}{\delta}\log\frac{T}{d}} \le \log\frac{T}{\delta}.$$

The second bound is a bit loose in the dependency in δ*, although whether this is important is arguable.*

C.4 Proof of Supporting Lemmas

C.4.1 Proof of Lemma 1

We overload the notation and let $\ell_t(\mu) := -r_t \log \mu - (1 - r_t)\log(1-\mu)$. In this case, we have the following:

$$\ell_t'(\mu) = -\frac{r_t}{\mu} + \frac{1-r_t}{1-\mu}, \quad \ell_t''(\mu) = \frac{r_t}{\mu^2} + \frac{1-r_t}{(1-\mu)^2}.$$

By Taylor's theorem with the integral form of the remainder,

$$\begin{aligned}
\ell_t(\mu_t) - \ell_t(\mu^\star) &= \ell_t'(\mu^\star)(\mu_t - \mu^\star) + \int_{\mu^\star}^{\mu_t} \ell_t''(z)(\mu_t - z)dz \\
&= \frac{\mu^\star - r_t}{\mu^\star(1-\mu^\star)}(\mu_t - \mu^\star) + \int_{\mu^\star}^{\mu_t} \left(\frac{r_t}{z^2} + \frac{1-r_t}{(1-z)^2}\right)(\mu_t - z)dz \\
&= -\xi_t \frac{\mu_t - \mu^\star}{\mu^\star(1-\mu^\star)} + \int_{\mu^\star}^{\mu_t} \left(\frac{r_t}{z^2} + \frac{1-r_t}{(1-z)^2}\right)(\mu_t - z)dz,
\end{aligned}$$

where we recall that $\mu^\star - r_t = -\xi_t$. Let us simplify the integral on the RHS:

$$\begin{aligned}
&\int_{\mu^\star}^{\mu_t} \left(\frac{r_t}{z^2} + \frac{1-r_t}{(1-z)^2}\right)(\mu_t - z)dz \\
&= r_t\left\{\frac{\mu_t}{\mu^\star} - 1 - \log\frac{\mu_t}{\mu^\star}\right\} + (1-r_t)\left\{\frac{1-\mu_t}{1-\mu^\star} - 1 - \log\frac{1-\mu_t}{1-\mu^\star}\right\} \\
&= -1 + \left\{r_t\frac{\mu_t}{\mu^\star} + (1-r_t)\frac{1-\mu_t}{1-\mu^\star}\right\} - \left\{r_t\log\frac{\mu_t}{\mu^\star} + (1-r_t)\log\frac{1-\mu_t}{1-\mu^\star}\right\} \\
&\overset{(*)}{=} -1 + \left\{\mu_t + \xi_t\frac{\mu_t}{\mu^\star} + (1-\mu_t) - \xi_t\frac{1-\mu_t}{1-\mu^\star}\right\} - \left\{\mu^\star\log\frac{\mu_t}{\mu^\star} + (1-\mu^\star)\log\frac{1-\mu_t}{1-\mu^\star} + \xi_t\log\frac{\mu_t}{\mu^\star} - \xi_t\log\frac{1-\mu_t}{1-\mu^\star}\right\} \\
&= \xi_t\frac{\mu_t - \mu^\star}{\mu^\star(1-\mu^\star)} + \mathrm{KL}(\mu^\star, \mu_t) + \xi_t\left(\log\frac{\mu^\star}{1-\mu^\star} - \log\frac{\mu_t}{1-\mu_t}\right),
\end{aligned}$$

where $(*)$ follows from the fact that $r_t = \mu^\star + \xi_t$. Plugging this back into the original expression and recalling the definition of μ_t and $\mu^\star$, we have that

$$\begin{aligned}
\ell_t(\mu_t) - \ell_t(\mu^\star) &= \mathrm{KL}(\mu^\star, \mu_t) + \xi_t\left(\langle \boldsymbol{x}_t, \theta^\star\rangle - \langle \boldsymbol{x}_t, \theta_t\rangle\right) \\
&= \mathrm{KL}(\mu^\star, \mu_t) - \xi_t\langle \boldsymbol{x}_t, \theta^\star - \theta_t\rangle
\end{aligned}$$

C.4.2 Proof of Lemma 2

By Lemma 1, we have the following:

$$\begin{aligned}
0 &= \sum_{s=1}^{t}\left\{\ell_s(\tilde{\boldsymbol{\theta}}_s) - \ell_s(\boldsymbol{\theta}_\star) - \mathrm{KL}(\mu_s(\boldsymbol{\theta}_\star), \mu_s(\tilde{\boldsymbol{\theta}}_s)) + \xi_s\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s\rangle\right\} \\
&= \sum_{s=1}^{t}\left\{\ell_s(\tilde{\boldsymbol{\theta}}_s) - \ell_s(\widehat{\boldsymbol{\theta}}_t) + \ell_s(\widehat{\boldsymbol{\theta}}_t) - \ell_s(\boldsymbol{\theta}^\star) - \mathrm{KL}(\mu_s(\boldsymbol{\theta}^\star), \mu_s(\tilde{\boldsymbol{\theta}}_s)) + \xi_s\langle \boldsymbol{x}_s, \boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}_s\rangle\right\} \\
&= \sum_{s=1}^{t}\left\{\ell_s(\widehat{\boldsymbol{\theta}}_t) - \ell_s(\boldsymbol{\theta}^\star) - \mathrm{KL}(\mu_s(\boldsymbol{\theta}^\star), \mu_s(\tilde{\boldsymbol{\theta}}_s)) + \xi_s\langle \boldsymbol{x}_s, \boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}_s\rangle\right\} + \mathrm{Reg}^O(T).
\end{aligned}$$

Rearranging gives the desired result.

C.4.3 Proof of Lemma 4

This follows from direct computation:

$$\begin{aligned}
D_m(z_1, z_2) &= m(z_1) - m(z_2) - m'(z_2)(z_1 - z_2) \\
&= \log(1 + e^{z_1}) - \log(1 + e^{z_2}) - \frac{e^{z_2}}{1 + e^{z_2}}(z_1 - z_2) \\
&= \log\frac{e^{z_2}}{1 + e^{z_2}} - \log\frac{e^{z_1}}{1 + e^{z_1}} + \left(1 - \frac{e^{z_2}}{1 + e^{z_2}}\right)(z_1 - z_2)
\end{aligned}$$

$$= \log \mu_2 - \log \mu_1 + (1-\mu_2) \log \frac{\mu_1(1-\mu_2)}{\mu_2(1-\mu_1)}$$
$$= \mu_2 \log \frac{\mu_2}{\mu_1} + (1-\mu_2) \log \frac{1-\mu_2}{1-\mu_1} = \mathrm{KL}(\mu_2, \mu_1).$$

C.4.4 Proof of Lemma 6

By Theorem 1, we have that with probability at least $1-\delta$, $\mathcal{L}_t(\boldsymbol{\theta}_\star) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \le \beta_t(\delta)^2$; throughout the proof let us assume that this event is true. Also, let $\boldsymbol{\theta} \in \mathcal{C}_t(\delta)$. Then, by second-order Taylor expansion of $\mathcal{L}_t(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_\star$,

$$\begin{aligned}\mathcal{L}_t(\boldsymbol{\theta}) &= \mathcal{L}_t(\boldsymbol{\theta}_\star) + \nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta} - \boldsymbol{\theta}_\star) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2_{\widetilde{\boldsymbol{G}}(\boldsymbol{\theta}_\star, \boldsymbol{\theta}) - \lambda_t \boldsymbol{I}} \\ &= \mathcal{L}_t(\boldsymbol{\theta}_\star) + \nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta} - \boldsymbol{\theta}_\star) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2_{\widetilde{\boldsymbol{G}}(\boldsymbol{\theta}_\star, \boldsymbol{\theta})} - \lambda_t \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_2^2.\end{aligned}$$

Lemma 5 implies that $\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \succeq \frac{1}{2+2S} \boldsymbol{H}_t(\boldsymbol{\theta}_1)$. Thus, we have that

$$\begin{aligned}\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)} &\le (2+2S) \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star, \boldsymbol{\theta})} \\ &= (2+2S) \left(\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\boldsymbol{\theta}_\star) + \nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}) + \lambda_t \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_2^2 \right) \\ &\le (2+2S) \left(\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) + \nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}) + \lambda_t \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_2^2 \right) \qquad (\mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \le \mathcal{L}_t(\boldsymbol{\theta}_\star)) \\ &\le 1 + (2+2S)\beta_t(\delta)^2 + (2+2S) \nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}), \quad \text{w.p. at least } 1-\delta, \end{aligned} \tag{18}$$

where we choose $\lambda_t = \frac{1}{4S^2(2+2S)}$. Here, there is no need to consider time-varying regularization as unlike Abeille et al. (2021), we do not explicitly use the regularization by λ_t in our algorithm.

Thus, it remains to bound $\nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta})$, which is done via a new concentration-type argument. Let $\mathcal{B}_d(2S)$ be a d-ball of radius $2S$ and $\boldsymbol{v} \in \mathcal{B}_d(2S)$.

First note that

$$\nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal \boldsymbol{v} = \sum_{s=1}^t (\mu(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star) - r_s) \boldsymbol{x}_s^\intercal \boldsymbol{v} = \sum_{s=1}^t \xi_s \boldsymbol{x}_s^\intercal \boldsymbol{v},$$

where here we overload the notation and denote $\xi_s := \mu(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star) - r_s$. Still, ξ_s is a martingale difference sequence w.r.t. $\mathcal{F}_{s-1} = \sigma\left(\{\boldsymbol{x}_1, r_1, \cdots, \boldsymbol{x}_{s-1}, r_{s-1}, \boldsymbol{x}_s\}\right)$, and thus so is $\xi_s \boldsymbol{x}_s^\intercal \boldsymbol{v}$.

As $|\xi_s \boldsymbol{x}_s^\intercal \boldsymbol{v}| \le 2S$ and $\mathbb{E}[(\xi_s \boldsymbol{x}_s^\intercal \boldsymbol{v})^2 | \mathcal{F}_{s-1}] = \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star)(\boldsymbol{x}_s^\intercal \boldsymbol{v})^2$, by Freedman's inequality (Lemma 3), for any $\eta \in \left[0, \frac{1}{2S}\right]$, the following holds:

$$\mathbb{P}\left[\sum_{s=1}^t \xi_s \boldsymbol{x}_s^\intercal \boldsymbol{v} \le (e-2)\eta \sum_{s=1}^t \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star)(\boldsymbol{x}_s^\intercal \boldsymbol{v})^2 + \frac{1}{\eta} \log \frac{t}{\delta} \right] \ge 1 - \delta. \tag{19}$$

Now for $\varepsilon_t \in (0,1)$ to be chosen later satisfying $\varepsilon_t < \varepsilon_{t+1}$, let $\hat{\mathcal{B}}_{\varepsilon_t}$ be an ε_t-cover of $\mathcal{B}_d(2S)$ (endowed with the usual Euclidean topology), i.e.,

$$\forall \boldsymbol{v} \in \mathcal{B}_d(2S),\ \exists \boldsymbol{w}(\boldsymbol{v}) \in \hat{\mathcal{B}}_{\varepsilon_t} : \|\boldsymbol{v} - \boldsymbol{w}(\boldsymbol{v})\|_2 \le \varepsilon_t.$$

By Corollary 4.2.13 of Vershynin (2018), we have that $|\hat{\mathcal{B}}_{\varepsilon_t}| \le \left(\frac{5S}{\varepsilon_t}\right)^d$. With this, we apply union bound for Eqn. (19) to both $t \ge 1$ and $\boldsymbol{v} \in \hat{\mathcal{B}}_{\varepsilon_t}$: with the choice of $\delta_t = \left(\frac{\varepsilon_t}{5S}\right)^d \frac{\delta}{t}$ and applying the union bound, for any $\eta \in [0, 2S]$, the following holds with probability at least $1-\delta$:

$$\sum_{s=1}^t \xi_s \boldsymbol{x}_s^\intercal \boldsymbol{v} \le (e-2)\eta \sum_{s=1}^t \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star)(\boldsymbol{x}_s^\intercal \boldsymbol{v})^2 + \frac{d}{\eta} \log \frac{5S}{\varepsilon_t} + \frac{1}{\eta} \log \frac{t}{\delta}, \quad \forall \boldsymbol{v} \in \hat{\mathcal{B}}(\varepsilon_t),\ \forall t \ge 1.$$

Let $\boldsymbol{v}_t \in \hat{\mathcal{B}}_{\varepsilon_t}$ be s.t. $\|(\boldsymbol{\theta}_\star - \boldsymbol{\theta}) - \boldsymbol{v}_t\|_2 \le \varepsilon_t$. Then,

$$\begin{aligned}&\nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}) \\ &= \sum_{s=1}^t \xi_s \boldsymbol{x}_s^\intercal \boldsymbol{v}_t + \sum_{s=1}^t \xi_s \boldsymbol{x}_s^\intercal \left((\boldsymbol{\theta}_\star - \boldsymbol{\theta}) - \boldsymbol{v}_t \right)\end{aligned}$$

$$\leq (e-2)\eta \sum_{s=1}^{t} \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star)(\boldsymbol{x}_s^\intercal \boldsymbol{v}_t)^2 + \frac{d}{\eta}\log\frac{5S}{\varepsilon_t} + \frac{1}{\eta}\log\frac{t}{\delta} + \varepsilon_t t \qquad \text{(w.p. at least } 1-\delta)$$

$$= (e-2)\eta \sum_{s=1}^{t} \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star)(\boldsymbol{x}_s^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}))^2 + (e-2)\eta \sum_{s=1}^{t} \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star)\left((\boldsymbol{x}_s^\intercal \boldsymbol{v}_t)^2 - (\boldsymbol{x}_s^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}))^2\right) + \frac{d}{\eta}\log\frac{5S}{\varepsilon_t} + \frac{1}{\eta}\log\frac{t}{\delta} + \varepsilon_t t$$

$$\overset{(*)}{\leq} (e-2)\eta \sum_{s=1}^{t} \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star)(\boldsymbol{x}_s^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}))^2 + \frac{(e-2)\eta}{4}(4S\varepsilon_t + \varepsilon_t^2)t + \frac{d}{\eta}\log\frac{5S}{\varepsilon_t} + \frac{1}{\eta}\log\frac{t}{\delta} + \varepsilon_t t$$

$$= (e-2)\eta \|\boldsymbol{\theta}_\star - \boldsymbol{\theta}\|^2_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)} + \frac{d}{\eta}\log\frac{5S}{\varepsilon_t} + \frac{1}{\eta}\log\frac{t}{\delta} + \left(\frac{(e-2)}{4}(4S\eta + \varepsilon_t\eta) + 1\right)\varepsilon_t t.$$

where $(*)$ follows from $\dot{\mu} \leq \frac{1}{4}$ and

$$(\boldsymbol{x}_s^\intercal \boldsymbol{a})^2 - (\boldsymbol{x}_s^\intercal \boldsymbol{b})^2 = 2\boldsymbol{x}_s^\intercal \boldsymbol{b}\boldsymbol{x}_s^\intercal(\boldsymbol{b}-\boldsymbol{a}) + (\boldsymbol{x}_s^\intercal(\boldsymbol{a}-\boldsymbol{b}))^2 \leq 4S\varepsilon_t + \varepsilon_t^2$$

for any $\boldsymbol{a}, \boldsymbol{b} \in \hat{\mathcal{B}}_{\varepsilon_t}$.

Choosing $\eta = \frac{1}{2(e-2)(2+2S)} < \frac{1}{2S}$, $\varepsilon_t = \frac{d}{t}$, and rearranging Eqn. (18) with Theorem 1, we finally have that

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)} \lesssim dS^2 \log\frac{St}{d} + S^2 \log\frac{t}{\delta}.$$

D PROOFS - MULTINOMIAL LOGISTIC BANDITS

D.1 Notations

To follow the regret analyses of Amani and Thrampoulidis (2021) for the remainder of the proof, we vectorize everything and denote $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta}) \in \mathbb{R}^{Kd}$. We first define the following quantities:

$$\boldsymbol{H}_t(\boldsymbol{\theta}) := \sum_{s=1}^{t-1} \boldsymbol{A}(\boldsymbol{x}_s, \boldsymbol{\theta}) \otimes \boldsymbol{x}_s \boldsymbol{x}_s^\intercal \tag{20}$$

$$\boldsymbol{B}(\boldsymbol{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \int_0^1 \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\theta}_1 + v(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)) dv, \tag{21}$$

$$\boldsymbol{G}_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \lambda \boldsymbol{I}_{Kd} + \sum_{s=1}^{t-1} \boldsymbol{B}(\boldsymbol{x}_s, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \otimes \boldsymbol{x}_s \boldsymbol{x}_s^\intercal, \tag{22}$$

$$\widetilde{\boldsymbol{B}}(\boldsymbol{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \int_0^1 (1-v) \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\theta}_1 + v(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)) dv, \tag{23}$$

$$\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \lambda \boldsymbol{I}_{Kd} + \sum_{s=1}^{t-1} \widetilde{\boldsymbol{B}}(\boldsymbol{x}_s, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \otimes \boldsymbol{x}_s \boldsymbol{x}_s^\intercal, \tag{24}$$

$$\boldsymbol{V}_t := 2\kappa\lambda \boldsymbol{I}_d + \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal, \tag{25}$$

where $\lambda > 0$ is to be chosen later.

We also recall all problem-dependent quantities as introduced in Amani and Thrampoulidis (2021), which we extend to time-varying arm-set:

$$\kappa(T) = \max_{\boldsymbol{x} \in \mathcal{X}_T} \max_{\boldsymbol{\Theta} \in \mathcal{B}^{K \times d}(S)} \frac{1}{\lambda_{\min}\left(\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\Theta})\right)}, \tag{26}$$

$$L_T = \max_{\boldsymbol{x} \in \mathcal{X}_T} \max_{\boldsymbol{\Theta} \in \mathcal{B}^{K \times d}(S)} \lambda_{\max}\left(\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\Theta})\right), \tag{27}$$

$$M_T \geq \max_{\boldsymbol{x} \in \mathcal{X}_T} \max_{\boldsymbol{\Theta} \in \mathcal{B}^{K \times d}(S)} \max_{k \in [K]} \left| \lambda_{\max}\left(\nabla^2 \mu_k(\boldsymbol{x}, \boldsymbol{\Theta})\right) \right|, \tag{28}$$

$$M'_T \geq \max_{\boldsymbol{x} \in \mathcal{X}_T} \max_{\boldsymbol{\Theta} \in \mathcal{B}^{K \times d}(S)} \max_{k,k' \in [K]} \left| \lambda_{\max}\left(\nabla[\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\Theta})_{k,k'}]\right) \right|. \tag{29}$$

D.2 Proof of Theorem 4 – MNL Loss-based Confidence Set

We can write

$$\boldsymbol{y}_s = \boldsymbol{\mu}(\boldsymbol{x}_s, \boldsymbol{\theta}_\star) + \boldsymbol{\xi}_s, \tag{30}$$

where $\boldsymbol{\xi}_s$ is some vector-valued martingale noise and $\boldsymbol{y}_s = (y_{s,1}, \cdots, y_{s,K}) \in \{0,1\}^K$.

We first establish an extension of Lemma 1 to the multiclass case:

Lemma 10. *The following holds for any $\boldsymbol{\theta}$:*

$$\ell_s(\boldsymbol{\theta}_\star) = \ell_s(\boldsymbol{\theta}) - \text{KL}(\boldsymbol{\mu}(\boldsymbol{x}_s, \boldsymbol{\theta}_\star), \boldsymbol{\mu}(\boldsymbol{x}_s, \boldsymbol{\theta})) + \boldsymbol{\xi}^\intercal (\boldsymbol{\Theta} - \boldsymbol{\Theta}_\star) \boldsymbol{x}_s, \tag{31}$$

where $\boldsymbol{\Theta}, \boldsymbol{\Theta}_\star \in \mathbb{R}^{K \times d}$ are the parameter matrices whose k-th row corresponds to $(\boldsymbol{\theta}^{(k)})^\intercal, (\boldsymbol{\theta}_\star^{(k)})^\intercal$, respectively.

From hereon, let us universally denote $\boldsymbol{\theta} \in \mathbb{R}^{Kd \times 1}$ to be the vectorized parameter vector, i.e., $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta}^\intercal)$.

Let $\{\widetilde{\boldsymbol{\Theta}}_s\} \subset \mathcal{B}^{K \times d}(\sqrt{K}S)$ be the output from an online learning algorithm of our choice (see Remark 2 and Proposition 1). The following lemma, whose proof is immediate from the above lemma (and is the same as that of Lemma 2), provides the necessary connection:

Lemma 11.

$$\sum_{s=1}^{t} \ell_s(\boldsymbol{\Theta}_\star) - \ell_s(\widehat{\boldsymbol{\Theta}}_t) \leq \text{Reg}^O(t) + \zeta_1(t) - \zeta_2(t), \tag{32}$$

where

$$\zeta_1(t) := \sum_{s=1}^{t} \boldsymbol{\xi}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s, \quad \zeta_2(t) := \sum_{s=1}^{t} \mathrm{KL}(\boldsymbol{\mu}(\boldsymbol{x}_s, \boldsymbol{\Theta}_\star), \boldsymbol{\mu}(\boldsymbol{x}_s, \widetilde{\boldsymbol{\Theta}}_s)).$$

For bounding $\mathrm{Reg}^O(T)$, we again consider the algorithm of Foster et al. (2018), which is also valid for online multiclass logistic regression:

Theorem 11 (Theorem 3 of Foster et al. (2018))**.** *There exists an (improper learning) algorithm for online multiclass logistic regression with the following regret:*

$$\mathrm{Reg}^O(t) \lesssim dK \log \frac{St}{dK}. \tag{33}$$

Remark 6. *Again, if one were to use the classical O2CS approach, then to take computational efficiency into account, one would have to use efficient variants of online multiclass logistic regression algorithm (Agarwal et al., 2022; Jézéquel et al., 2021). These, however, incur an online regret that scales in S, again, which leads to no improvement in the final regret.*

D.2.1 Upper Bounding $\zeta_1(t)$: Martingale Concentrations

Again, let $\mathcal{F}_{s-1}$ be the σ-field generated by $(\boldsymbol{x}_1, \boldsymbol{y}_1, \cdots, \boldsymbol{x}_{s-1}, \boldsymbol{y}_{s-1}, \boldsymbol{x}_s)$. Then, $\boldsymbol{x}_s$ and $\widetilde{\boldsymbol{\Theta}}_s$ are $\mathcal{F}_{s-1}$-measurable, and $\boldsymbol{\xi}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s$ is martingale difference w.r.t. $\mathcal{F}_{s-1}$. We also have that $|\boldsymbol{\xi}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s| \le 2\sqrt{K}S$ and

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{\xi}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s\right)^2 | \mathcal{F}_{s-1}\right] &= \boldsymbol{x}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)^\intercal \mathbb{E}[\boldsymbol{\xi}_s\boldsymbol{\xi}_s^\intercal|\mathcal{F}_{s-1}](\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s \\ &= \boldsymbol{x}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)^\intercal \underbrace{\left(\mathrm{diag}(\{\mu_k((\boldsymbol{\theta}_\star^{(k)})^\intercal\boldsymbol{x}_s)\}_{k=1}^K) - \boldsymbol{\mu}_s\boldsymbol{\mu}_s^\intercal\right)}_{\triangleq \boldsymbol{A}(\boldsymbol{x}_s, \boldsymbol{\Theta}_\star)} (\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s \triangleq \sigma_s^2.\end{aligned}$$

By Freedman's concentration inequality (Lemma 3), the following holds for any $\eta \in \left[0, \frac{1}{2\sqrt{K}S}\right]$:

$$\mathbb{P}\left[\zeta_1(t) = \sum_{s=1}^{t} \boldsymbol{\xi}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s \le (e-2)\eta \sum_{s=1}^{t} \sigma_s^2 + \frac{1}{\eta}\log\frac{t}{\delta}, \quad \forall t \ge 1\right] \ge 1 - \delta. \tag{34}$$

D.2.2 Lower bounding $\zeta_2(t)$: Multivariate second-order expansion of the KL Divergence

The following lemmas are multivariate version of Lemma 4 and 5

Lemma 12. *Let $m(\boldsymbol{z}) := \log\left(1 + \sum_{k=1}^K e^{z_k}\right)$ be the log-exp-sum function (which is known to be the log-partition function for Categorical distribution), and $\boldsymbol{\mu}(\boldsymbol{z}) = (\mu_1, \cdots, \mu_K)$ with $\mu_k := \frac{e^{z_k}}{1+\sum_{k=1}^K e^{z_i}}$. Then we have that $\mathrm{KL}(\boldsymbol{\mu}(\boldsymbol{z}^{(2)}), \boldsymbol{\mu}(\boldsymbol{z}^{(1)})) = D_m(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)})$.*

Definition 2 (Definition 1 of Tran-Dinh et al. (2015); Definition 2 of Sun and Tran-Dinh (2019))**.** *For a given function $f : \mathbb{R}^d \to \mathbb{R}$, define $\varphi_{\boldsymbol{x},\boldsymbol{u}}(t) := f(\boldsymbol{x} + t\boldsymbol{u})$ for $\boldsymbol{x} \in \mathrm{dom}(f)$ and $\boldsymbol{u} \in \mathbb{R}^d$. Then, we say that f is M_f-**generalized self-concordant** if the following is true for any $\boldsymbol{x}, \boldsymbol{u}$:*

$$|\varphi'''_{\boldsymbol{x},\boldsymbol{u}}(t)| \le M_f \varphi''_{\boldsymbol{x},\boldsymbol{u}}(t)\|\boldsymbol{u}\|_2, \quad \forall t \in \mathbb{R}, M_f > 0.$$

Lemma 13. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is M_f-generalized self-concordant, and let $\mathcal{Z} \subset \mathbb{R}^d$ be bounded. Then, the following holds for any $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}$:*

$$\int_0^1 (1-v)\nabla^2 f(\boldsymbol{z}_1 + v(\boldsymbol{z}_2 - \boldsymbol{z}_1))dv \succeq \frac{1}{2 + M_f\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2}\nabla^2 f(\boldsymbol{z}_1). \tag{35}$$

This further implies that $\boldsymbol{H}_t(\boldsymbol{\theta}_\star) \preceq (2 + 2\sqrt{6}S)\widetilde{\boldsymbol{G}}(\boldsymbol{\theta}_\star, \boldsymbol{\theta})$ and $\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star, \boldsymbol{\theta}) \succeq \frac{1}{2\kappa(T)}(\boldsymbol{I}_K \otimes \boldsymbol{V}_t)$.

By Lemma 4 of Tran-Dinh et al. (2015), m is $\sqrt{6}$-generalized self-concordant so that we can apply the above generalized self-concordant result. Via a similar second-order expansion argument, we have that

$$\mathrm{KL}(\boldsymbol{\mu}_\star, \tilde{\boldsymbol{\mu}}) = D_m(\widetilde{\boldsymbol{\Theta}}_s\boldsymbol{x}_s, \boldsymbol{\Theta}_\star\boldsymbol{x}_s)$$

$$
\begin{aligned}
&= \boldsymbol{x}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)^\intercal \left\{ \int_0^1 (1-v)\nabla^2 m(\boldsymbol{\Theta}_\star \boldsymbol{x}_s + v(\widetilde{\boldsymbol{\Theta}}_s \boldsymbol{x}_s - \boldsymbol{\Theta}_\star \boldsymbol{x}_s))(\boldsymbol{z}_1 - \boldsymbol{z}_2) dv \right\} (\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s \\
&\geq \frac{1}{2 + \sqrt{6}\|(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s\|_2} \boldsymbol{x}_s^\intercal(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)^\intercal \nabla^2 m(\boldsymbol{\Theta}_\star \boldsymbol{x}_s)(\widetilde{\boldsymbol{\Theta}}_s - \boldsymbol{\Theta}_\star)\boldsymbol{x}_s \\
&\geq \frac{1}{2 + 2\sqrt{6KS}} \sigma_s^2,
\end{aligned}
$$

and thus,

$$
\zeta_2(t) \geq \frac{1}{2 + 2\sqrt{6KS}} \sum_{s=1}^{t} \sigma_s^2. \tag{36}
$$

Proof of Theorem 4. Combining Eqn. (32)(33)(34)(36) with the choice of $\eta = \frac{1}{2(e-2)+2\sqrt{6KS}} < \frac{1}{2\sqrt{KS}}$ and the fact that $-\frac{1}{2+2\sqrt{6KS}} + \frac{e-2}{2(e-2)+2\sqrt{6KS}} < 0$, we have the desired result. □

D.3 Full Theorem Statements for Regret Bounds

We state the full versions of Theorem 5, including all the logarithmic factors:

Theorem 12. *MNL-UCB+ and its improved version attain the following regret bounds, respectively, w.p. at least* $1-\delta$*:*

$$
\mathrm{Reg}^B(T) \lesssim L_T R_T \sqrt{dKS} \left(\sqrt{d} \log \frac{ST}{dK} + \log \frac{T}{\delta} \right) \sqrt{\max\left(\frac{S^2}{K}, \kappa(T) \right) T}, \tag{37}
$$

$$
\begin{aligned}
\mathrm{Reg}^B_{imp}(T) &\lesssim R_T \sqrt{d} KS \left(\sqrt{d} \log \frac{ST}{dK} + \log \frac{T}{\delta} \right) \sqrt{T} \\
&\quad + R_T d K^{\frac{3}{2}} S \left(M_T' \sqrt{KS} + M_T \right) \left(d \left(\log \frac{ST}{dK} \right)^2 + \left(\log \frac{T}{\delta} \right)^2 \right) \max\left(\frac{S^2}{K}, \kappa(T) \right).
\end{aligned} \tag{38}
$$

We now provide the (only) previous state-of-the-art regret bound that we compare ourselves to:

Theorem 13 (Theorem 2, 3 of Amani and Thrampoulidis (2021))**.** *MNL-UCB and its improved version with* $\lambda = \frac{dK^{\frac{3}{2}}}{S} \log \frac{ST}{dK\delta}$ *attain the following regret bounds, respectively, w.p.* $1-\delta$*:*

$$
\mathrm{Reg}^B(T) \lesssim L_T R_T d K^{\frac{3}{4}} S \left(\log \frac{ST}{dK} + \log \frac{T}{\delta} \right) \sqrt{\max\left(\frac{S}{dK^{\frac{3}{4}} \log \frac{ST}{dK\delta}}, \kappa(T) \right) T}, \tag{39}
$$

$$
\begin{aligned}
\mathrm{Reg}^B_{imp}(T) &\lesssim R_T d K^{\frac{5}{4}} S^{\frac{3}{2}} \left(\log \frac{ST}{dK} + \log \frac{T}{\delta} \right) \sqrt{T} \\
&\quad + R_T d^2 K^2 S^2 (M_T' \sqrt{KS} + M_T) \left(\left(\log \frac{ST}{dK} \right)^2 + \left(\log \frac{T}{\delta} \right)^2 \right) \max\left(\frac{S}{dK^{\frac{3}{2}} \log \frac{ST}{dK\delta}}, \kappa(T) \right).
\end{aligned} \tag{40}
$$

D.4 Proof of Theorem 12 – Regret Bound of (Improved) MNL-UCB+

Again, we start with the following crucial lemma, whose proof is provided in Appendix D.5.4:

Lemma 14. *For any* $\boldsymbol{\theta} \in \mathcal{C}_t(\delta)$*, the following holds with probability at least* $1-\delta$*:*

$$
\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star, \boldsymbol{\theta})} \lesssim \gamma_t(\delta)^2 \triangleq dKS \log \frac{St}{dK} + \sqrt{KS} \log \frac{t}{\delta} + dKL_T, \tag{41}
$$

For simplicity, we assume that the last term, dKL_T, is negligible.

Now, assume that we have some bonus term $\epsilon_t(\boldsymbol{x})$ s.t. the following holds w.h.p. for each $\boldsymbol{x} \in \mathcal{X}_t$ and $t \in [T]$:

$$
\Delta(\boldsymbol{x}, \boldsymbol{\Theta}_t) := \left| \boldsymbol{\rho}^\intercal \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\theta}_\star) - \boldsymbol{\rho}^\intercal \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\theta}_t) \right| \leq \epsilon_t(\boldsymbol{x}), \tag{42}
$$

and assume that the learner follows the following UCB algorithm:

$$
\boldsymbol{x}_t = \operatorname*{arg\,max}_{\boldsymbol{x} \in \mathcal{X}_t} \boldsymbol{\rho}^\intercal \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\theta}_t) + \epsilon_t(\boldsymbol{x}). \tag{43}
$$

Then, we have that

$$\begin{aligned}\mathrm{Reg}^B(T) &= \sum_{t=1}^{T}\left\{\boldsymbol{\rho}^\intercal\boldsymbol{\mu}(\boldsymbol{x}_{t,\star},\boldsymbol{\theta}_\star) - \boldsymbol{\rho}^\intercal\boldsymbol{\mu}(\boldsymbol{x}_t,\boldsymbol{\theta}_\star)\right\}\\ &\le \sum_{t=1}^{T}\left\{\boldsymbol{\rho}^\intercal\boldsymbol{\mu}(\boldsymbol{x}_{t,\star},\boldsymbol{\theta}_t) + \epsilon_t(\boldsymbol{x}_{t,\star}) - \boldsymbol{\rho}^\intercal\boldsymbol{\mu}(\boldsymbol{x}_t,\boldsymbol{\theta}_\star)\right\}\\ &\le \sum_{t=1}^{T}\left\{\boldsymbol{\rho}^\intercal\boldsymbol{\mu}(\boldsymbol{x}_t,\boldsymbol{\theta}_t) + \epsilon_t(\boldsymbol{x}_t) - \boldsymbol{\rho}^\intercal\boldsymbol{\mu}(\boldsymbol{x}_t,\boldsymbol{\theta}_\star)\right\}\\ &\le 2\sum_{t=1}^{T}\epsilon_t(\boldsymbol{x}_t).\end{aligned}$$

We also recall a simple technical lemma:

Lemma 15 (Lemma 10 of Amani and Thrampoulidis (2021))**.**

$$\boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{\theta}_1) - \boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{\theta}_1) = \left[\boldsymbol{B}(\boldsymbol{x},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\otimes\boldsymbol{x}^\intercal\right](\boldsymbol{\theta}_1-\boldsymbol{\theta}_2). \tag{44}$$

D.4.1 $\sqrt{\kappa T}$-type regret – Algorithm 2

Here, we follow the proof provided in Appendix B of Amani and Thrampoulidis (2021). We start with the following lemma,

Lemma 16 (Improved Lemma 1 of Amani and Thrampoulidis (2021))**.** *For $\boldsymbol{\theta}\in\mathcal{C}_t(\delta)$ and $\boldsymbol{x}\in\mathcal{X}_t$, the following holds with probability at least $1-\delta$:*

$$\Delta(\boldsymbol{x},\boldsymbol{\theta}) \le \sqrt{2\kappa}R_T L_T \gamma_t(\delta)\|\boldsymbol{x}\|_{\boldsymbol{V}_t^{-1}}. \tag{45}$$

Proof. We have that

$$\begin{aligned}\Delta(\boldsymbol{x},\boldsymbol{\theta}) &\le R_T\left\|\left[\boldsymbol{B}(\boldsymbol{x},\boldsymbol{\theta}_\star,\boldsymbol{\theta})\otimes\boldsymbol{x}^\intercal\right](\boldsymbol{\theta}_\star-\boldsymbol{\theta})\right\|_2 && \text{(Assumption 4, CS, Lemma 15)}\\ &\le R_T\left\|\left[\boldsymbol{B}(\boldsymbol{x},\boldsymbol{\theta}_\star,\boldsymbol{\theta})\otimes\boldsymbol{x}^\intercal\right]\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta})^{-1/2}\right\|_2\|\boldsymbol{\theta}_\star-\boldsymbol{\theta}\|_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta})} && \text{(CS)}\\ &\overset{(*)}{\le} R_T L_T\sqrt{\lambda_{\max}\left(\left[\boldsymbol{I}_K\otimes\boldsymbol{x}^\intercal\right]\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta})^{-1}\left[\boldsymbol{I}_K\otimes\boldsymbol{x}\right]\right)}\|\boldsymbol{\theta}_\star-\boldsymbol{\theta}\|_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta})}\\ &\le R_T L_T\sqrt{2\kappa\lambda_{\max}\left(\left[\boldsymbol{I}_K\otimes\boldsymbol{x}^\intercal\right]\left[\boldsymbol{I}_K\otimes\boldsymbol{V}_t^{-1}\right]\left[\boldsymbol{I}_K\otimes\boldsymbol{x}\right]\right)}\|\boldsymbol{\theta}_\star-\boldsymbol{\theta}\|_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta})} && \text{(Lemma 13)}\\ &= \sqrt{2\kappa(T)}R_T L_T\gamma_t(\delta)\|\boldsymbol{x}\|_{\boldsymbol{V}_t^{-1}}, && (\boldsymbol{\theta}\in\mathcal{C}_t(\delta)\text{, Theorem 4})\end{aligned}$$

where CS refers to Cauchy-Schwartz inequality and $(*)$ is when the hidden computations are precisely the same as done in the chain of inequalities in Appendix B.2 of Amani and Thrampoulidis (2021). □

Thus by elliptical potential lemma (Abbasi-Yadkori et al., 2011, Lemma 11) and recalling that we've chosen $\lambda = \frac{K}{4S^2}$, we have the following regret bound:

$$\begin{aligned}\mathrm{Reg}^B(T) &\lesssim L_T R_T\sqrt{dKS\log\frac{ST}{dK} + S\sqrt{K}\log\frac{T}{\delta}}\sqrt{\max\left(1,\frac{S^2}{\kappa(T)K}\right)\kappa(T)dT\log\left(1+\frac{TS}{\kappa dK}\right)}\\ &\lesssim L_T R_T\sqrt{dKS}\left(\sqrt{d}\log\frac{ST}{dK}+\log\frac{T}{\delta}\right)\sqrt{\max\left(\frac{S^2}{K},\kappa(T)\right)T}.\end{aligned}$$

D.4.2 $\sqrt{T}+\kappa$-type regret – Algorithm 3

Here, we follow the proof provided in Appendix D of Amani and Thrampoulidis (2021). With similar reasoning as previous, we first have the following:

Lemma 17 (Improved Lemma 17 of Amani and Thrampoulidis (2021)). *For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{W}_t(\delta)$[8] and any $t \in [T]$, with probability at least $1-\delta$ we have that*

$$\boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\theta}_1) - \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\theta}_2) \leq \left[\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\Theta}_2) \otimes \boldsymbol{x}^\intercal\right] (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) + 2\kappa(T) M_T \gamma_t(\delta)^2 \|\boldsymbol{x}\|^2_{\boldsymbol{V}_t^{-1}} \mathbf{1}, \tag{46}$$

where $\leq$ holds elementwise.

Proof. In their chain of inequalities for their proof of Lemma 17 in their Appendix D (Amani and Thrampoulidis, 2021), we alternatively proceed as follows:

$$\begin{aligned}
M_T \left\|\left[\boldsymbol{I}_K \otimes \boldsymbol{x}^\intercal\right](\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\right\|_2^2 &\leq M_T \left\|\left[\boldsymbol{I}_K \otimes \boldsymbol{x}^\intercal\right] \widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{-1/2}\right\|_2^2 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2_{\widetilde{\boldsymbol{G}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} && \text{(CS)} \\
&\leq M_T \left\|\left[\boldsymbol{I}_K \otimes \boldsymbol{x}^\intercal\right] \widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{-1/2}\right\|_2^2 \gamma_t(\delta)^2 && \text{(Lemma 14)} \\
&\overset{(*)}{\leq} 2\kappa(T) M_T \gamma_t(\delta)^2 \|\boldsymbol{x}\|^2_{\boldsymbol{V}_t^{-1}}
\end{aligned}$$

where CS refers to Cauchy-Schwartz inequality w.r.t. $\widetilde{\boldsymbol{G}}_t$ instead of $\boldsymbol{G}_t$, and $(*)$ is when the hidden computations are precisely the same as done in the chain of inequalities in Appendix D of Amani and Thrampoulidis (2021). The rest of the proof is then the same. □

Lemma 18 (Improved Lemma 18 of Amani and Thrampoulidis (2021)).

$$\Delta(\boldsymbol{x}, \boldsymbol{\theta}_t) \leq \bar{\epsilon}_t(\boldsymbol{x}, \boldsymbol{\theta}_t) := R_T \sqrt{2 + 2\sqrt{6} S} \gamma_t(\delta) \left\|\left[\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\Theta}_t) \otimes \boldsymbol{x}^\intercal\right] \boldsymbol{H}_t(\boldsymbol{\theta}_t)^{-1/2}\right\|_2 + 2\kappa(T) M_T \left(\sum_{k=1}^K \rho_k\right) \gamma_t(\delta)^2 \|\boldsymbol{x}\|^2_{\boldsymbol{V}_t^{-1}}. \tag{47}$$

Proof. In their chain of inequalities for their proof of Lemma 18 in their Appendix D (Amani and Thrampoulidis, 2021), we alternatively proceed as follows:

$$\begin{aligned}
\Delta(\boldsymbol{x}, \boldsymbol{\theta}_t) &\leq R_T \left\|\left[\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\theta}_t) \otimes \boldsymbol{x}^\intercal\right] (\boldsymbol{\theta}_\star - \boldsymbol{\theta}_t)\right\|_2 + 2\kappa(T) M_T \left(\sum_{k=1}^K \rho_k\right) \gamma_t(\delta)^2 \|\boldsymbol{x}\|^2_{\boldsymbol{V}_t^{-1}} \\
&\leq R_T \left\|\left[\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\theta}_t) \otimes \boldsymbol{x}^\intercal\right] \widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star, \boldsymbol{\theta}_t)^{-1/2}\right\|_2 \|\boldsymbol{\theta}_\star - \boldsymbol{\theta}_t\|_{\widetilde{\boldsymbol{G}}(\boldsymbol{\theta}_\star, \boldsymbol{\theta}_t)} + 2\kappa(T) M_T \left(\sum_{k=1}^K \rho_k\right) \gamma_t(\delta)^2 \|\boldsymbol{x}\|^2_{\boldsymbol{V}_t^{-1}} && \text{(CS)} \\
&\leq R_T \gamma_t(\delta) \left\|\left[\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\theta}) \otimes \boldsymbol{x}^\intercal\right] \widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star, \boldsymbol{\theta}_t)^{-1/2}\right\|_2 + 2\kappa(T) M_T \left(\sum_{k=1}^K \rho_k\right) \gamma_t(\delta)^2 \|\boldsymbol{x}\|^2_{\boldsymbol{V}_t^{-1}} && \text{(Lemma 14)} \\
&\leq R_T \sqrt{2 + 2\sqrt{6} S} \gamma_t(\delta) \left\|\left[\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\theta}_t) \otimes \boldsymbol{x}^\intercal\right] \boldsymbol{H}_t(\boldsymbol{\theta}_t)^{-1/2}\right\|_2 + 2\kappa(T) M_T \left(\sum_{k=1}^K \rho_k\right) \gamma_t(\delta)^2 \|\boldsymbol{x}\|^2_{\boldsymbol{V}_t^{-1}}, \\
& && \text{(Lemma 13)}
\end{aligned}$$

where CS refers to Cauchy-Schwartz inequality w.r.t. $\widetilde{\boldsymbol{G}}_t$ instead of $\boldsymbol{G}_t$. □

After some tedious computations, again following through proof of their Theorem 3 in Appendix D of Amani and Thrampoulidis (2021), while applying Cauchy-Schwartz inequality w.r.t. $\widetilde{\boldsymbol{G}}_t$ instead of $\boldsymbol{G}_t$, we have the following regret bound:

$$\begin{aligned}
\mathrm{Reg}^B(T) \lesssim\ & R_T \sqrt{d} K S \left(\sqrt{d} \log \frac{ST}{dK} + \log \frac{T}{\delta}\right) \sqrt{T} \\
& + R_T d K^{\frac{3}{2}} S \left(M'_T \sqrt{KS} + M_T\right) \left(d \left(\log \frac{ST}{dK}\right)^2 + \left(\log \frac{T}{\delta}\right)^2\right) \max\left(\frac{S^2}{K}, \kappa(T)\right).
\end{aligned} \tag{48}$$

[8] $\mathcal{W}_t$ is the new feasible set of estimators, defined in Eqn. (72) of Amani and Thrampoulidis (2021).

D.5 Proof of Supporting Lemmas

D.5.1 Proof of Lemma 10

We overload the notation and let $\ell(\boldsymbol{\mu}) = -y_0 \log\left(1 - \sum_{k=1}^K \mu_k\right) - \sum_{k=1}^K y_k \log \mu_k$, where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)$. For simplicity denote $\mu_0(\boldsymbol{\mu}) = \mu_0 = 1 - \sum_{k=1}^K \mu_k$ and $\mu_0^\star = \mu_0(\boldsymbol{\mu}^\star)$. Then we first have that for $k \neq k' \in [K]$,

$$\partial_k \ell(\boldsymbol{\mu}) = \frac{y_0}{\mu_0} - \frac{y_k}{\mu_k}, \quad \partial_{kk} \ell(\boldsymbol{\mu}) = \frac{y_0}{\mu_0^2} + \frac{y_k}{\mu_k^2}, \quad \partial_{kk'} \ell(\boldsymbol{\mu}) = \frac{y_0}{\mu_0^2}.$$

Let α be multi-index. By multivariate Taylor's theorem with the integral form of remainder,

$$\begin{aligned}
\ell(\boldsymbol{\mu}) - \ell(\boldsymbol{\mu}^\star) &= \nabla \ell(\boldsymbol{\mu}^\star)^\intercal (\boldsymbol{\mu} - \boldsymbol{\mu}^\star) + 2 \sum_{|\alpha|=2} \frac{(\boldsymbol{\mu} - \boldsymbol{\mu}^\star)^\alpha}{\alpha!} \int_0^1 (1-t) \partial^\alpha \ell(\boldsymbol{\mu}^\star + t(\boldsymbol{\mu} - \boldsymbol{\mu}^\star)) dt \\
&= \nabla \ell(\boldsymbol{\mu}^\star)^\intercal (\boldsymbol{\mu} - \boldsymbol{\mu}^\star) + \sum_{k=1}^K (\mu_k - \mu_k^\star)^2 \int_0^1 (1-t) \left\{ \frac{y_0}{(\mu_0^\star + t(\mu_0 - \mu_0^\star))^2} + \frac{y_k}{(\mu_k^\star + t(\mu_k - \mu_k^\star))^2} \right\} dt \\
&\quad + 2 \sum_{1 \le k < k' \le K} (\mu_k - \mu_k^\star)(\mu_{k'} - \mu_{k'}^\star) \int_0^1 (1-t) \frac{y_0}{(\mu_0^\star + t(\mu_0 - \mu_0^\star))^2} dt \\
&= \nabla \ell(\boldsymbol{\mu}^\star)^\intercal (\boldsymbol{\mu} - \boldsymbol{\mu}^\star) + \sum_{k=1}^K (\mu_k - \mu_k^\star)^2 \int_0^1 (1-t) \frac{y_k}{(\mu_k^\star + t(\mu_k - \mu_k^\star))^2} dt \\
&\quad + \left(\sum_{k=1}^K (\mu_k - \mu_k^\star) \right)^2 \int_0^1 (1-t) \frac{y_0}{(\mu_0^\star + t(\mu_0 - \mu_0^\star))^2} dt \\
&= \underbrace{\nabla \ell(\boldsymbol{\mu}^\star)^\intercal (\boldsymbol{\mu} - \boldsymbol{\mu}^\star)}_{(a)} + \underbrace{\sum_{k=0}^K (\mu_k - \mu_k^\star)^2 \int_0^1 (1-t) \frac{y_k}{(\mu_k^\star + t(\mu_k - \mu_k^\star))^2} dt}_{(b)}.
\end{aligned}$$

(a)

$$\begin{aligned}
\nabla \ell(\boldsymbol{\mu}^\star)^\intercal (\boldsymbol{\mu} - \boldsymbol{\mu}^\star) &= \sum_{k=1}^K \left(\frac{y_0}{\mu_0^\star} - \frac{y_k}{\mu_k^\star} \right) (\mu_k - \mu_k^\star) \\
&= \sum_{k=1}^K \left(\frac{y_0}{\mu_0^\star} (\mu_k - \mu_k^\star) - \frac{y_k}{\mu_k^\star} \mu_k + y_k \right).
\end{aligned}$$

(b)

$$\begin{aligned}
\sum_{k=0}^K (\mu_k - \mu_k^\star)^2 \int_0^1 (1-t) \frac{y_k}{(\mu_k^\star + t(\mu_k - \mu_k^\star))^2} dt &= \sum_{k=0}^K (\mu_k - \mu_k^\star)^2 \int_{\mu_k^\star}^{\mu_k} \left(1 - \frac{v - \mu_k^\star}{\mu_k - \mu_k^\star} \right) \frac{y_k}{v^2} \frac{1}{\mu_k - \mu_k^\star} dv \\
&= \sum_{k=0}^K y_k \int_{\mu_k^\star}^{\mu_k} \frac{\mu_k - v}{v^2} dv \\
&= \sum_{k=0}^K y_k \left\{ \frac{\mu_k}{\mu_k^\star} - 1 - \log \frac{\mu_k}{\mu_k^\star} \right\}.
\end{aligned}$$

Recall that $\sum_{k=0}^K y_k = \sum_{k=0}^K \mu_k = \sum_{k=0}^K \mu_k^\star = 1$ and $y_k = \mu_k^\star + \xi_k$ for $k \in [K]$. Denoting $\xi_0 = -\sum_{k=1}^K \xi_k$, we then also have that $y_0 = \mu_0^\star + \xi_0$. Then, we have that

$$\ell(\boldsymbol{\mu}) - \ell(\boldsymbol{\mu}^\star) = y_0 \left\{ \frac{\mu_0}{\mu_0^\star} - 1 - \log \frac{\mu_0}{\mu_0^\star} \right\} + \sum_{k=1}^K \left\{ \frac{y_0}{\mu_0^\star} (\mu_k - \mu_k^\star) - y_k \log \frac{\mu_k}{\mu_k^\star} \right\}$$

$$
\begin{aligned}
&= \frac{y_0}{\mu_0^\star}\sum_{k=0}^K \mu_k - y_0 + y_0 \log\frac{\mu_0^\star}{\mu_0} + \sum_{k=1}^K \left\{ -\frac{y_0}{\mu_0^\star}\mu_k^\star + y_k \log\frac{\mu_k^\star}{\mu_k}\right\} \\
&= \frac{y_0}{\mu_0^\star} - \frac{y_0}{\mu_0^\star}\sum_{k=1}^K \mu_k^\star - y_0 + \sum_{k=0}^K y_k \log\frac{\mu_k^\star}{\mu_k} \\
&= \sum_{k=0}^K \mu_k^\star \log\frac{\mu_k^\star}{\mu_k} + \sum_{k=0}^K \xi_k \log\frac{\mu_k^\star}{\mu_k} \\
&= \sum_{k=0}^K \mu_k^\star \log\frac{\mu_k^\star}{\mu_k} + \sum_{k=0}^K \xi_k \log\frac{\mu_k^\star}{\mu_k} \\
&= \mathrm{KL}(\boldsymbol{\mu}^\star, \boldsymbol{\mu}) + \sum_{k=1}^K \xi_k \left(\log\frac{\mu_k^\star}{\mu_0^\star} - \log\frac{\mu_k}{\mu_0}\right) \\
&\overset{(*)}{=} \mathrm{KL}(\boldsymbol{\mu}^\star, \boldsymbol{\mu}) + \sum_{k=1}^K \xi_k \langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star^{(k)} - \boldsymbol{\theta}_t^{(k)}\rangle,
\end{aligned}
$$

where at $(*)$, we let $\boldsymbol{\mu}^\star = \boldsymbol{\mu}(\boldsymbol{x}_t, \boldsymbol{\theta}_\star)$ and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$. Then, with proper matrix notations, the statement follows.

D.5.2 Proof of Lemma 12

Denote $\mu_k^{(i)} = \mu_k(\boldsymbol{z}^{(i)})$ and $C_k^{(i)} := 1 + \sum_{j\neq k} e^{z_j^{(i)}}$. Then we have the following conversion between μ, C, and z:

$$
\mu_k^{(i)} = \frac{e^{z_k^{(i)}}}{C_k^{(i)} + e^{z_k^{(i)}}}, \quad z_k^{(i)} = \frac{\mu_k^{(i)} C_k^{(i)}}{1 - \mu_k^{(i)}}.
$$

The statement then follows from direct computation:

$$
\begin{aligned}
&D_m(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}) \\
&= m(\boldsymbol{z}^{(1)}) - m(\boldsymbol{z}^{(2)}) - \nabla m(\boldsymbol{z}^{(2)})^\intercal (\boldsymbol{z}^{(1)} - \boldsymbol{z}^{(2)}) \\
&= \log\left(1 + \sum_{k=1}^K e^{z_k^{(1)}}\right) - \log\left(1 + \sum_{k=1}^K e^{z_k^{(2)}}\right) - \sum_{k=1}^K \frac{e^{z_k^{(2)}}}{1 + \sum_{k=1}^K e^{z_k^{(2)}}}(z_k^{(1)} - z_k^{(2)}) \\
&= \log\frac{1 - \sum_{k=1}^K \mu_k^{(2)}}{1 - \sum_{k=1}^K \mu_k^{(1)}} - \sum_{k=1}^K \mu_k^{(2)} \log\frac{\mu_k^{(1)}(1-\mu_k^{(2)})C_k^{(1)}}{\mu_k^{(2)}(1-\mu_k^{(1)})C_k^{(2)}} \\
&= \left(1 - \sum_{k=1}^K \mu_k^{(2)}\right) \log\frac{1 - \sum_{k=1}^K \mu_k^{(2)}}{1 - \sum_{k=1}^K \mu_k^{(1)}} + \sum_{k=1}^K \mu_k^{(2)} \log\frac{\mu_k^{(2)}}{\mu_k^{(1)}} + \sum_{k=1}^K \mu_k^{(2)} \left\{ \log\frac{1 - \sum_{j=1}^K \mu_j^{(2)}}{1 - \sum_{j=1}^K \mu_j^{(1)}} - \log\frac{(1-\mu_k^{(2)})C_k^{(1)}}{(1-\mu_k^{(1)})C_k^{(2)}}\right\} \\
&= \mathrm{KL}(\boldsymbol{\mu}(\boldsymbol{z}^{(2)}), \boldsymbol{\mu}(\boldsymbol{z}^{(1)})) + \sum_{k=1}^K \mu_k^{(2)} \left\{ \log\frac{\sum_{j=1}^K e^{z_j^{(1)}}}{\sum_{j=1}^K e^{z_j^{(2)}}} - \log\frac{C_k^{(1)} + e^{z_k^{(1)}}}{C_k^{(2)} + e^{z_k^{(2)}}}\right\} \\
&= \mathrm{KL}(\boldsymbol{\mu}(\boldsymbol{z}^{(2)}), \boldsymbol{\mu}(\boldsymbol{z}^{(1)})).
\end{aligned}
$$

D.5.3 Proof of Lemma 13

By Proposition 8 of Sun and Tran-Dinh (2019), we have that for any $\boldsymbol{z}_1, \boldsymbol{z}_2$,

$$
\nabla^2 f(\boldsymbol{z}_1 + v(\boldsymbol{z}_2 - \boldsymbol{z}_1)) \succeq e^{-M_f \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 v} \nabla^2 f(\boldsymbol{z}_1).
$$

Multiplying both sides by $(1-v)$ and integrating over $[0,1]$ w.r.t. v, the statement follows:

$$
\begin{aligned}
\int_0^1 (1-v)\nabla^2 f(\boldsymbol{z}_1 + v(\boldsymbol{z}_2 - \boldsymbol{z}_1))dv &\succeq \int_0^1 (1-v) e^{-M_f\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 v} \nabla^2 f(\boldsymbol{z}_1) dv \\
&= \left(\frac{1}{M_f\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2} + \frac{\exp(-M_f\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2) - 1}{(M_f\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2)^2}\right) \nabla^2 f(\boldsymbol{z}_1)
\end{aligned}
$$

$$\succeq \frac{1}{2 + M_f \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2} \nabla^2 f(\boldsymbol{z}_1),$$

where the last inequality follows from the elementary inequality $\frac{1}{z} + \frac{e^{-z}-1}{z^2} \geq \frac{1}{2+z}$ for any $z \geq 0$.

D.5.4 Proof of Lemma 14

By Theorem 4, we have that with probability at least $1-\delta$, $\mathcal{L}_t(\boldsymbol{\theta}_\star) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \leq \beta_t(\delta)^2$, which we assume to be true throughout the proof. Let $\boldsymbol{\theta} \in \mathcal{C}_t(\delta)$. Then, we first have that via second-order Taylor expansion of $\mathcal{L}_t(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_\star$,

$$\begin{aligned}
\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star, \boldsymbol{\theta})} &= \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\boldsymbol{\theta}_\star) + \nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}) + \lambda_t \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_2^2 \\
&\leq \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) + \nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}) + \lambda_t \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_2^2 \\
&\leq K + \beta_t(\delta)^2 + \nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}), \quad \text{w.p. at least } 1-\delta, \qquad (49)
\end{aligned}$$

where we chose $\lambda = \frac{K}{4S^2}$.

Now observe that

$$\nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal \boldsymbol{v} = \sum_{s=1}^t \left[\left(\boldsymbol{\mu}(\boldsymbol{x}_s, \boldsymbol{\theta}_\star) - \boldsymbol{y}_s \right) \otimes \boldsymbol{x}_s \right]^\intercal \boldsymbol{v} = \sum_{s=1}^t \boldsymbol{\xi}_s^\intercal \mathrm{vec}^{-1}(\boldsymbol{v}) \boldsymbol{x}_s$$

where vec^{-1} is the matricization operator, and we overload the notation and define $\boldsymbol{\xi}_s := \boldsymbol{\mu}(\boldsymbol{x}_s, \boldsymbol{\theta}_\star) - \boldsymbol{y}_s$.

Let $\mathcal{B}^{dK}(2S)$ be a dK-ball of radius $2S$, and $\boldsymbol{v} \in \mathcal{B}^{dK}(2S)$. It can be easily checked that $\boldsymbol{\xi}_s^\intercal \mathrm{vec}^{-1}(\boldsymbol{v}) \boldsymbol{x}_s$ is also a martingale difference sequence that satisfies

$$\left| \boldsymbol{\xi}_s^\intercal \left(\mathrm{vec}^{-1}(\boldsymbol{v}) \boldsymbol{x}_s \right) \right| \leq 2S,$$

$$\mathbb{E}\left[\left(\boldsymbol{\xi}_s^\intercal \left(\mathrm{vec}^{-1}(\boldsymbol{v}) \boldsymbol{x}_s \right) \right)^2 \middle| \mathcal{F}_{s-1} \right] = \|\mathrm{vec}^{-1}(\boldsymbol{v}) \boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)}.$$

where for simplicity we denote $\boldsymbol{A}_\star(\boldsymbol{x}_s) := \boldsymbol{A}(\boldsymbol{x}_s, \boldsymbol{\Theta}_\star)$. Thus, by Freedman's inequality (Lemma 3), for any $\eta \in \left[0, \frac{1}{2S}\right]$, the following holds:

$$\mathbb{P}\left[\sum_{s=1}^t \boldsymbol{\xi}_s^\intercal \left(\mathrm{vec}^{-1}(\boldsymbol{v}) \boldsymbol{x}_s \right) \leq (e-2)\eta \sum_{s=1}^t \|\mathrm{vec}^{-1}(\boldsymbol{v}) \boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)} + \frac{1}{\eta} \log \frac{t}{\delta} \right] \geq 1 - \delta. \qquad (50)$$

Then, via similar reasoning (ε-net and union bound) as in the proof of Lemma 6, we have the following: for $\boldsymbol{v}_t$ s.t. $\|\boldsymbol{v}_t\|_2 \leq 2S$ and $\|(\boldsymbol{\theta}_\star - \boldsymbol{\theta}) - \boldsymbol{v}_t\|_2 \leq \varepsilon_t$,

$$\begin{aligned}
&\nabla \mathcal{L}_t(\boldsymbol{\theta}_\star)^\intercal (\boldsymbol{\theta}_\star - \boldsymbol{\theta}) \\
&= \sum_{s=1}^t \boldsymbol{\xi}_s^\intercal \left(\mathrm{vec}^{-1}(\boldsymbol{v}_t) \boldsymbol{x}_s \right) + \sum_{s=1}^t \boldsymbol{\xi}_s^\intercal \left(\mathrm{vec}^{-1}((\boldsymbol{\theta}_\star - \boldsymbol{\theta}) - \boldsymbol{v}_t) \boldsymbol{x}_s \right) && \text{(linearity of } \mathrm{vec}^{-1}\text{)} \\
&\leq (e-2)\eta \sum_{s=1}^t \|\mathrm{vec}^{-1}(\boldsymbol{v}_t) \boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)} + \frac{dK}{\eta} \log \frac{5S}{\varepsilon_t} + \frac{1}{\eta} \log \frac{t}{\delta} + \varepsilon_t t && \text{(w.p. at least } 1-\delta\text{)} \\
&= (e-2)\eta \left\{ \sum_{s=1}^t \|\mathrm{vec}^{-1}(\boldsymbol{\theta}_\star - \boldsymbol{\theta}) \boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)} + \sum_{s=1}^t \left(\|\mathrm{vec}^{-1}(\boldsymbol{v}_t) \boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)} - \|\mathrm{vec}^{-1}\left(\boldsymbol{\theta}_\star - \boldsymbol{\theta}\right) \boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)} \right) \right\} \\
&\quad + \frac{dK}{\eta} \log \frac{5S}{\varepsilon_t} + \frac{1}{\eta} \log \frac{t}{\delta} + \varepsilon_t t \\
&\overset{(*)}{\leq} (e-2)\eta \sum_{s=1}^t \left\| (\boldsymbol{\Theta}_\star - \boldsymbol{\Theta}) \boldsymbol{x}_s \right\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)} + (e-2)\eta L \left(4S + \varepsilon_t\right) \varepsilon_t t + \frac{dK}{\eta} \log \frac{5S}{\varepsilon_t} + \frac{1}{\eta} \log \frac{t}{\delta} + \varepsilon_t t \\
&\overset{(**)}{=} (e-2)\eta \|\boldsymbol{\theta}_\star - \boldsymbol{\theta}\|^2_{\boldsymbol{H}_t(\boldsymbol{\theta}_\star)} + \frac{dK}{\eta} \log \frac{5S}{\varepsilon_t} + \frac{1}{\eta} \log \frac{t}{\delta} + \left((e-2)L \left(4S\eta + \varepsilon_t \eta\right) + 1 \right) \varepsilon_t t
\end{aligned}$$

$$\leq (e-2)(2+2\sqrt{6}S)\eta\|\boldsymbol{\theta}_\star-\boldsymbol{\theta}\|^2_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta})}+\frac{dK}{\eta}\log\frac{5S}{\varepsilon_t}+\frac{1}{\eta}\log\frac{t}{\delta}+\left((e-2)L\left(4S\eta+\varepsilon_t\eta\right)+1\right)\varepsilon_t t,$$
$$(\boldsymbol{H}_t(\boldsymbol{\theta}_\star)\preceq(2+2\sqrt{6}S)\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta}))$$

where $(*)$ follows from the observation that

$$\begin{aligned}
\|\boldsymbol{C}\boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)}-\|\boldsymbol{D}\boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)}&=\|\boldsymbol{D}\boldsymbol{x}_s+(\boldsymbol{C}-\boldsymbol{D})\boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)}-\|\boldsymbol{D}\boldsymbol{x}_s\|^2_{\boldsymbol{A}_\star(\boldsymbol{x}_s)}\\
&=2\boldsymbol{x}_s^\intercal\boldsymbol{D}^\intercal\boldsymbol{A}_\star(\boldsymbol{x}_s)(\boldsymbol{C}-\boldsymbol{D})\boldsymbol{x}_s+\boldsymbol{x}_s^\intercal(\boldsymbol{C}-\boldsymbol{D})^\intercal\boldsymbol{A}_\star(\boldsymbol{x}_s)(\boldsymbol{C}-\boldsymbol{D})\boldsymbol{x}_s\\
&\leq 2\|\boldsymbol{D}^\intercal\boldsymbol{A}_\star(\boldsymbol{x}_s)(\boldsymbol{C}-\boldsymbol{D})\boldsymbol{x}_s\|_2+L\varepsilon_t^2 && \text{(Definition of } L \text{ (Eqn. (27)))}\\
&\leq 2\|\boldsymbol{D}^\intercal\|_2\|\boldsymbol{A}_\star(\boldsymbol{x}_s)\|_2\|(\boldsymbol{C}-\boldsymbol{D})\|_2+L\varepsilon_t^2\\
&\leq 2L\|\boldsymbol{D}^\intercal\|_F\|(\boldsymbol{C}-\boldsymbol{D})\|_F+L\varepsilon_t^2 && \text{(Definition of } L \text{ (Eqn. (27)))}\\
&\leq L\left(4S+\varepsilon_t\right)\varepsilon_t
\end{aligned}$$

for any $\boldsymbol{C},\boldsymbol{D}\in\mathbb{R}^{d\times K}$ with $\|\boldsymbol{C}\|_F,\|\boldsymbol{D}\|_F\leq 2S$ and $\|\boldsymbol{C}-\boldsymbol{D}\|_F\leq\varepsilon_t$. $(**)$ follows from the observation that for $\boldsymbol{\theta}=\mathrm{vec}(\boldsymbol{\Theta}^\intercal)$,

$$\begin{aligned}
\boldsymbol{\theta}^\intercal(\boldsymbol{A}\otimes\boldsymbol{x}\boldsymbol{x}^\intercal)\boldsymbol{\theta}&=\mathrm{vec}(\boldsymbol{\Theta}^\intercal)^\intercal(\boldsymbol{A}\otimes\boldsymbol{x}\boldsymbol{x}^\intercal)\mathrm{vec}(\boldsymbol{\Theta}^\intercal)\\
&\overset{(a)}{=}\mathrm{vec}(\boldsymbol{\Theta}^\intercal)^\intercal\mathrm{vec}\left(\boldsymbol{x}\boldsymbol{x}^\intercal\boldsymbol{\Theta}^\intercal\boldsymbol{A}^\intercal\right)\\
&\overset{(a)}{=}\mathrm{vec}(\boldsymbol{\Theta}^\intercal)^\intercal\left(\boldsymbol{A}\boldsymbol{\Theta}\otimes\boldsymbol{x}\right)\boldsymbol{x}\\
&\overset{(b)}{=}\boldsymbol{x}^\intercal\left(\boldsymbol{\Theta}^\intercal\boldsymbol{A}^\intercal\otimes\boldsymbol{x}^\intercal\right)\mathrm{vec}(\boldsymbol{\Theta}^\intercal)\\
&\overset{(a)}{=}\boldsymbol{x}^\intercal\mathrm{vec}(\boldsymbol{x}^\intercal\boldsymbol{\Theta}^\intercal\boldsymbol{A}\boldsymbol{\Theta})\\
&=\boldsymbol{x}^\intercal\boldsymbol{\Theta}^\intercal\boldsymbol{A}\boldsymbol{\Theta}\boldsymbol{x},
\end{aligned}$$

where (a) follows from the mixed Kronecker matrix-vector product property, $(\boldsymbol{C}\otimes\boldsymbol{D})\mathrm{vec}(\boldsymbol{E})=\mathrm{vec}(\boldsymbol{D}\boldsymbol{E}\boldsymbol{C}^\intercal)$, and (b) follows from the tranpose property of the Kronecker product, $(\boldsymbol{C}\otimes\boldsymbol{D})^\intercal=\boldsymbol{C}^\intercal\otimes\boldsymbol{D}^\intercal$.

Choosing $\eta=\frac{1}{2(e-2)(2+2\sqrt{6}S)}<\frac{1}{2S}$, $\varepsilon_t=\frac{dK}{t}$, and rearranging Eqn. (49) with Theorem 4, we finally have that

$$\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|^2_{\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}_\star,\boldsymbol{\theta})}\lesssim dKS\log\frac{St}{dK}+\sqrt{K}S\log\frac{t}{\delta}+dKL.$$

E EXPERIMENTS – LOGISTIC BANDITS

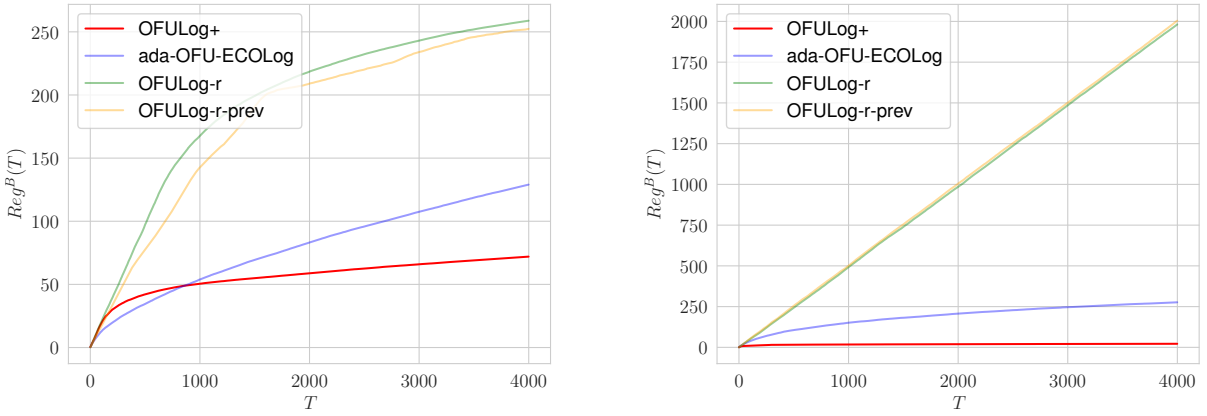
E.1 Setting

We follow the experimental setting of Faury et al. (2022) and compare our OFULog+ with three other *tractable* versions of the state-of-the-art algorithms: `ada-OFU-ECOLog` (Faury et al., 2022), `OFULog-r-prev` (Abeille et al., 2021), and `OFULog-r` (Abeille et al., 2021). Here, `OFULog-r` refers to the tractable algorithm of Abeille et al. (2021) with the improved $\lambda_t = \frac{d}{S}\log\frac{St}{d\delta}$, and `OFULog-r-prev` refers to the same algorithm with the original $\lambda_t = d\log t$. Also, the implementation of Abeille et al. (2021) as presented in Faury et al. (2022) was a bit inaccurate as the authors used only a few steps of Newton's method to approximate the MLE. For a fair comparison, we replace this with Sequential Least SQuares Programming (SLSQP) implemented in SciPy (Virtanen et al., 2020). Throughout the experiments, we fix $T = 4000$, $d = 2$, $|\mathcal{A}| = 20$, and $\delta = 0.01$. We use $\boldsymbol{\theta}_\star = \frac{S-1}{\sqrt{d}}\mathbf{1}$ and time-varying arm-set by sampling in the unit ball at random at each t.

E.2 Comparing Regrets

For the experiments, we consider $S \in \{2, 10\}$, which results in $\kappa = 9$ and $\kappa = 22028$, respectively. As κ often scales exponentially in S, such drastically large κ is to be expected for $S = 10$. The results, averaged over 10 independent runs, are shown in Figure 1(a) and 1(b). It is clear that `OFULog+` outperforms other algorithms significantly.

Interestingly, it can be observed that for both cases, at the initial phase (e.g., $T \leq 1000$ for $S = 2$), `OFULog+` underperforms compared to `ada-OFU-ECOLog`, but then the regret of `OFULog+` flattens much faster than the other algorithms, leading to the best (cumulative) regret at the end. This is more pronounced when compared to the existing confidence-set-based algorithms, `OFULog-r` and `OFULog-r-prev`.



(a) $S = 2$ (b) $S = 10$

Figure 1: Plot of $\mathrm{Reg}^B(T)$ for all considered algorithms.

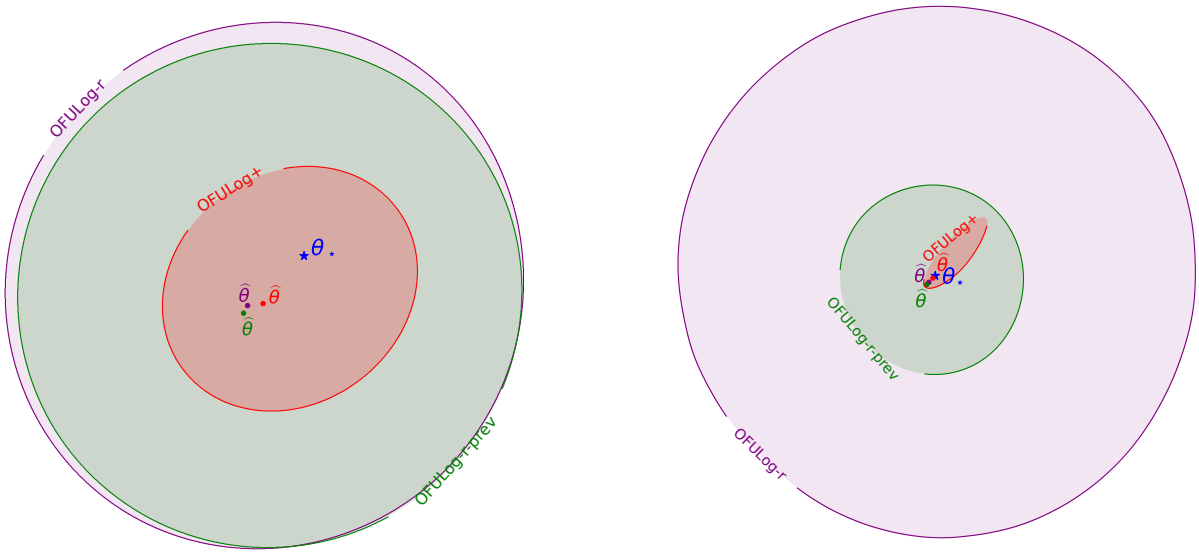
E.3 Comparing Confidence Sets

In Figure 2, we plot the confidence sets at $t = 500$ resulting from `OFULog-r-prev`, `OFULog-r`, and `OFULog+`, for $S \in \{2, 10\}$. Indeed, it can be seen that our confidence set is significantly tighter than that of the other confidence-set-based algorithms! Especially for $S = 10$, several crucial observations can be made:

- Distance-wise, the MLE resulting from our `OFULog+` is the closest to the unknown parameter $\boldsymbol{\theta}_\star$.
- Our confidence set is skewed towards $\boldsymbol{\theta}_\star$, i.e., `OFULog+` oversamples near $\boldsymbol{\theta}_\star$. This is reminiscent of the nonconvex (and non-tractable) confidence sets proposed in Abeille et al. (2021); Faury et al. (2020). Thus,

our O2CS is the first approach to result in the tightest yet tractable loss-based confidence set that displays similar adaptivity.

- Despite the fact that the theory predicts that `OFULog-r` should perform better than `OFULog-r-prev` due to improved λ_t, `OFULog-r` has a much looser confidence set.



(a) $S = 2$ (b) $S = 10$

Figure 2: Confidence sets at $t = 500$.