Transfer Learning in Bandits with Latent Continuity

Hyejin Park, Seiyun Shin, Kwang-Sung Jun, and Jungseul Ok

***Abstract*—A continuity structure of correlations among arms in multi-armed bandit can bring a significant acceleration of exploration and reduction of regret, in particular, when there are many arms. However, it is often latent in practice. To cope with the latent continuity, we consider a transfer learning setting where an agent learns the structural information, parameterized by a Lipschitz constant and an embedding of arms, from a sequence of past tasks and *transfers* it to a new one. We propose a simple but provably-efficient algorithm to accurately estimate and fully exploit the Lipschitz continuity at the same asymptotic order of lower bound of sample complexity in the previous tasks. The proposed algorithm is applicable to estimate not only a latent Lipschitz constant given an embedding, but also a latent embedding, while the latter requires slightly more sample complexity. To be specific, we analyze the efficiency of the proposed framework in two folds: (i) our regret bound on the new task is close to that of the oracle algorithm with the full knowledge of the Lipschitz continuity under mild assumptions; and (ii) the sample complexity of our estimator matches with the information-theoretic fundamental limit. Our analysis reveals a set of useful insights on transfer learning for latent Lipschitz continuity. From a numerical evaluation based on real-world dataset of rate adaptation in time-varying wireless channel, we demonstrate the theoretical findings and show the superiority of the proposed framework compared to baselines.**

***Index Terms*—Multi-armed bandits, Lipschitz continuity, transfer learning, wireless rate adaptation**

I. Introduction

THE classical stochastic multi-armed bandit (MAB) [1] of independent K arms in time-horizon T has the fundamental limit of regret $O(K \log T)$ which scales linearly with K and thus is infeasible in case of having large K. To cope with large number of arms, one may exploit a structural assumption on underlying correlations among the arms. This idea has been demonstrated with a wide spectrum of structural assumptions, e.g., Lipschitz continuity [2], [3], weakly Lipschitz ($\mathcal{X}$-armed bandits) [4], linearity [5], convexity [6], or graphical unimodality [7]. These studies indeed derive bandit algorithms whose regret does not scale in K under some canonical scenarios, i.e., *scale-free* regret. To be specific, in case of the Lipschitz continuity structure described by a Lipschitz constant L and an embedding $\boldsymbol{x} = (x(1), ..., x(K))$ of K arms on some metric space, it is possible to obtain a regret bound of $O(\min\{K, \text{poly}(L)\} \log T)$ [8].

Such a benefit from the continuity structure, however, requires prior knowledge on the Lipschitz constant L and the embedding of arms. In this case, it is natural to estimate the latent knowledge from similar tasks. For example, rate adaptation over non-stationary wireless channels [9], [10] faces a sequence of bandit problems with similar structural properties, while one can find a continuity structure among transmission rates. The channel often changes discretely over time, and such a discrete change distinguishes tasks in learning scenarios. This motivates us to study a *transfer learning* problem, where we aim to learn the latent Lipschitz continuity from previous tasks and use it for the next task.

In order to build provably sample-efficient algorithms for the transfer bandit problem, we investigate the risks of using wrong estimation of Lipschitz constant L in two main failure scenarios (Section II-C), where we focus on estimating the latent L for simplicity. Overestimating L leads to an unnecessary regret, attenuating the benefit from the structure, such as scale-free regret. On the other hand, underestimating L can cause a catastrophic failure of suffering a linear regret. Accordingly, we design an estimator for L, which balances between the two extremes. Using this estimator, we show that one can nearly achieve the minimal regret of the oracle algorithm knowing L exactly (Section III-B). Furthermore, our estimator is asymptotically optimal in that the sample complexity on previous tasks matches to the one in a fundamental limit analysis (Section III-D).

In addition to transferring latent L, we also consider the problem with latent embedding $\boldsymbol{x}$, as the embedding is frequently unknown in practice (Section III-F). Interestingly, our asymptotic analysis reveals that the sample complexity to learn latent embedding is only slightly larger than that to learn latent L. Finally, using the real-world dataset of rate adaptation to time-varying wireless channel [9], we provide a numerical justification for not only our theoretical findings but also the superiority over baseline algorithms in a realistic scenario of sequential bandit problem with latent continuity.

Related work. In [11], the authors provide a generic approach to constructing optimal algorithms when we have the complete knowledge on the structural property, including but not limited to Lipschitz [2], linear [5], convex [6], or unimodal [7]. The structural knowledge, however, is often incomplete in practice as aforementioned. In this regard, the work of [12] has shown that it is possible to achieve the minimax optimal regret of $\Theta(L^{D/(D+2)}T^{(D+1)/(D+2)})$ without any knowledge of the Lipschitz constant for the continuum of arms, i.e., infinite candidate structures, where D is the dimension of the embedding space. Observing that the minimax regret is mostly generated by neighboring arms of the optimal arm (although they have only thimbleful gaps to the best arm), the minimax analysis is extended to the case with unknown local smoothness (a weaker notion of Lipschitz continuity) [13]. Moreover, a similar minimax analysis can be conducted for online optimization problems to find a point x_T in an embedded space so as to minimize its regret $R_T^{\pi} = \sup_x f(x) - f(x_T)$

*Correspondence to Jungseul Ok. Author emails: parkebbi2@postech.ac.kr, seiyuns2@illinois.edu, kjun@cs.arizona.edu, and jungseul@postech.ac.kr. A part of this paper was presented at IEEE International Symposium on Information Theory (ISIT), 2021.

given a payoff function f, and establish stochastic adaptive strategies with optimal *simple* regret of $\tilde{O}(T^{-1/2})$, e.g., [14]–[19]. Another line of works on the incomplete knowledge of structures formulate model selection problems to identify the best structure providing the minimax regret given a finite set of candidate structures [20], [21]. While minimax regret analysis, which yields asymptotic regret $O(\sqrt{T})$ from infinitesimally sub-optimal arms in worst-case scenarios, provides meaningful insights into these scenarios, it often overlooks the substantial improvements possible through instance-dependent analyses of $O(\log T)$ regret for various (known) structures [2], [5]–[7], [11], [22]–[24]. Given these insights, we focus on instance-dependent regret analysis with unknown structures, exploiting latent patterns without prior structural knowledge.

For the instance-dependent regret analysis with unknown structures, the pioneering work of [25] has proposed a transfer learning framework where the learner faces a sequence of bandit tasks that are randomly drawn from a distribution over a *finite* set of problem instances. They suggest an algorithm that leverages the robust tensor power method to learn the underlying set of instances while repeatedly solving each task. Their algorithms, however, only consider a finite number of instances, in contrast to the Lipschitz structure we consider, which consists of infinitely many instances. Although a few follow-up studies [26], [27] explore infinite instance sets, they focus on simple linear structures and minimax regret rather than instance-dependent regret. Additionally, the work of [28] addresses a meta-learning problem in non-stationary bandit environments. While they focus on efficiently identifying optimal arms that evolve with each task, our approach targets minimizing instance-dependent regret by exploiting latent patterns, emphasizing the need to adapt to unknown structures. To our knowledge, we are the first to study transfer learning in bandits with instance-dependent optimality beyond the simple case of the finite instance set.

Contribution. Our main contributions of this article are summarized below:

- We introduce the transfer learning problem for learning the latent Lipschitz continuity where we aim to fully exploit the benefit, e.g., scale-free regret for canonical settings (Property 1), of the Lipschitz continuity structure, parameterized by constant L and embedding $\boldsymbol{x}$.
- In the problem of estimating the latent L, we propose the provably efficient estimator $\hat{L}_\beta$ from the knowledge of previous tasks. Given rich enough experience from past tasks, a bandit algorithm using the estimator $\hat{L}_\beta$ closely achieves the problem-dependent regret lower bound of oracle knowing true L (Theorem 3). Additionally, the asymptotic order of the sample complexity that $\hat{L}_\beta$ requires coincides with the information-theoretic fundamental limit (Theorem 6).
- We postulate a plausible distribution of generating tasks and establish a theoretical performance guarantee on the proposed method (Theorem 5). This provide not only useful guidance to tune the hyperparamter β of our framework $\hat{L}_\beta$ but also a heuristic algorithm automatically adopting β.
- Building upon the estimation of latent L with known embedding, we also study the case of latent embedding of arms that have need of slightly more sample complexity (Section III-F).
- We numerically demonstrate our insights from analysis and justify the superiority of the proposed schemes compared to baselines in a realistic scenario of sequential bandit problems with latent continuity, built based on the real-world dataset of rate adaptation over time-varying channel [9] (Section IV).

Outline. The paper is organized as follows. In Section II, we provide useful insights from the fundamental limit and optimal algorithm with prior knowledge on Lipschitz constant L and discuss the importance and challenges of estimating L. In Section III, we describe the transfer learning setting and the proposed framework with performance guarantees. Section IV contains results of our numerical evaluation in practical scenarios. We conclude our paper with exciting future work in Section V.

II. Preliminary

A. Lipschitz bandit model

Let $\boldsymbol{\theta} = (\theta(1), ..., \theta(K))$ denote a multi-armed bandit instance, where each play of arm $i \in [K] := \{1, 2, ..., K\}$ generates a random variable $X(i,t)$ drawn i.i.d. from Bernoulli distribution with mean $\theta(i) \in [0,1]$ in round t. For ease of exposition, we restrict our attention to Bernoulli distribution, while our analysis can be generalized to the exponential family with a single parameter [29], [30]. At each round $t = 1, 2, ...$, the decision maker π selects an arm $i_t \in [K]$, pulls it, and then receives a reward $X(i_t, t)$ drawn from the distribution associated with the arm i_t. Let $\theta_* := \max_{i\in[K]} \theta(i)$ and $\mathcal{K}_*(\boldsymbol{\theta}) := \{i \in [K] : \theta(i) = \theta_*\}$ denote the best mean reward and the set of best arms, respectively. For a given multi-armed bandit instance $\boldsymbol{\theta}$, an algorithm π aims to maximize the expected cumulative rewards over the time horizon T. This aim is equivalent to minimizing the regret defined as follows:

$$R_T^\pi(\boldsymbol{\theta}) := \sum_{i\in[K]} \big(\theta_* - \theta(i)\big)\, \mathbb{E}_\pi[n_T(i)]\ ,$$

where $n_T(i)$ is the number of pulling arm i up to time T, and the expectation $\mathbb{E}_\pi$ is taken w.r.t. the randomness induced by both the rewards and the algorithm π. The regret $R_T^\pi(\boldsymbol{\theta})$ can also be viewed as the expected opportunity cost for selecting sub-optimal arms.

We consider the set of mean rewards where the arms are constrained to satisfy Lipschitz condition w.r.t. an embedding of the arms $\boldsymbol{x} = (x(1), ..., x(K)) \in [0,1]^{D\times K}$, which is commonly referred to as the Lipschitz structure with a constant L [2]. Formally, denoting by $\boldsymbol{d} := (d(i,j)) \in \mathbb{R}^{K\times K}$ the distance matrix for each pair of arms such that $d(i,j) := \|x(i) - x(j)\|$ for the embedding $\boldsymbol{x}$, the Lipschitz structure $\Phi(L; \boldsymbol{d})$ is defined as follows:

$$\Phi(L;\boldsymbol{d}) := \Big\{\boldsymbol{\theta}\in[0,1]^K\colon |\theta(i)-\theta(j)| \le L\cdot d(i,j)\ \forall i,j\in[K]\Big\}\ . \quad (1)$$

As indicated in (1), the structure is encoded by Lipschitz constant L and embedding distance $d(i,j)$. To reduce complexity, we assume that the Lipschitz constant $L>0$ is latent, while we know the embedding $\boldsymbol{x}$, i.e., the relative similarity $(d(i,j))_{i,j\in[K]}$ among the arms. We first analyze the case of a latent Lipschitz constant L with known embedding. Then, we expands to the case of a latent embedding for arbitrary known L, which requires learning each pair of arms. A detailed description of the latter case, unknown embedding distance learning, is postponed to Section III-F.

B. Optimal regret with known structure

In this subsection, for ease of exposition, we discuss the gain from Lipschitz structure when varying L for given $\boldsymbol{d}$ (or $\boldsymbol{x}$). By omitting $\boldsymbol{d}$, we assume that the learner knows that the instance $\boldsymbol{\theta}$ conforms to the structure $\Phi(L)$. We say an algorithm π is uniformly good for $\Phi(L)$ if $\mathbb{E}_\pi[n_T(i)]=o(T^\rho)$ for all $\rho>0$, $\boldsymbol{\theta}\in\Phi(L)$ and $i\notin\mathcal{K}_*(\boldsymbol{\theta})$. That is, a uniformly good algorithm has ability to adapt to any $\boldsymbol{\theta}\in\Phi(L)$ and enjoys a sublinear regret in T. Then, any uniformly good algorithm has the following fundamental limit:

Theorem 1 (Regret lower bound with known L [2])**.** *Let π be a uniformly good algorithm for $\Phi(L)$. For any $\boldsymbol{\theta}\in\Phi(L)$,*

$$\liminf_{T\to\infty}\frac{R_T^\pi(\boldsymbol{\theta})}{\log T}\geq C(\boldsymbol{\theta},L)\ , \tag{2}$$

where $C(\boldsymbol{\theta},L)$ is the optimal value of the following linear programming (LP):

$$\min_{\boldsymbol{\eta}\succcurlyeq 0}\ \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})}\big(\theta_*-\theta(i)\big)\eta(i) \tag{3a}$$

$$\text{s.t.}\ \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})}\mathrm{KL}(\theta(i)\|\lambda^j(i;\boldsymbol{\theta},L))\eta(i)\geq 1,\forall j\notin\mathcal{K}_*(\boldsymbol{\theta})\ . \tag{3b}$$

Here $\lambda^j(i;\boldsymbol{\theta},L):=\max\{\theta(i),\theta_*-L\cdot d(i,j)\}$ for all $i,j\in[K]$, and $\mathrm{KL}(\theta\|\lambda)$ is the Kullback-Leibler divergence between Bernoulli distributions with mean θ and λ, i.e., $\mathrm{KL}(\theta\|\lambda):=\theta\log\big(\theta/\lambda\big)+(1-\theta)\log((1-\theta)/(1-\lambda))$.

Consider $\eta(i)\simeq n_T(i)\log T$ as the rate of exploration arm i and define $\mathcal{D}(\boldsymbol{\theta},L)$ as the set of all feasible $\boldsymbol{\eta}\in\bar{\mathbb{R}}_+^K$ verifying all the constraints in (3b) where $\bar{\mathbb{R}}_+:=\{x\in\mathbb{R}:x\geq 0\}\cup\{\infty\}$. Then, $C(\boldsymbol{\theta},L)$ which is the optimal value of LP (3) is the minimal rate of regret while $\boldsymbol{\eta}\in\mathcal{D}(\boldsymbol{\theta},L)$. Furthermore, the condition with $\boldsymbol{\lambda}^j$ in (3b) can be interpreted as a necessary condition to statistically distinguish $\boldsymbol{\lambda}^j$ from $\boldsymbol{\theta}$, where the construction of $\boldsymbol{\lambda}^j$ from $\boldsymbol{\theta}$ is the minimal perturbation to make (originally suboptimal) arm $j\notin\mathcal{K}_*(\boldsymbol{\theta})$ best in $\boldsymbol{\lambda}^j$ while verifying the Lipschitz continuity, i.e., $\boldsymbol{\lambda}^j\in\Phi(L)$. To sum up, it is the most confusing bandit problem which needs to be distinguished from $\boldsymbol{\theta}$ to identify the best arm under $\boldsymbol{\lambda}^j$.

Optimal algorithms for known L. We set $\eta(i;\boldsymbol{\theta},L)=\infty$ for optimal arm $i\in\mathcal{K}_*(\boldsymbol{\theta})$. Then, for $i\notin\mathcal{K}_*(\boldsymbol{\theta})$, $\eta(i;\boldsymbol{\theta},L)\log t$ provides a suggestion on minimal exploration at time t. This motivates to an algorithm that keeps tracking the estimated LP solution $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}_t,L)$ (if we knew L a priori) where $\hat{\boldsymbol{\theta}}_t$ is the estimation of $\boldsymbol{\theta}$ at time t. Indeed, there have been a number of algorithms that use this framework to achieve the asymptotic lower bound in Theorem 1; e.g. OSSB (Optimal Sampling for Structured Bandit) [11] and DEL (Directed Exploration Learning) [8]. In this paper, we use Algorithm 1, denoted by $\pi(L)$ for a given L, that is a simplified version of DEL algorithm* [8], which is originally designed for structured Markov decision process.

Algorithm 1 consists of four phases: monotonization, estimation, exploitation, and exploration. In the monotionization phase, the algorithm aims to accurately identify the true optimal arm for exploitation phase, along with restrictions on set of well-choosen arms using $\zeta_t=1/(1+\log\log t)$ [31]. Before the exploitation and exploration phase, updating $\hat{\theta}_t(i)=\hat{\theta}_*$ within the margin ζ_t effectively expands the set of arms considered to be near-optimal, treating them as potential optimal arms. This ensures that the estimated mean rewards are closely aligned with the true mean rewards, facilitating accurate identification of the best arms. The estimation phase ensures that every arm is sampled at least $\Omega(\log t/\log\log t)$. Although this additional sampling does not significantly impact the regret asymptotically $\log T/\log\log T=o(\log T)$, it ensures concentrations of the estimates $\hat{\boldsymbol{\theta}}_t$ to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}_t,L)$ to $\boldsymbol{\eta}(\boldsymbol{\theta},L)$, respectively. In the exploitation and exploration phase, the algorithm utilizes a clipped LP solution $\boldsymbol{\eta}_t$ as follows:

$$\eta_t(i):=\min\{\log t,\eta(i;\hat{\boldsymbol{\theta}}_t,L)\},\forall i\in[K]\ .$$

In the exploitation phase, the algorithm exploits the current best arm if the current exploration is statistically sufficient to identify the best arm. In other words, the number of pulling arm up to time t is sufficient to satisfy the minimal exploration at time t, i.e., $n_t(i)\geq(1+\lambda)\eta_t(i)\log t$ for all arms with some positive margin $\lambda>0$. In the exploration phase, it explores the most under-explore arm toward clipped LP solution $\boldsymbol{\eta}_t$.

The following theorem shows the asymptotic optimality of Algorithm 1 up to an arbitrarily small constant factor $(1+\lambda)$.

Theorem 2 (Regret upper bound with known L [2])**.** *Consider $\boldsymbol{\theta}\in\Phi(L)$ such that for each $i\notin\mathcal{K}_*(\boldsymbol{\theta})$, the LP solution $\eta(i;\boldsymbol{\theta},L)$ is unique and continuous at $\boldsymbol{\theta}$. Then, for any given $\lambda>0$, an algorithm $\pi=\pi(L)$ has*

$$\limsup_{T\to\infty}\frac{R_T^\pi(\boldsymbol{\theta})}{\log T}\leq(1+\lambda)C(\boldsymbol{\theta},L)\ . \tag{5}$$

Theorem 2 implies that when using exact value of L, the algorithm can asymptotically achieve the regret lower bound derived in Theorem 1. It is remarkable to the fundamental limit of regret $C(\boldsymbol{\theta},L)$ in that there exist matching upper bounds including our algorithm $\pi(L)$. However, $C(\boldsymbol{\theta},L)$ is implicitly defined in that it hides its scaling with the number of arms K, the dimension of the embedding space D, and other instance-dependent characteristics such as the smallest gap $\Delta_{\boldsymbol{\theta}}$. To gain more understanding, we provide an upper bound on $C(\boldsymbol{\theta},L)$ as a function of these parameters:

*We choose DEL algorithm as it has better empirical behavior than OSSB thanks to the careful handling (such as monotization) for cases where the assumptions for the analysis are broken.

Algorithm 1 $\pi(L)$

if $\exists i \in [K]$ s.t. $n_1(i) = 0$ **then**
  {Initialization} Play an arm chosen uniformly at random from $\{i \in [K] : n_1(i) = 0\}$
end if
for $t = 1, 2, \ldots$ **do**
  if $\forall i \in \mathcal{K}_*(\hat{\boldsymbol{\theta}}_t)$ s.t. $n_t(i) \leq \log t + 1$ **then**
    {Monotonization} Play the most under-sampled current best arm $i_t \in \arg\min_{i \in \mathcal{K}_*(\hat{\boldsymbol{\theta}}_t)} n_t(i)$
  else if $\exists i \in [K]$ s.t. $n_t(i) \leq \frac{\log t}{\log \log t}$ † **then**
    {Estimation} Play the most under-sampled arm $i_t \in \arg\min_{i \in [K]} n_t(i)$
  else
    Compute $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}_t, L)$ and set $\boldsymbol{\eta}_t$ as follows:

$$\eta_t(i) = \min\{\log t^{\dagger}, \eta(i; \hat{\boldsymbol{\theta}}_t, L)\} \quad \forall i \in [K] \tag{4}$$

    Update $\hat{\theta}_t(i) = \hat{\theta}_*$ s.t. $|\hat{\theta}_t(i) - \hat{\theta}_*| \leq \zeta_t$ where $\zeta_t := \frac{1}{1+\log\log t^{\dagger}}$
    if $\forall i \in [K]$ s.t. $n_t(i) \geq (1+\lambda)\eta_t(i) \log t$ **then**
      {Exploitation} Exploit the most under-sampled current best arm $i_t \in \arg\min_{i \in \mathcal{K}_*(\hat{\boldsymbol{\theta}}_t)} n_t(i)$
    else
      {Exploration} Explore the most under-explored arm $i_t \in \arg\max_{i \in [K]} \left(\eta_t(i) \log t - n_t(i)\right)$
    end if
  end if
  Update statistics $\hat{\boldsymbol{\theta}}_{t+1}$ based on new reward r_t corresponding to arm i_t for each $i \in [K]$, set:

$$n_{t+1}(i) = n_t(i) + \mathbb{1}[i_t = i]; \qquad \text{and} \qquad \hat{\theta}_{t+1}(i) = \hat{\theta}_t(i) + \mathbb{1}[i_t = i] \left(\frac{r_t - \hat{\theta}_t(i)}{n_t(i) + 1} \right).$$

end for

† Positive constants can be multiplied to these terms for stabilizing the empirical behavior without harming the asymptotic optimality in Theorem 2.

Property 1. *Let* $\Delta_{\boldsymbol{\theta}} := \min_{i \notin \mathcal{K}_*(\boldsymbol{\theta})} \theta_* - \theta(i)$ *denote the smallest suboptimality gap. Then, for* $L > 0$ *and* $\boldsymbol{\theta} \in \Phi(L)$*,*

$$C(\boldsymbol{\theta}, L) \leq \frac{8}{\Delta_{\boldsymbol{\theta}}^2} \min \left\{ K, \left(\frac{8L\sqrt{D}}{\Delta_{\boldsymbol{\theta}}} + 1 \right)^D \right\}, \tag{6}$$

where D *is the dimension of the embedding space.*

By an upper bound on $C(\boldsymbol{\theta}, L)$, it is worth noting that exploiting the exact Lipschitz structure provides a drastic reduction in regret. Given a fixed $\Delta_{\boldsymbol{\theta}}$ and D, the fundamental limit of $C(\boldsymbol{\theta}, L)$ is scale-free, meaning that it does not scale with the number of arms K. In particular, this is a dramatic advantage over no continuity structure, i.e., $L = \infty$, whose optimal regret does scale with K.

C. Impact of incorrect estimation of L

In the context of *latent* Lipschitz constant L, one needs to estimate L from observed samples, which can be quite inaccurate. We study the impact of using an incorrect estimator L' in the following two cases: (i) $L' > L$; and (ii) $L' < L$.

(i) $L' > L$: When L' is overestimated, we have $\Phi(L) \subset \Phi(L')$, thus yielding:

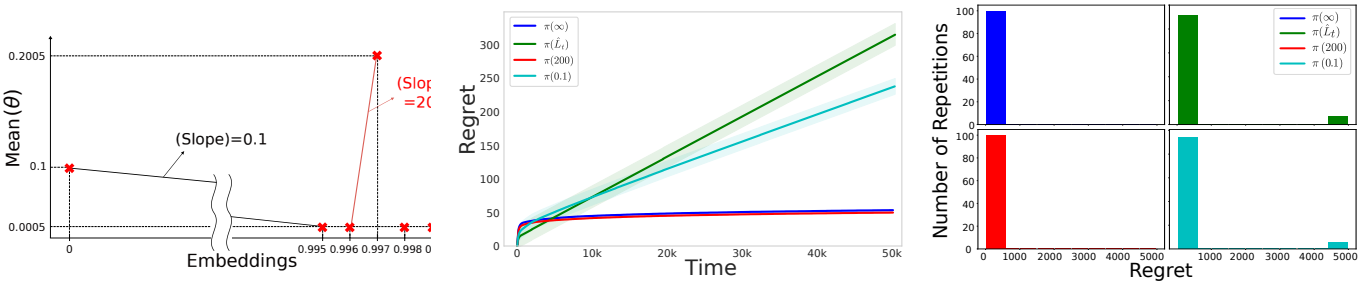
$$C(\boldsymbol{\theta}, L) \leq C(\boldsymbol{\theta}, L') \quad \forall \boldsymbol{\theta} \in \Phi(L) \,. \tag{7}$$

In other words, *a small L implies a stronger structure and thus a smaller regret.* In this case, the regret of algorithm $\pi(L')$ is provably bounded from above $C(\boldsymbol{\theta}, L')$, and thus causes a larger regret. Due to the conservative choice of L', the regret rate of algorithm $\pi(L')$ can be degenerated into that of the unstructured case. Then, in the most extreme case, the regret scales with K as discussed above, which is problematic for larger K.

(ii) $L' < L$: It can be highly risky to explore with a smaller Lipschitz constant L' (i.e., a stronger structure) than the true L. L' is underestimated, which implies that $\boldsymbol{\theta} \in \Phi(L) \setminus \Phi(L') \neq \emptyset$. The true parameter $\boldsymbol{\theta} \notin \Phi(L')$ does not appear in the LP constraints that the algorithm solves. This results that once the algorithm starts exploiting an incorrect best arm, it may not be able to collect sufficient statistical evidence to correct itself. Therefore, the algorithm $\pi(L')$ has a considerable risk of suffering a linear regret.

We confirm this phenomenon by a numerical simulation where the mean rewards of 6 arms $\boldsymbol{\theta} = (0.1, 0.0005, 0.0005, 0.2005, 0.0005, 0.0005)$ and their embedding $\boldsymbol{x} = (0, 0.995, 0.996, 0.997, 0.998, 0.999)$ are illustrated in Fig. 1(a). We set the true Lipschitz constant L to be 200 and the second steepest slope to be 0.1. For each time step t, we define the estimator $\hat{L}_t := \max_{i \neq j \in [K]} \frac{|\hat{\theta}_t(i) - \hat{\theta}_t(j)|}{d(i,j)}$. According to this setting, we compare four algorithms: $\pi(\infty)$, $\pi(\hat{L}_t)$, $\pi(200)$, and $\pi(0.1)$. Fig. 1(b). shows the mean and variance of 100 iterations of the experiment with the four algorithms, and time horizon of each iteration is $T = 5 \times 10^4$. We observe that the most conservative choice of $L = \infty$ and the exact choice of $L = 200$ show similar logarithmic regrets. On the other hand, the aggressive choice of $L = 0.1$ can suffer from an almost

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3441669



(a) Bandit parameter $\boldsymbol{\theta}$ and embedding $\boldsymbol{x}$ with $L = 200$ (b) Regret of 4 algorithms over time (c) Histogram of regret at $T = 50k$ of 4 algorithms

Fig. 1. Comparison among $\pi(\infty)$, $\pi(200)$, $\pi(0.1)$ and $\pi(\hat{L}_t)$ for given $\boldsymbol{\theta}$ and $\boldsymbol{x}$ shown in Fig. 1(a).

linear regret due to the mismatch between the true structure and its belief. The last method $\pi(\hat{L}_t)$ also shows an almost linear regret, despite continuing to update its estimator for L after each time step. This is because the Lipschitz constant is often underestimated by undersampling the hidden true best arm 4. The underestimated $\hat{L}_t$ reinforces the algorithm to reduce the exploration rate on the true best arm, thus losing the opportunity to correct $\hat{L}_t$. Indeed, in Fig. 1(c), each of $\pi(\hat{L}_t)$ and $\pi(0.1)$ suffers a certain portion of catastrophic failures out of 100 sample paths, that also explains relatively large regrets in Fig. 1(b).

We further analyze the failure of $\pi(\hat{L}_t)$ with Fig. 2. Fig. 2(a) presents the histogram of $\hat{L}_T$ at $T = 50k$ for 100 iterations. A sustainable portion of sample paths (6 out of 100; the first bin in Fig. 2(a)) has the estimation $\hat{L}_T$ concentrated around the second-steepest slope (0.1) in Fig. 1(a). Suppose that at certain iteration t, the second best arm 1 (accidentally) has a much higher empirical mean than the others, i.e., $\hat{\theta}(1) \gg \max_{i=2,\dots,6} \hat{\theta}(i) \approx 0$. Then, the value of $\hat{L}_t$ is much lower than true L as arm 1 is far from every other point. The underestimated $\hat{L}_t$ forces the algorithm overgeneralize and excessively reduces the exploration rate on arms 2-6 and also the chance of correcting $\hat{L}_t$. This is demonstrated in Fig. 2(b) and Fig. 2(c) which compare two groups of sample paths with the bottom-5% or top-5% values of $\hat{L}_T$ at $T = 50k$ in Fig. 2(a). As shown in Fig. 2(b), the bottom-5% group with under-estimated $\hat{L}_t$ misidentifies the second best arm 1 (i.e., $x = 0$) as the best arm and mainly contributes the linear regret. In Fig. 2(c), the best arm is hidden from the fact that the Lipschitz constant of the bottom-5% group is estimated to be almost 0.1 continuously over time. In other words, the bottom-5% group can hardly recover from the estimation error in $\hat{L}_t$, which provides an explanation on such a catastrophic failure. To summarize, our simulation shows that simultaneously learning structure and minimizing regret in $\pi(\hat{L}_t)$ in transfer learning setting is highly nontrivial and cannot be done via naive methods.

III. Main Results

A. Transfer learning model

To learn latent L, we consider a scenario of transfer learning illustrated in Fig. 3. where one aims to transfer the knowledge on L extracted from M past episodes with the mean rewards $(\boldsymbol{\theta}_m)_{m\in[M]}$ satisfying $\boldsymbol{\theta}_m \in \Phi(L)$ to a new episode $M+1$ with mean rewards $\boldsymbol{\theta} \in \Phi(L)$. For simplicity, we assume that each episode has the same length T.

Let $L_m := \max_{i\neq j\in[K]} \frac{|\theta_m(i)-\theta_m(j)|}{\|x(i)-x(j)\|}$ denote the tightest Lipschitz constant of episode m. For the transfer learning framework, we let $L = \max_{m\in[M]} L_m$ be the maximum value of Lipschitz constant in episode M and the $(M+1)$-th episode shares the same L, i.e., L is the smallest Lipschitz constant that explains all the episode up to M. We make the following assumption on L_m's:

Assumption 1 (Learnability)**.** *At least α-portion of the previous M episodes have their L_m close to L with certain margin $\varepsilon_\alpha > 0$. Formally, there exist $\alpha > 0$ and $\varepsilon_\alpha > 0$ such that*

$$|\{m \in [M] : L_m \geq L - \varepsilon_\alpha\}| \geq \alpha M \ .$$

The parameters α and ε_α in Assumption 1 quantify the difficulty of estimating L on the new episode tightly, where larger α and smaller ε_α imply sharper concentration of L_m's around L and thus easier setting. Recalling that smaller L implies smaller regret, we aim to estimate the smallest possible L. Notice that $\max(\hat{L}_m)$ can be easily manipulated with any large L_m. Thus, through the assumption on tail distribution of L_m's which allows a robust estimation of L, it is possible to securely recover the Lipschitz constant. Our analysis under Assumption 1 can be easily applied to the case assuming a plausible family of prior distribution of bandit parameter over structure $\Phi(L)$. We provide a guideline for the choice of the parameters α and ε_α controlling this tail distribution in Section IV.

Additionally, the difficulty of estimating L also depends on the sampling scheme in the prior tasks. Hence, we make the following assumption:

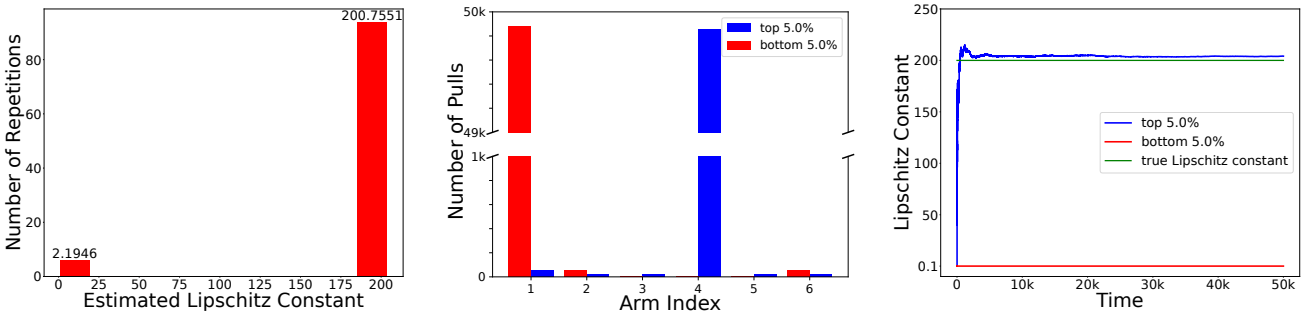
Assumption 2 (Minimal exploration)**.** *For each episode m, every arm $i \in [K]$ is pulled at least $\tau > 0$.*

In Assumption 2, a small value of τ implies a high risk of having insufficient samples for estimation of L. Thus, when T is sufficiently large, the conditions on τ are naturally verified.

B. Extracting Lipschitz constant

To analyze with latent L, we take a two-step approach. First, we estimate L from extracting structural information in previous M episodes, and then use this value to run $\pi(L)$. Let $\hat{\boldsymbol{\theta}}_m$ be the estimated mean rewards in episode m and $\hat{L}_m := \max_{i\neq j\in[K]} \frac{|\hat{\theta}_m(i)-\hat{\theta}_m(j)|}{\|x(i)-x(j)\|}$ be the estimated Lipschitz

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3441669



(a) Histogram of $\hat{L}_{T=50k}$ (b) Arm pulls $n_T(i)$'s at $T = 50k$ (c) Estimated $\hat{L}_t$ over time

Fig. 2. Behavior of $\pi(\hat{L}_t)$ for given $\boldsymbol{\theta}$ and $\boldsymbol{x}$ shown in Fig. 1(a).

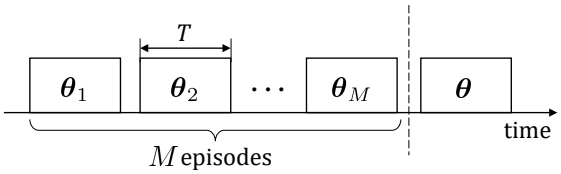


Fig. 3. Transfer learning framework; the knowledge of latent Lipschitz constant is learned from previous M episodes and transferred to a new $M+1$ episode.

constant in episode m. For an efficient estimate of L, we introduce two hyperparameters $\beta \in (0, \alpha)$ and $\varepsilon_\beta > \varepsilon_\alpha$ to set our estimator as an upper confidence bound on L:

$$\hat{L}_\beta := \ell_\beta + \varepsilon_\beta \ . \tag{8}$$

Here $\ell_\beta := \lceil \beta M \rceil\text{-}\max_{m\in[M]} \hat{L}_m$, where k-max denotes the operator taking the k-th largest element. For better understanding, we present a table of description of hyperparameters in Table I. Note that the margin ε_β is imposed to reduce the risk of underestimating L since structured bandit algorithms with the underestimated Lipschitz constant can induce almost linear regret as shown in Section II-C. While algorithms with any overestimated Lipschitz constant may generate additional logarithmic regret to the fundamental limit, it does not make such catastrophic failures. Let $\Delta_{\boldsymbol{x}} := \min_{i\neq j} d(i,j) > 0$. By running $\pi(\hat{L}_\beta)$, we get the following performance guarantee:

Theorem 3. *Suppose Assumptions 1 and 2 hold for $\alpha, \varepsilon_\alpha$ and τ. Let $\beta \in (0, \alpha)$ and $\tau \geq \frac{4}{\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta - \varepsilon_\alpha)^2}\left(\ln(2K) + \frac{1}{\min\{\beta, \alpha-\beta\}}\right)$. If $\tau M \geq 4Z\ln(T)$ with $Z = \frac{1}{\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2\min\{\beta,\alpha-\beta\}}$, for any $\boldsymbol{\theta} \in \Phi(L)$, Algorithm $\pi(\hat{L}_\beta)$ with $\lambda > 0$ has*

$$\limsup_{T\to\infty} \frac{R_T^\pi(\boldsymbol{\theta})}{\log T} \leq (1+\lambda)\ C\big(\boldsymbol{\theta}, L + 2\varepsilon_\beta - \varepsilon_\alpha\big) \ .$$

The proof of Theorem 3 is provided in Section III-E. The upper bound of $C(\boldsymbol{\theta}, L')$ in Property 1 is continuous L' for $L' \geq L$ and $\boldsymbol{\theta} \in \Phi(L)$. Under mild additional assumptions, we also have a continuity of $C(\boldsymbol{\theta}, L')$ in L' for $L' \in [L, L+\varepsilon)$ and $\boldsymbol{\theta} \in \Phi(L)$. Then, we provide a formal description of continuity of $C(\boldsymbol{\theta}, L)$ in L, implying a near optimality of the regret upper bound of $\pi(\hat{L}_\beta)$, and more details in Appendix C:

Theorem 4. *For given $\boldsymbol{\theta} \in \Phi(L)$, the optimal value of* (3a)–(3b)*, $C(\boldsymbol{\theta}, L')$ is locally upper-continuous at L (to be specified in the proof), provided that the optimal arm $x^*(\boldsymbol{\theta})$ and the solution to problem* (3a)–(3b) *in Theorem 1 are unique.*

Hence, Theorem 3 implies that when τ and M are sufficiently large, i.e., we have rich experiences with prior tasks, the algorithm $\pi(\hat{L}_\beta)$ closely achieves the fundamental limit of oracle performance with known L in advance. One can interpret τM as the sample complexity per arm for some probably approximately correct (PAC) learning of L. The sample complexity per arm required in Theorem 3 can be written as follows:

$$\tau M = \Omega\left(\frac{1}{\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta - \varepsilon_\alpha)^2\alpha}\log T\right) \ , \tag{9}$$

which matches with the information-theoretic lower bound for the PAC learning obtained later (Section III-D). In Section III-C below, we discuss on how to select hyperparamters β and ε_β of $\hat{L}_\beta$.

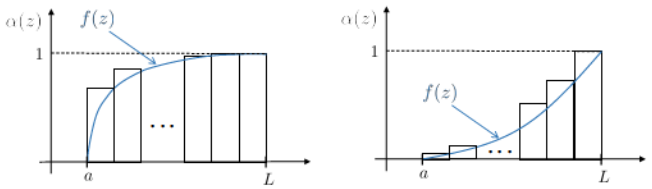
C. Hyperparameter choice for $\hat{L}_\beta$

General guideline. Theorem 3 includes an intrinsic trade-off for the choice of β, which appears in the requirement of $\tau M \geq \tilde{\Theta}(\max\{\frac{1}{\beta}, \frac{1}{\alpha-\beta}\})$ where $\tilde{\Theta}$ hides logarithmic factors. The value of β should not be too close to α nor 0. When β is too small (choosing near the top of $\{\hat{L}_m\}$), the estimate $\hat{L}_\beta$ can be too large and overshoot whereas when β is too large, the estimate falls below L and risks incurring linear regret. Furthermore, the requirement on τM becomes $\Omega(\frac{1}{\alpha})$. When multiple valid $(\alpha, \varepsilon_\alpha)$'s are known, one should pick α that is not too small. Our theorem also suggests that when a valid $(\alpha, \varepsilon_\alpha)$ is available, one should set $\varepsilon_\beta \approx 2\varepsilon_\alpha$, which prevents the requirement on τ from exploding, and $\beta \approx \alpha/2$, which prevents the requirement on τM from exploding. Such a choice implies that the Lipschitz constant $\hat{L}_\beta$ used by our algorithm is at most $3\varepsilon_\alpha$ away from L.

Exploiting distributional assumption. While Assumption 1 characterizes the difficulty of transfer learning, there are many pairs of $(\alpha, \varepsilon_\alpha)$ that satisfy the assumption. Since these pairs

TABLE I
DESCRIPTION OF HYPERPARAMETERS α, ε_α, β, AND ε_β.

Hyperparameter	Description
α	The proportion of episodes where L_m is greater than $L-\varepsilon_\alpha$ in all episodes M
ε_α	The margin for setting the degree of concentration of L_m to L
β	The proportion of episodes for the efficient estimation of Lipschitz constant on the new episode
ε_β	The margin for preventing catastrophic failures in the estimation of Lipschitz constant on the new episode



(a) $\gamma < 1$ (b) $\gamma > 1$

Fig. 4. Example plot of the cumulative histogram function $\alpha(z)$ and a lower approximation $f(z)$ defined in Assumption 3.

are a function of the past Lipschitz constants $\{L_m\}$ given arbitrarily from the environment, it is not easy to understand which pair would provide the strongest theoretical guarantee.

To get around this issue, we consider a parametric assumption on $\{L_m\}$ that characterizes the density of $\{L_m\}$ near the true Lipschitz constant L. Let $\xi_m := L - L_m, \forall m \in [M]$, be the *Lipschitz gap*. We define

$$\alpha(z) := \frac{|\{m \in [M] : \xi_m \le z\}|}{M} ,$$

which computes how much fraction of the past episodes have their Lipschitz constant L_m within distance z from L. Note that $(\alpha, \varepsilon_\alpha)$ with $\alpha = \alpha(\varepsilon_\alpha)$ satisfies Assumption 1. In other words, $\alpha(z)$ is a cumulative histogram of the Lipschitz gaps $\{\xi_m\}$. Examples of $\alpha(z)$ can be found in Fig. 4. We assume that $\alpha(z)$ is under-approximated by a polynomial function:

Assumption 3. *There exist $a \ge 0$ and $\gamma > 0$ such that*

$$\alpha(z) \ge f(z) := \left(\frac{z-a}{L-a}\right)^\gamma , \forall z \in [0, L] .$$

Examples of such functions can be found in Fig. 4. When chosen tightly, the parameter a characterizes the smallest Lipschitz gap $\min_m \xi_m$, and the parameter γ characterizes how densely Lipschitz gaps are located around $\min_m \xi_m$. A smaller γ means a higher such density, which should help transfer learning. Such a parameterization resembles the Tsybakov noise condition [32], [33] and β-regularity in infinite-armed bandits [34].

Theorem 5. *Suppose Assumptions 2 and 3 hold for some $a \in (0, L)$ and $\gamma > 0$. There exists $\beta_0 = \Theta\left(\left(\frac{\min\{L,\varepsilon_\beta\}-a}{L-a}\right)^\gamma\right)$, $\tau_0 = \Theta\left(\frac{\ln(K)}{\Delta_*^2} \cdot \frac{(2+\gamma)^2}{(\min\{L,\varepsilon_\beta\}-a)^2}\right)$, and*

$$Z = \Theta\left(\max\left\{\frac{(2+\gamma)^2}{(\min\{L,\varepsilon_\beta\}-a)^2}\cdot\frac{1}{\beta_0}, \frac{1}{(\min\{L,\varepsilon_\beta\})^2}\cdot\frac{1}{\beta}\right\}\right)$$

where $f(z)$ in Assumption 3 such that, for all $\beta \le \beta_0$, $\tau \ge \tau_0$, and $\tau M \ge Z \ln(T)$, Algorithm $\pi(\hat{L}_\beta)$ with $\gamma > 0$ has

$$\limsup_{T\to\infty} \frac{R_T^\pi(\boldsymbol{\theta})}{\log T} \le (1+\gamma) C(\boldsymbol{\theta}, L + 2\varepsilon_\beta) \quad \forall \boldsymbol{\theta} \in \Phi(L) .$$

For simplicity, assume $a = 0$ and $\varepsilon_\beta < L$. Theorem 5 shows that the requirement on M is $\tilde{\Omega}\left(\left(\frac{1}{(\frac{\varepsilon_\beta}{L})^\gamma}\right)\frac{1}{\beta}\right)$. This means that as γ goes to 0, the transfer learning becomes easy in the sense that the requirement on M becomes close to $\tilde{\Omega}(1/\beta)$. On the other hand, as γ goes to infinity, the requirement of the number of past episodes M increases exponentially. This is not too surprising since even with $\gamma = 2$ the density of Lipschitz gaps around 0 becomes very low. This aligns well with the intuition that the density of $\{L_m\}$ around L, which is encoded by γ, should determine the difficulty of transfer learning.

γ-tuning algorithm. In addition to general guideline, we present a heuristic Algorithm 2, denoted by $\hat{L}(\gamma)$, that can automatically tune the appropriate hyperparameter β and ε_β. Since the range of β and ε_β depends on α and ε_α, we first assume $\alpha(z)$ using the density of $\{\hat{L}_m\}$. Based on distributional assumption $\alpha(z)$, we can estimate γ which encodes approximated function in Assumption 3. In other words, Algorithm 2 estimates an optimal parameter γ of polynomial function $f(z)$, which approximates to the cumulative histogram of the Lipschitz gap $\{\xi_m\}$ from previous episodes as shown in Fig. 4. We use the maximum value of $\{\hat{L}_k\}_{k\in[m]}$ obtained from $\pi(\infty)$, which means no continuity structure, as pre-estimated Lipschitz constant to calculate the Lipschitz gap in advance. Then, the algorithm computes β based on Theorem 5. The detailed procedure of the algorithm is as follows: there are two initialization constant a and $c \in (0, 1)$, which is a L_m that satisfies $L - L_m = 0$ and which is a scale factor for calculating ε_β proportion to true Lipschitz constant, respectively. The hyperparmeter q and the pre-estimated Lipschitz L' are calculated via $\{\hat{L}_k\}_{k\in[m]}$. If $m = 1$, meaning that there is no history of episode, β is 1 and ε_β is $\hat{L}_m$ multiplied by c. For $m \ge 2$, to obtain a temporary pre-estimated L', the maximum value of the cumulative Lipschitz constant is multiplied by some constant $q \ge 1$. To calculate q, we split k-fold and measure $q_j = \max_{m\in[M]} \hat{L}_m / \max_{m\in[M]\setminus \text{fold}_j} \hat{L}_m$, and take average over j. As k increases, q also increases, so that the algorithm can be robust to the suddenly large $\hat{L}_m$. However, if q is too large, the pre-estimated Lipschitz constant L becomes large, $k = 5$ is set as the default. Notice that $L + 2\varepsilon_\beta$ is within a constant factor of L. Therefore, we set $\varepsilon_\beta = c \cdot L'$ for some $c \in (0, 1)$. We use a least square for finding the best value of γ for fitting $f(z)$ utilizing the cumulative estimated Lipschitz constant. Note that the β_0 of Theorem 5 requires knowledge of $(a, L, \varepsilon_\beta)$ and $\beta \le \beta_0$. Based on this,

Algorithm 2 $\hat{L}(\gamma)$

Initialize the constant $a = 0$ and $c \in (0, 1)$
for $m = 1, 2, ...$ **do**
  if $m = 1$ **then**
    Set $\beta_m = 1$ and $\varepsilon_{\beta,m} = c \cdot \hat{L}_m$
  else
    Set the constant q using k-fold as follows:

$$q = \Sigma_j q_j / k \;\text{ where } q_j = \max_{k\in[m]} \hat{L}_k / \max_{k\in[m]/\text{fold}_j} \hat{L}_k \;\forall j \in [k];$$

    Compute pre-estimated Lipschitz constant $L' = q \cdot \max_{k\in[m]} \hat{L}_k$
    Calculate the Lipschitz gap $\xi_k := L' - \hat{L}_k, \forall k \in [m]$ and $\varepsilon_{\beta,m} = c \cdot \max_{k\in[m]} \hat{L}_k$
    Find γ using a least square:

$$\alpha(z) := \frac{|\{k \in [m] : \xi_k \le z\}|}{m}; \qquad \text{and} \qquad f(z) := \left(\frac{z-a}{L'-a}\right)^\gamma, \; \forall z \in [a, \hat{L}_m].c$$

    Set β_m as follows:

$$\beta_m = \begin{cases} \frac{\Sigma_{k\in[m-1]}\beta_k}{m-1}, & \text{if } \gamma > 1; \\ \left(\frac{\varepsilon_{\beta,m}-a}{L'-a}\right)^\gamma, & \text{otherwise.} \end{cases}$$

  end if
  Calculate $\hat{L}_m(\gamma) = \lceil \beta_m M \rceil\text{-}\max_{m\in[M]} \; \hat{L}_m + \varepsilon_{\beta,m}$
end for

we calculate $\beta_m = (\frac{\varepsilon_{\beta,m}-a}{L'-a})^\gamma$. However, as γ increases, β has a very small value, e.g., $\beta < 0.1$. This can occur with arbitrarily large $\hat{L}$, resulting in γ being greater than 1, which is the case in Fig. 4(b). Therefore, if $\gamma > 1$, the average of the β of the previous episode is used as β_m. Lastly, $\hat{L}(\gamma)$ is calculated using β_m and $\varepsilon_{\beta,m}$, which are obtained through Algorithm 2. In summary, Algorithm 2 provides a heuristic approach to automatically tune the hyperparameters β and ϵ_β by estimating the optimal parameter γ for appproximating the cumulate histogram of the Lipschitz gap. The experiment results using hyperparameters calculated through the γ-tuning algorithm can be found in Section IV-B.

D. Lower bound of sample complexity τM

We study a fundamental limit of sample complexity in estimating L from M prior tasks. For the concentration of $\hat{L}$ to L, we define an event $\mathcal{L}_\varepsilon := \{\hat{L} \notin [L, L + \varepsilon]\}$. Let τ_m denote the number of playing the most under-sampled arm in episode m. Then, Assumption 2 can be equivalently written as $\min_{m\in[M]} \tau_m \ge \tau$. For given $\alpha > 0, \varepsilon_\alpha > 0, \tau > 0$, and $\varepsilon > \varepsilon_\alpha$, we say an estimator $\hat{L}$ is uniformly good for $(\alpha, \varepsilon_\alpha, \tau, \varepsilon)$ if the estimator $\hat{L}$ verifies the following for any L, $(\boldsymbol{\theta}_m \in \Phi(L))_{m\in[M]}$ and $(\tau_m)_{m\in[M]}$ satisfying Assumptions 1 and 2:

$$\mathbb{P}[\mathcal{L}_\varepsilon] \le O(T^{-c}) \quad \exists c > 0 \; . \tag{10}$$

The concentration in (10) can be interpreted as a minimal condition to conclude the desired regret upper bound in Theorem 3 recalling that $\pi(L')$ with wrong $L' \notin [L, L+\varepsilon]$ can generate linear regret. Using a change-of-measure argument based on Lemma 2 (Lemma 19 in [35]) in Appendix E, under the supposition that estimator $\hat{L}$ from M prior tasks is uniformly good for $(\alpha, \varepsilon_\alpha, \tau, \varepsilon)$, one can prove the following theorem:

Theorem 6. *Suppose that an estimator $\hat{L}$ is uniformly good for $\alpha > 0, \varepsilon_\alpha > 0, \tau > 0$, and $\varepsilon > \varepsilon_\alpha$. Then, we must have*

$$\Delta_{\boldsymbol{x}}^2 (\varepsilon - \varepsilon_\alpha)^2 \alpha \tau M = \Omega(\log T) \; . \tag{11}$$

The proof of Theorem 6 is provided in Appendix E. Notice that when we set $\varepsilon = 2\varepsilon_\beta - \varepsilon_\alpha$, the concentration of $\hat{L}$ in (10) becomes the one required to conclude the regret bound of $\pi(\hat{L})$ in Theorem 3. Hence, Theorem 6 provides a lower bound on $K\tau M$ to obtain the desired concentration of $\hat{L}$ as (9) which asymptotically matches with the lower bound on τM in Theorem 3 with $\beta = c\alpha$ for any positive constant $c < 1$ since the choice of β implies $\min\{\beta, \alpha - \beta\} = \min\{c\alpha, (1-c)\alpha\}$.

E. Proof of Theorem 3

As defined above, we set $\Delta_{\boldsymbol{x}} := \min_{i\neq j} d(i, j)$ and $\hat{L}_\beta := l_\beta + \varepsilon_\beta$ in (8). The key analysis of Theorem 3 lies in deriving the concentration of $\hat{L}_\beta$ to L:

Lemma 1. *Take Assumptions 1 and 2 with α, ε_α, and τ. Let $\beta \in (0, \alpha)$ and $\varepsilon_\beta > \varepsilon_\alpha$. If $\tau \ge \frac{4}{\Delta_{\boldsymbol{x}}^2 (\varepsilon_\beta - \varepsilon_\alpha)^2} \big(\ln(2K) + \frac{1}{\min\{\beta, \alpha-\beta\}} \big)$,*

$$\begin{aligned} &\mathbb{P}[L > \ell_\beta + \varepsilon_\beta] + \mathbb{P}[\ell_\beta > L + \varepsilon_\beta - \varepsilon_\alpha] \\ &\le 2 \exp\left(-\frac{\Delta_{\boldsymbol{x}}^2 (\varepsilon_\beta - \varepsilon_\alpha)^2}{4} \min\{\beta, \alpha - \beta\} \tau M \right) \; . \end{aligned} \tag{12}$$

We leave the detailed proof of Lemma 1 in Appendix A. To avoid linear regret, we desire to control $\mathbb{P}[L > \ell_\beta + \varepsilon_\beta] + \mathbb{P}[\ell_\beta >$

$L+\varepsilon_\beta-\varepsilon_\alpha] \le O(1/T)$ by bounding the RHS from Lemma 1. Then, it suffices to control

$$\exp\left(-\frac{\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2}{4}\min\{\beta,\alpha-\beta\}\tau M\right) \le \frac{1}{T}\ .$$

One can show that a sufficient condition on τM to satisfy this inequality is $\tau M \ge 4Z\ln(T)$ where $Z = \frac{1}{\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2\min\{\beta,\alpha-\beta\}}$. Assuming the condition on τM holds true, the following regret decomposition concludes the proof:

$$\begin{aligned}
R_T^\pi(\boldsymbol{\theta}) &\le \mathbb{E}_\pi\Bigg[\sum_{i\in[K]}(\theta_*-\theta(i))n_T(i)\mid L\le\hat{L}_\beta\le L+2\varepsilon_\beta-\varepsilon_\alpha\Bigg] \\
&\quad+\tilde{\Delta}_{\boldsymbol{\theta}}T\left(\mathbb{P}[\hat{L}_\beta<L]+\mathbb{P}[\hat{L}_\beta>L+2\varepsilon_\beta-\varepsilon_\alpha]\right) \\
&\le (1+\lambda)C(\boldsymbol{\theta},L+2\varepsilon_\beta-\varepsilon_\alpha)\log T+o(\log T) \\
&\quad+\tilde{\Delta}_{\boldsymbol{\theta}}T\cdot 2\exp\left(-\frac{\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2}{4}\min\{\beta,\alpha-\beta\}\tau M\right) \\
&\le (1+\lambda)C(\boldsymbol{\theta},L+2\varepsilon_\beta-\varepsilon_\alpha)\log T+o(\log T)\ ,
\end{aligned}$$

where we use Theorem 2 and $\tilde{\Delta}_{\boldsymbol{\theta}} := \max_{i\in[K]}\theta_*-\theta(i)\le 1$ for the second and last inequality, respectively.

F. Embedding distance learning

Heretofore, we assumed that embedding $\boldsymbol{x}$ is a fixed vector, but sometimes it is unknown in practice. In this subsection, we aim to learn a latent embedding distance $(d(i,j))_{i,j\in[K]}$, c.f., estimating latent L with known embedding $\boldsymbol{x}$. To avoid confusion, along with arbitrary fixed L, we set $L=1$ for embedding distance learning and omit the notation L. Then, we reuse the notation $\boldsymbol{d}$ for embedding distance matrix. Let $d_m(i,j)=|\theta_m(i)-\theta_m(j)|$ be the embedding distance of episode $m\in[M]$ for each pair of arms $i,j\in[K]$ and $\boldsymbol{d}_m := (d_m(i,j))\in\mathbb{R}^{K\times K}$ be the embedding distance matrix in episode m. The hat operator denotes an estimated value. We assume a α'-portion concentration of $d(i,j) := \max_{m\in[M]}d_m(i,j)$ for all $i,j\in[K]$ to ensure learning of all pairs of arms, and make the following assumption analog to Assumption 1:

Assumption 1′. *At least α'-portion of the previous M episodes have their $d_m(i,j)$ close to $d(i,j)$ with certain margin $\varepsilon_{\alpha'}>0$. Formally, for any $i,j\in[K]$, there exist $\alpha'>0$ and $\varepsilon_{\alpha'}>0$ such that*

$$|\{m\in[M]: d_m(i,j)\ge d(i,j)-\varepsilon_{\alpha'}\}|\ge\alpha' M\ .$$

Based on Assumption 1′, we suggest an estimator $(\hat{d}(i,j))_{\beta'}$ using two hyperparameters β' and $\varepsilon_{\beta'}$ for each pair of arms individually, which is the similar way to $\hat{L}_\beta$:

$$(\hat{d}(i,j))_{\beta'} := \lceil\beta' M\rceil\text{-}\max_{m\in[M]}\ \hat{d}_m(i,j)+\varepsilon_{\beta'}\ . \tag{13}$$

Our estimator $(\hat{d}(i,j))_{\beta'}$ adds margin $\varepsilon_{\beta'}$ to top-β' portion of M episodes to avoid incurring linear regret. Then, one can apply our algorithm to estimate a latent embedding distance independently for each pair (i,j) by using the estimator $(\hat{d}(i,j))_{\beta'}$. As discussed on Theorem 2, we get the following regret upper bound which is asymptotically close to fundamental limit for the true embedding distance matrix $\boldsymbol{d}$ by running $\pi(\hat{\boldsymbol{d}}_{\beta'})$ where $\hat{\boldsymbol{d}}_{\beta'} := ((\hat{d}(i,j))_{\beta'})\in\mathbb{R}^{K\times K}$:

Theorem 3′. *Suppose Assumptions 1 and 2 hold for $\alpha',\varepsilon'_\alpha$ and τ. Let $\beta'\in(0,\alpha')$ and $\tau\ge\frac{4}{(\varepsilon_{\beta'}-\varepsilon_{\alpha'})^2}\left(\ln(4)+\frac{1}{\min\{\beta',\alpha'-\beta'\}}\right)$. If $\tau M\ge Z'\ln(KT)$ with $Z'=\frac{1}{(\varepsilon_{\beta'}-\varepsilon_{\alpha'})^2\min\{\beta',\alpha'-\beta'\}}$, for any $\boldsymbol{\theta}\in\Phi(\boldsymbol{d})$, the algorithm $\pi(\hat{\boldsymbol{d}}_{\beta'})$ with $\lambda>0$ has*

$$\limsup_{T\to\infty}\frac{R_T^\pi(\boldsymbol{\theta})}{\log T}\le(1+\lambda)\ C\big(\boldsymbol{\theta},\boldsymbol{d}+(2\varepsilon_{\beta'}-\varepsilon_{\alpha'})I_K\big)\ ,$$

where I_K is $K\times K$ identity matrix.

The following statement shows a continuity of $C(\boldsymbol{\theta},\boldsymbol{d})$ in $\boldsymbol{d}$:

Theorem 4′. *For given bandit structure $\Phi(L,\boldsymbol{d}')$ with fixed $\boldsymbol{\theta}$ and $\boldsymbol{x}$, the optimal value of* (3a)–(3b)*, $\boldsymbol{d}'\to C(\boldsymbol{\theta},\boldsymbol{d}')$ is locally continuous at $\boldsymbol{d}'$, provided that the optimal arm $x^*(\boldsymbol{\theta})$, the solution to problem* (3a)–(3b) *in Theorem 1, is unique.*

Henceforward, we present a concentration bound for verifying the minimal condition of desired upper bound in Theorem 3′. Notice that a certain (i',j') pair of arms quantifies the accuracy of estimates of i' and j'. To well estimate the expected reward for all arms, it is sufficient to track the distance between one of any fixed arm $i\in[K]$ and all others. Accordingly (and similarity to the notion of uniformly good estimator $\hat{L}$), for given $\alpha'>0$, $\varepsilon_{\alpha'}$, $\tau>0$ and $\varepsilon>\varepsilon_{\alpha'}$, we say an estimator matrix $\hat{\boldsymbol{d}}$ is uniformly good for $(\alpha',\varepsilon'_\alpha,\tau,\varepsilon)$, if the estimator matrix $\hat{\boldsymbol{d}}$ verifies the following for any $\boldsymbol{d}$, $(\theta_m\in\Phi(\boldsymbol{d}))_{m\in[M]}$, and $(\tau_m)_{m\in[M]}$ satisfying Assumption 1′ and 2:

$$\begin{aligned}
&\mathbb{P}\left[\bigcup_{j\in[K]\setminus\{i\}}\left(\{\hat{d}(i,j)<d(i,j)\}\cup\{\hat{d}(i,j)>d(i,j)+\varepsilon\}\right)\right] \\
&\le o\left(\frac{\log T}{T}\right)\quad\forall i\in[K]\ .
\end{aligned}\tag{14}$$

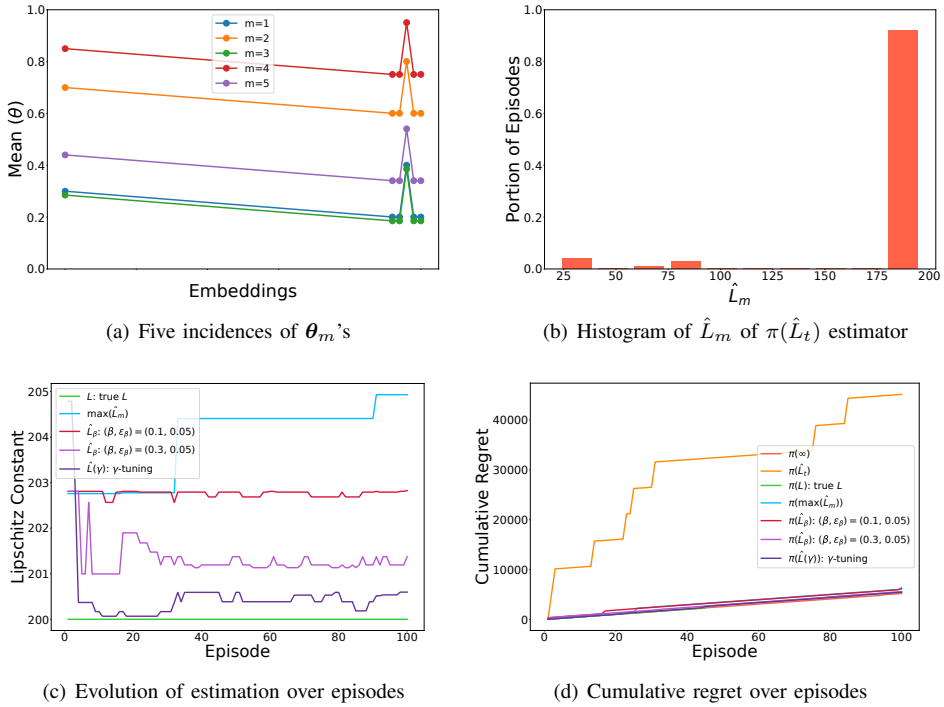
We define the event for any fixed $i\in[K]$, $\mathcal{X} = \{\cap_{j\in[K]\setminus\{i\}}(d(i,j)\le(\hat{d}(i,j))_{\beta'}\le d(i,j)+2\varepsilon_{\beta'}-\varepsilon'_\alpha)\}$. Then, we derive the concentration of $(\hat{d}(i,j))_{\beta'}$ to $d(i,j)$ using the following lemma:

Lemma 1′. *Take Assumptions 1′ and 2 with α', $\varepsilon_{\alpha'}$, and τ. Let $\beta'\in(0,\alpha')$. If $\tau\ge\frac{4}{(\varepsilon_{\beta'}-\varepsilon_{\alpha'})^2}\left(\ln(4)+\frac{1}{\min\{\beta',\alpha'-\beta'\}}\right)$, for any $i,j\in[K]$,*

$$\begin{aligned}
&\mathbb{P}\Big[d(i,j)>(\hat{d}(i,j))_{\beta'}\Big]+\mathbb{P}\Big[(\hat{d}(i,j))_{\beta'}>d(i,j)+2\varepsilon_{\beta'}-\varepsilon_{\alpha'}\Big] \\
&\le 2\exp\left(-\frac{(\varepsilon_{\beta'}-\varepsilon_{\alpha'})^2}{4}\min\{\beta',\alpha'-\beta'\}\tau M\right)\ .
\end{aligned}\tag{15}$$

Using Lemma 1′ to bound $\mathbb{P}[\mathcal{X}^{\mathcal{C}}]$,

$$\begin{aligned}
\mathbb{P}[\mathcal{X}^{\mathcal{C}}] &\le \mathbb{P}\left[\bigcup_{j\in[K]\setminus\{i\}}\Big\{d(i,j)>(\hat{d}(i,j))_{\beta'}\Big\}\right] \\
&\quad+\mathbb{P}\left[\bigcup_{j\in[K]\setminus\{i\}}\Big\{(\hat{d}(i,j))_{\beta'}>d(i,j)+2\varepsilon_{\beta'}-\varepsilon_{\alpha'}\Big\}\right]
\end{aligned}$$



(a) Five incidences of $\boldsymbol{\theta}_m$'s

(b) Histogram of $\hat{L}_m$ of $\pi(\hat{L}_t)$ estimator

(c) Evolution of estimation over episodes

(d) Cumulative regret over episodes

Fig. 5. Analysis of experienced episodes for the importance of the conservative estimation.

$$\begin{aligned}
&\overset{(a)}{\leq} \sum_{j\in[K]\setminus\{i\}} \mathbb{P}\left[d(i,j) > (\hat{d}(i,j))_{\beta'}\right] \\
&\quad + \sum_{j\in[K]\setminus\{i\}} \mathbb{P}\left[(\hat{d}(i,j))_{\beta'} > d(i,j) + 2\varepsilon_{\beta'} - \varepsilon_{\alpha'}\right] \\
&\leq K\exp\left(-\frac{(\varepsilon_{\beta'}-\varepsilon_{\alpha'})^2}{4}\min\{\beta',\alpha'-\beta'\}\tau M\right) ,
\end{aligned}$$

where for (a), we use Boole's inequality. Then, we control

$$K\exp\left(-\frac{(\varepsilon_{\beta'}-\varepsilon'_{\alpha})^2}{4}\min\{\beta',\ \alpha'-\beta'\}\tau M\right) \leq \frac{1}{T} ,$$

in order to prevent linear regret. A sufficient condition on τM is $\tau M \geq Z'\ln(KT)$ where $Z' = \frac{4}{(\varepsilon_{\beta'}-\varepsilon_{\alpha'})^2\min\{\beta',\alpha'-\beta'\}}$. Therefore, the sample complexity for all arms which can be interpreted as $K\tau M$ in embedding distance learning required

$$K\tau M = \Omega\left(\frac{K}{(\varepsilon_{\beta'}-\varepsilon_{\alpha'})^2\alpha'}\log(KT)\right) . \tag{16}$$

For the desired concentration of each pair of arms, we drive a lower bound on $K\tau M$,

Theorem 6′. *Suppose that an estimator matrix $\hat{\boldsymbol{d}}$ is uniformly good for $\alpha > 0, \varepsilon_{\alpha'} > 0, \tau > 0$, and $\varepsilon > \varepsilon_{\alpha'}$. Then, we must have*

$$(\varepsilon-\varepsilon_{\alpha'})^2\alpha'\tau M = \Omega(\log T) . \tag{17}$$

Setting $\varepsilon = 2\varepsilon_{\beta'} - \varepsilon_{\alpha'}$ makes the concentration of each of $d(i,j)$ which is required to conclude the desired regret upper bound in Theorem 3′. It is confirmed that the sample complexity bound of order in embedding distance learning, which transfers for all (i,j) pairs of arms, is the similar asymptotic growth rate to Lipschitz constant learning case (9), which transfers only L. Notice that the logarithmic order of growth can be neglected by the linear scale of K in case of having large K. The detailed proofs of embedding distance learning are provided in Appendix F.

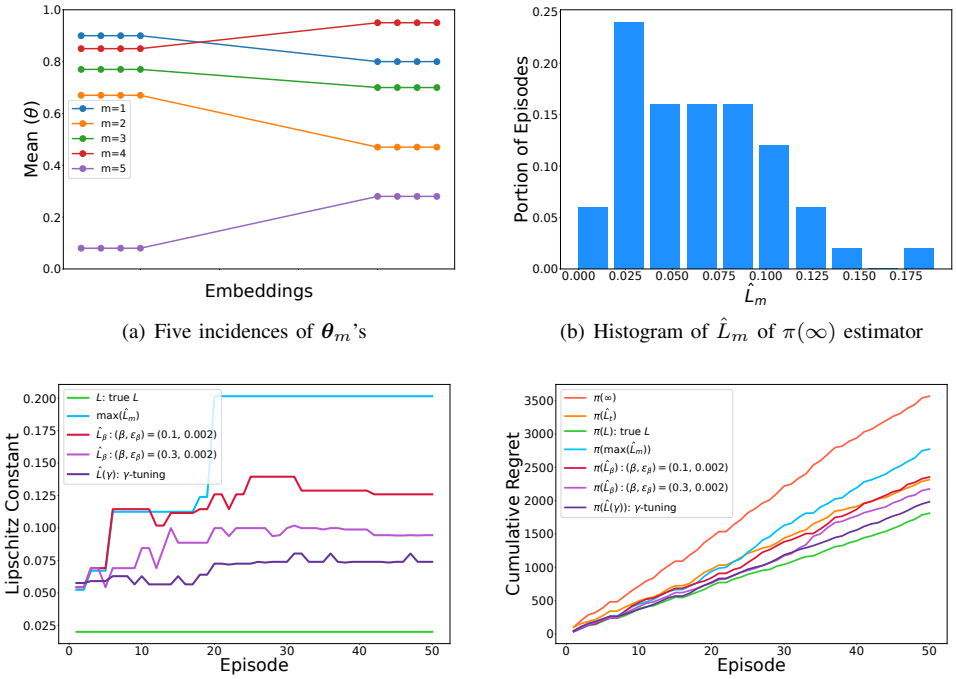
IV. NUMERICAL EVALUATION

In this section, we present experiments to compare the performance of estimators with an asymptotically optimal algorithm. We validate the proposed framework for estimating latent Lipschitz constant L and latent embedding distance $(d(i,j))_{i,j\in[K]}$ in various settings. We first observe two episodic cases to scrutinize the estimation of latent L: a case where it is hard to estimate true L with one arm hidden, as shown in Fig. 1(a) (Section IV-A), and a case where it is relatively easy to estimate L through the specific process of generating instances (Section IV-B). For analyzing the estimation of latent embedding distance, real-world dataset [9] in an application of rate adaptation is used (Section IV-C). The simulation is built on SMPyBandits package for MAB algorithms [36].*

A. Importance of the conservative estimation

Setup for episodic framework. We consider a scenario of transfer learning, where one episode corresponds to the setting that one arm is hidden as in Fig. 1(a), i.e., $\boldsymbol{\theta} = (0.1, 0.0005, 0.0005, 0.2005, 0.0005, 0.0005)$ and true Lipschitz constant L is 200. Based on $\boldsymbol{\theta}$, $\boldsymbol{\theta}_m \in \Phi(L)$ is independently generated by adding a constant, which is randomly selected in $[0, 0.7995]$, to all arms. Fig. 5(a) shows

*The detailed source code for this project is available on Github: https://github.com/hyejin-s/transfer-learning-bandits.git.



(a) Five incidences of $\boldsymbol{\theta}_m$'s

(b) Histogram of $\hat{L}_m$ of $\pi(\infty)$ estimator

(c) Evolution of estimation over episodes

(d) Cumulative regret over episodes

Fig. 6. Analysis of experienced episodes for the benefit from exploiting the continuity structure.

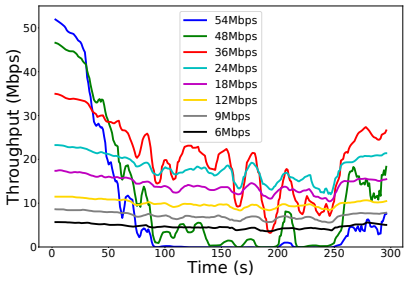
five incidences of $\boldsymbol{\theta}_m$'s. For each episode $m \in [M = 100]$, we set time horizon $T = 5 \times 10^4$ and multiply 5 to the estimation term in algorithm π for more accurate Lipschitz constant estimation and transfer learning.

Cumulative regret. Fig. 5(c) and Fig. 5(d) show evolution of estimation of Lipschitz constant and cumulative regret over episodes, respectively, using various estimators on L —$\hat{L}_t$'s taking the Lipschitz constant which is estimated at each time step $t \in [T]$, $\max(\hat{L}_m)$'s taking the maximum of previous estimated $\hat{L}_m$, two $\hat{L}_\beta$'s with $(\beta, \varepsilon_\beta) \in \{(0.1, 0.05), (0.3, 0.05)\}$, and $\hat{L}(\gamma)$'s with γ-tuning of L. Fig. 5(b) presents the histogram of empirical $\hat{L}_m$ of $\pi(\hat{L}_t)$ estimator. Compared with $\hat{L}_\beta$ and $\hat{L}(\gamma)$ estimators, which transfers structural knowledge using history of previous episodes $\{\hat{L}_m\}_{m=1}^{M-1}$, in the case of $\pi(\hat{L}_t)$ updating L at each time step t, it sometimes shows an unstably estimated Lipschitz constant in Fig. 5(b). As discussed in Section II-C, the hidden optimal arm is often undersampled, indicating that the estimated Lipschitz constant is smaller than the true Lipschitz constant 200 and causing linear regret. As a result, $\pi(\hat{L}_t)$ has a large cumulative regret. Even if the optimal arm is hidden, all other estimators except for $\pi(\hat{L}_t)$ estimate L close to 200 as shown in Fig. 5(c) and plot the curves of similar cumulative regret as shown in Fig. 5(d). This shows the risk of estimating L at each time step, implying the importance of conservative estimation in transfer learning scenarios.

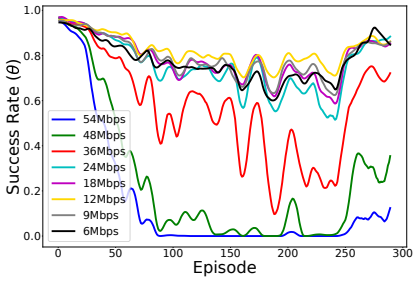
B. Benefit from exploiting the continuity structure

In this subsection, we demonstrate the importance of utilizing the continuity structure by comparison of various hyperparameter estimators.

Setup for episodic framework. For numerical simulations, we configure an environmental setting to take advantage of the structure for a clear performance comparison of algorithms. We assume that there are C clusters in which c_k arms with the same mean form a cluster. Notice that the number of arms in the bandit instance is $C \times c_k$. However, if there is the knowledge of structure among arms arms in the cluster is restricted by the continuity. Therefore, it can be considered as C arms in the bandit instance. Let the distance between the arms in a cluster be constant ρ. The constant ρ should be small enough to utilize the structural information. However, if ρ is too small than 0.1, the estimated Lipschitz constant $L = |\theta(i) - \theta(j)|/\rho$ can be large by an erroneously predicted mean of arms. In addition, the distance between clusters y, which is greater than ρ, is set within the range that structure benefits. We consider Lipschitz structure $\Phi(L)$ with $L = 0.02$, the length of episodes $M = 50$, and time horizon $T = 10^5$. We set the time horizon T for sufficient exploration in order to estimate L more accurately. For each episode $m \in [M = 50]$, $\boldsymbol{\theta}_m \in \Phi(L)$ is independently generated by the following procedure: starting with cluster 1, $\theta_1 \in [0.05, 0.95]$ for selected uniformly at random, for $i = 2, 3, ..., C$, select θ_i uniformly at random from $[\theta_{i-1} - L \cdot d(i-1, i), \theta_{i-1} + L \cdot d(i-1, i)] \cap [0.05, 0.95]$. The arms in a cluster have the same mean. We choose $C = 2$ clusters that is sufficient to show the results, although we can make several clusters. We place $c_k = 4$ arms in each cluster, which have the same mean with a distance of $\rho = 1$ from each other. We set the distance of each cluster $y = 10$. Fig. 6(a) shows five incidences of $\boldsymbol{\theta}_m$'s generated from the above procedure. Fig. 6(b) presents the histogram of empirical $\hat{L}_m$. For every estimator (stated below), we use the same sequence of $\hat{L}_m$'s



(a) Throughput $\boldsymbol{\mu}$ over time in [9]



(b) Success probability $\boldsymbol{\theta}$ over episodes

Fig. 7. Instantaneous throughput and success probability of *eight* rates.

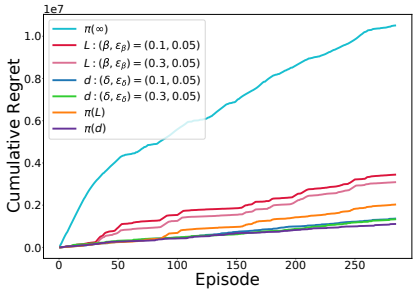


Fig. 8. Cumulative regret over episodes for estimators L and $\boldsymbol{d}$.

generated by $\pi(\infty)$ that uses no continuity structures.

Stable estimator $\hat{L}_\beta$. Fig. 6(c) illustrates the comparison between the accuracy of *five* estimators on L —$\hat{L}_t$'s taking the Lipschitz constant which is estimated at each time step $t \in [T]$, two $\hat{L}_\beta$'s with $(\beta, \varepsilon_\beta) \in \{(0.1, 0.002), (0.3, 0.002)\}$, $\max(\hat{L}_m)$'s taking the maximum of previous estimated $\hat{L}_m$, and $\hat{L}(\gamma)$'s with γ-tuning of L. For estimator $\hat{L}(\gamma)$, we set initial value $a = 0$ and $c = 0.1$. As expected, the most conservative estimation of $\max(\hat{L}_m)$ has monotonically increasing estimation of Lipschitz constant as the past episodes piling up. Theoretically, $\hat{L}_{\max}$ can explode up to $1/\Delta_{\boldsymbol{x}} = 1$ in a finite number of episodes with positive probability. Indeed, $\max(\hat{L}_m)$ is easily manipulated by a single episode with arbitrarily large $\hat{L}_m$, e.g., $0.175 \leq \hat{L}_m \leq 0.2$ in Fig. 6(b). Therefore, it is important to appropriately select β to exclude the right tail of histogram of $\hat{L}_m$ and to avoid the aggressive choice that lead to linear regret. The choice of $\beta = 0.1$ prevents some large estimated Lipschitz constant of $\hat{L}_m$, but also shows a difference from the true Lipschitz value. In this case, a larger β is more helpful to reduce regret. As expected, the hyperparameter choice of γ-tuning algorithm leads us to obtain an accurate and stable estimation of L, yielding the lower regret. Notice that in γ-tuning estimator $\hat{L}(\gamma)$, when an estimated Lipschitz constant value deviates from the distribution, $\gamma > 1$ is used to exclude the outlier by using the mean of β in previous episodes.

Cumulative regret. Fig. 6(d) plots the curves of cumulative regret over time under various estimators. The regret of five estimators on L take on some value between $\pi(\infty)$ which means no continuity structures and $\pi(L)$ which uses the full knowledge of the Lipschitz continuity. As discussed in our theoretical analysis, we observe that the more accurate the estimator leads the greater reduction in regret. Notice that $\pi(\max(\hat{L}_m))$, which uses the most conservative choice of L, gets closer to $\pi(\infty)$. Employing small value of β provides much conservative estimation of L, but it can be too conservative to exploit the Lipschitz constant. This gives insight into the *importance* of transfer the knowledge of L via delivering appropriate hyperparameters (β, ε).

C. Latent optimal embedding

Recall that in the previous subsection (Section IV-B), we consider a fixed embedding setting. However, it is often the case that the embedding is latent in practice. This motivates us to consider *embedding distance learning* in Section III-F, where we assume that embedding $\boldsymbol{x}$ is unknown. In particular, in this subsection, we consider the problem of rate adaptation (RA) in which non-stationary radio channels can be differentiated into tasks as the channel changes discretely over time and can capture the availability of embedding learning. We start by briefly recapitulate as to how the RA problem can be reduced to multi-armed bandit problem, as in [9].

Rate adaptation. Rate adaptation (RA) aims to maximize a throughput of packets given wireless link conditions. Let $\boldsymbol{r} = (r(1), ..., r(K))$ denote the set of rates and $\boldsymbol{\theta} = (\theta(1), ..., \theta(K))$ denote a probability of success transmission from here. For some $i \in [K]$, we assume that a rate $r(i)$ with which each packet is desired to be transmitted is first chosen. This packet is then transmitted with a probability of success transmission $\theta(i)$. For each round t, in particular, the success probability is modeled by a binary random variable $X(i,t)$ which is drawn i.i.d. from the Bernoulli distribution with mean $\theta(i) \in [0,1]$. The throughput $\boldsymbol{\mu}$, representing the amount of data transferred per unit hour, is then defined as follows:

$$\mu(i) = r(i) \cdot \theta(i) \; i \in [K] \; .$$

From this, it can be seen that the rates correspond to arms in MAB problems. In general, notice that a reward in MAB is given directly from observations $X(i,t)$. On the contrary, RA aims to maximize the throughput $\boldsymbol{\mu}$ rather than the success transmission probability $\boldsymbol{\theta}$ itself. For simplicity, when a rate is chosen for a certain packet, we assume that this packet transmission lasts during one slot. All packets are assumed to have unit sizes, implying that a packet with rate r_i is then transmitted over a period of $1/r_i$. Let $\theta_* := \max_{i \in [K]} \theta(i)$, $\mathcal{K}_*(\boldsymbol{\theta}) := \{i \in [K] : \theta(i) = \theta_*\}$ and r_* denote the best success probability, the set of best arms, and the corresponding rate to θ_*, respectively. From the assumptions, the regret that we aim to minimize can be rewritten as follows:

$$R_T^\pi(\boldsymbol{\theta}) := \sum_{i \in [K]} \left(r_*\theta_* - r(i)\theta(i)\right) \mathbb{E}_\pi[s_T(i)] \; ,$$

where $s_T(i)$ denotes the number of decision for choosing rate $r(i)$ for each packet up to time T. We assume the continuity of θ for the RA problem.* Accordingly, as in

*The continuity of throughput $\boldsymbol{\mu}$ was considered as well, but the gain for structure did not appear significantly.

Theorem 1, we can write the asymptotic lower bound of regret as $\liminf_{T\to\infty} R_T^\pi(\boldsymbol{\theta})/\log T \geq C(\boldsymbol{\theta}, L)$ for uniformly good algorithm. Here $C(\boldsymbol{\theta}, L)$ indicates the solution of the following minimization problem:

$$\min_{\boldsymbol{\eta}\succcurlyeq 0} \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \big(r_*\theta_* - r(i)\theta(i)\big)\eta(i) \tag{18}$$

$$\text{s.t.} \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \mathrm{KL}(\theta(i)\|\lambda^j(i;\boldsymbol{\theta},L))\eta(i) \geq 1, \forall j \notin \mathcal{K}_*(\boldsymbol{\theta}) \ , \tag{19}$$

where

$$\lambda^j(i;\boldsymbol{\theta},L) := \begin{cases} \frac{\mu_*}{r(i)}, & \text{if } i=j; \\ \max\{\theta(i), \frac{\mu_*}{r(i)} - L\cdot d(i,j)\}, & \text{otherwise.} \end{cases}$$

Setup for episodic framework. We note that non-stationary test-bed trace dataset in [9] is used in our simulation, demonstrating instantaneous throughput over time of *eight* rates (6Mbps, 9Mbps, 12Mbps, 18Mbps, 24Mbps, 36Mbps, 48Mbps, 54Mbps) in Fig. 7(a). We first define a single task to reproduce the episodic simulation in wireless link. We set m episodes in a sequential manner over $5s$ sliding window, as if the algorithm selects transmission rate based on estimated success probability every $1ms$. The throughput corresponding to these episodes is calculated as the average value of that time. Notice that the probability of success transmission $\boldsymbol{\theta}$ affects the agent's decision of choosing rates (among eight arms). Fig. 7(b) plots the curves of success probability over episodes.

Cumulative regret in rate adaptation. We compare (i) the case where the embedding is arbitrarily fixed (estimation of L) and (ii) the case where embedding distance is learned (estimation of $\boldsymbol{d}$) with real-data. For the first case, we assume fixed embedding as $\boldsymbol{x} = [6, 9, 12, 18, 24, 36, 48, 54]$ which is proportional to rate. We compare $\pi(\infty)$, (i) $\pi(\hat{L}_\beta)$ with $(\beta, \varepsilon_\beta) \in \{(0.1, 0.05), (0.3, 0.05)\}$, (ii) $\pi(\hat{\boldsymbol{d}}_{\beta'})$ with $(\beta', \varepsilon_{\beta'}) \in \{(0.1, 0.05), (0.3, 0.05)\}$, and each of true values L and $\boldsymbol{d}$ to see if we can leverage the Lipschitz structure. For each episode, we set time horizons $T = 5\times 10^3$. We utilize the structural information from the sequence generated by $\pi(\infty)$, which is no continuity structure algorithm for estimator $\hat{L}_\beta$ and $\hat{\boldsymbol{d}}_{\beta'}$. As observed in Fig. 8, the case of finding the optimal embedding distance outperforms the case of arbitrary fixed embedding $\boldsymbol{x}$. In addition, if we select the parameters well, it shows almost optimal regret in each case. Therefore, even if embedding distance learning may require more complexity (16), it shows better performance than when the embedding is set arbitrarily.

V. Conclusion

We have investigated the role of *transfer learning* with incomplete knowledge of Lipschitz continuity. Our main contribution lies in our estimator $\hat{L}_\beta$, its information-theoretic optimality, and regret analysis when using $\hat{L}_\beta$ for future tasks, that is shown to be close to the one with known Lipschitz structures. We have reported useful insights on transfer learning with latent Lipschitz constants, as well as demonstrating the superiority of the proposed framework via numerical evaluations.

In a further direction, by analysis of estimating the optimal embedding distance with $(\beta', \varepsilon_{\beta'})$, we confirmed that our proposed algorithm is applicable even when the embedding $\boldsymbol{x}$ is not available. One exciting future research is to extend our setup to fully *adaptive* sequential transfer setting where we require the learner to adjust ε_β automatically to a given precision (e.g., within a constant factor of L). Moreover, while our current focus is on finite-armed bandits with Lipschitz continuity, our approach can be extended to a broader class of continuous-armed bandits and continuous reward functions. Such extensions could make our approach applicable to wider range of real-world problems.

Appendix A
Proof of Lemma 1

We begin with a concentration analysis on $\hat{L}_m$ to L_m. Notice that Assumption 2 guarantees that every arm is played at least τ times in each episode. Hence, using Hoeffding's inequality, it follows that for each $m \in [M]$,

$$\mathbb{P}\left[|\theta_m(i) - \hat{\theta}_m(i)| \geq \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2\tau\right) \ , \tag{20}$$

which implies

$$\begin{aligned} &\mathbb{P}\left[|L_m - \hat{L}_m| \geq \varepsilon\right] \\ &\leq \mathbb{P}\left[\exists i \in [K] : |\theta_m(i) - \hat{\theta}_m(i)| \geq \frac{\varepsilon\Delta_{\boldsymbol{x}}}{2}\right] \\ &\leq 2K\exp\left(-\frac{\varepsilon^2\Delta_{\boldsymbol{x}}^2\tau}{2}\right) \ . \end{aligned} \tag{21}$$

For the following proofs, without loss of generality, we assume that $L_1 \geq \cdots \geq L_M$ and define $\xi_m = L - L_m$ as the Lipschitz gaps.

Bound of $\mathbb{P}[L > \ell_\beta + \varepsilon_\beta]$. Let us first show the upper confidence bound derivation. Let $[x]_+ = \max\{0, x\}$. Let S_k be the k-th smallest element of the set S and $m_\beta := \lceil \beta M \rceil$. Define $\mathcal{S} := \{S \subseteq [M] : |S| = M - m_\beta + 1\}$. Then,

$$\begin{aligned} &\mathbb{P}\left(\underset{m\in[M]}{m_\beta\text{-}\max}\, \hat{L}_m \leq L - \varepsilon_\beta\right) \\ &\leq \sum_{S\in\mathcal{S}} \prod_{k=1}^{M-m_\beta+1} \mathbb{P}(\hat{L}_{S_k} \leq L - \varepsilon_\beta) \\ &= \sum_{S\in\mathcal{S}} \prod_{k=1}^{M-m_\beta+1} \mathbb{P}\left(\hat{L}_{S_k} - L_{S_k} \leq -(\varepsilon_\beta - \xi_{S_k})\right) \\ &\leq \sum_{S\in\mathcal{S}} \prod_{k=1}^{M-m_\beta+1} 1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[\varepsilon_\beta - \xi_{S_k}]_+^2) \ . \end{aligned}$$

Notice that $|S| = M - m_\beta + 1$ and S_k is the k-th smallest element of the set S. For all $k \in [M - m_\beta + 1]$, we have $S_k \leq M - (M - m_\beta + 1) + k = m_\beta + k - 1$. Recalling that

ξ_m is non-decreasing by definition, we have $\xi_{S_k} \le \xi_{m_\beta+k-1}$. Then, the RHS above is bounded by

$$\begin{aligned}
&\sum_{S\in\mathcal{S}} \prod_{k=1}^{M-m_\beta+1} 1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[\varepsilon_\beta - \xi_{m_\beta+k-1}]_+^2) \\
&= \sum_{S\in\mathcal{S}} \prod_{m=m_\beta}^{M} 1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[\varepsilon_\beta - \xi_{m}]_+^2) \\
&= \binom{M}{M-m_\beta+1} \prod_{m=m_\beta}^{M} 1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[\varepsilon_\beta - \xi_{m}]_+^2) \,.
\end{aligned}$$

Note that

$$\begin{aligned}
&\prod_{m=m_\beta}^{M} 1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[\varepsilon_\beta - \xi_{m}]_+^2) \\
&\overset{(a)}{\le} \min_s \left(1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[\varepsilon_\beta - s]_+^2)\right)^{|\{m\ge m_\beta:\, \xi_m \le s\}|} \\
&\overset{(b)}{\le} \left(1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta - \varepsilon_\alpha)^2)\right)^{\alpha M - m_\beta + 1} \\
&\le \exp\left(\left(\ln(2K) - \frac{\tau}{2}\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2\right)\cdot(\alpha-\beta)M\right) \,,
\end{aligned}$$

where (a) is for the tightest bound using ξ_m with monotonically increasing property; and (b) is by Assumption 1. Using the fact that $\binom{M}{M-m_\beta+1} \le 2^M \le \exp(M)$, we finally get:

$$\begin{aligned}
&\mathbb{P}\left(\min_{m\in[M]} m_\beta\text{-}\max \hat{L}_m \le L - \varepsilon_\beta\right) \\
&\le \exp\left(\left(1 + (\alpha-\beta)\ln(2K) - \frac{\tau}{2}\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2(\alpha-\beta)\right)M\right) \\
&\le \exp\left(-\frac{\tau}{4}\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2(\alpha-\beta)M\right) \,,
\end{aligned}$$

where the last inequality is from the assumption $\tau \ge \frac{4}{\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2}\left(\ln(2K) + \frac{1}{\alpha-\beta}\right)$.

Bound of $\mathbb{P}[\ell_\beta > L + \varepsilon_\beta - \varepsilon_\alpha]$. The proof is analogous to the proof of bound of $\mathbb{P}[L > \ell_\beta + \varepsilon_\beta]$. From the definition of ℓ_β in (8), it follows that

$$\begin{aligned}
&\mathbb{P}\left(\min_{m\in[M]} m_\beta\text{-}\max \hat{L}_m \ge L + \varepsilon_\beta - \varepsilon_\alpha\right) \\
&\le \mathbb{P}\big(\exists S\subseteq[M]: |S| = m_\beta \ \forall S_k \in S, \hat{L}_{S_k} - (\varepsilon_\beta - \varepsilon_\alpha) \ge L\big) \\
&= \mathbb{P}\big(\exists S\subseteq[M]: \\
&\quad |S| = m_\beta \ \forall S_k \in S, \hat{L}_{S_k} - L_{S_k} \ge -\big((\varepsilon_\alpha - \varepsilon_\beta) - \xi_{S_k}\big)\big) \\
&\overset{(c)}{\le} \sum_{S\in\mathcal{S}} \prod_{k=1}^{m_\beta} 1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[(\varepsilon_\alpha-\varepsilon_\beta) - \xi_{S_k}]_+^2) \\
&\le \binom{M}{m_\beta} \prod_{m=M-m_\beta+1}^{M} 1 \wedge 2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[(\varepsilon_\alpha-\varepsilon_\beta) - \xi_{m}]_+^2) \\
&\le \binom{M}{m_\beta}\left(2K\exp\left(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2\right)\right)^{m_\beta} \\
&\overset{(d)}{\le} \exp\left(\left(\frac{1}{\beta} + \ln(2K)\right)\beta M - \frac{\tau}{2}\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2\beta M\right) \\
&\overset{(e)}{\le} \exp\left(-\frac{\tau}{4}\Delta_{\boldsymbol{x}}^2(\varepsilon_\beta-\varepsilon_\alpha)^2\beta M\right),
\end{aligned}$$

where (c) is from $S_k \le M - m_\beta + k$ for all $k \in [m_\beta]$; (d) obtained by the fact that $\binom{M}{m_\beta} \le 2^M \le \exp(M)$; and (e) follows from our assumption on τ. □

Appendix B
Proof of Theorem 2

We note that our analysis can be concluded with any other algorithm than Algorithm 1 if it achieves the asymptotic optimality provided in Theorem 2. Algorithm 1 is a simplification of DEL algorithm in [8] originally designed for Markov decision process (MDP). It is straightforward to correspond the bandit problem with Lipschitz continuity to an MDP of single state and K actions with Lipschitz continuity. Hence, the proof will be concluded by Theorem 4 in [8] once we correspond the following linear programming to the one in (3):

$$\min_{\boldsymbol{\eta}\succcurlyeq 0} \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \big(\theta_* - \theta(i)\big)\eta(i) \tag{22a}$$

$$\text{s.t.} \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \mathrm{KL}(\theta(i)\|\lambda(i))\eta(i) \ge 1 \ \forall \boldsymbol{\lambda} \in \Psi(\boldsymbol{\theta}, L) \,, \tag{22b}$$

where $\Psi(\boldsymbol{\theta}, L) \subset \Phi(L)$ is the set of confusing parameters to $\boldsymbol{\theta}$ defined as

$$\begin{aligned}
\Psi(\boldsymbol{\theta}, L) := \{\boldsymbol{\lambda} \in \Phi(L) : \mathcal{K}_*(\boldsymbol{\theta}) \cap \mathcal{K}_*(\boldsymbol{\lambda}) = \emptyset \ \text{ and} \\
\theta(i) = \lambda(i) \ \forall i \in \mathcal{K}_*(\boldsymbol{\theta})\} \,.
\end{aligned}$$

The correspondence is provided by Theorem 1 in [2]. This completes the proof. □

Appendix C
Proof of Theorem 4

To prove the local upper-continuity of $C(\boldsymbol{\theta}, L')$ in (3a)–(3b) for $L' \in [L, L+\delta]$, we consider a modified linear programming (LP) of which the optimal value is denoted by $C'(\boldsymbol{\theta}, L')$:

$$\min_{\boldsymbol{\eta}\succcurlyeq 0} \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \big(\theta_* - \theta(i)\big)\eta(i) \tag{23a}$$

$$\text{s.t.} \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \mathrm{KL}(\theta(i)\|\lambda^j(i;\boldsymbol{\theta}, L))\eta(i) \ge 1, \forall j \notin \mathcal{K}_*(\boldsymbol{\theta}) \tag{23b}$$

$$\sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \big(\theta_* - \theta(i)\big)\eta(i) \le C(\theta, \infty) \,. \tag{23c}$$

The only modification is the additional constraint (23c).

We will first show the equivalence between $C(\boldsymbol{\theta}, L')$ and $C'(\boldsymbol{\theta}, L')$, and then the local upper-continuity of $C'(\boldsymbol{\theta}, L')$, which completes the proof. Recalling that $C(\boldsymbol{\theta}, L')$ is increasing in $L'' \ge L'$, it follows that $C(\boldsymbol{\theta}, L') \le C(\boldsymbol{\theta}, \infty)$, i.e., the solution $\boldsymbol{\eta}^*$ of $C(\boldsymbol{\theta}, L')$ should verify $\sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \big(\theta_* - \theta(i)\big)\eta^*(i) \le C(\theta, \infty)$. This shows the equivalence, i.e., $C(\boldsymbol{\theta}, L') = C'(\boldsymbol{\theta}, L')$

To show the local upper-continuity of $C'(\boldsymbol{\theta}, L')$ for $L' \in [L, L+\delta]$, we will use Berge's maximum theorem. To do

so, we check the conditions for Berge's maximum theorem: (i) the objective function (23a) is continuous with respect to $L' \in [L, L+\delta]$ and $\boldsymbol{\eta} \geq 0$ and (ii) the constraint set (23b)-(23c) is non-empty and compact. The continuity of the objective function is straightforward as it is a linear function in $\boldsymbol{\eta}$. To show that the constraint set is non-empty, we construct $\boldsymbol{\eta}'$ such that $\eta'(i) = \frac{1}{\mathrm{KL}(\theta(i)\|\theta^*)}$ for $i \notin \mathcal{K}_*(\boldsymbol{\theta})$. Noting that $\boldsymbol{\eta}'$ is the solution for $C(\boldsymbol{\theta}, \infty)$, we have $\sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \big(\theta_* - \theta(i)\big)\eta'(i) = C(\theta, \infty)$ and $\forall j \notin \mathcal{K}_*(\boldsymbol{\theta})$,

$$
\begin{aligned}
1 &= \mathrm{KL}(\theta(i)\|\theta_*)\eta'(j) && (24)\\
&= \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \mathrm{KL}(\theta(i)\|\lambda^j(i;\boldsymbol{\theta},\infty))\eta'(i) && (25)\\
&\leq \sum_{i\notin\mathcal{K}_*(\boldsymbol{\theta})} \mathrm{KL}(\theta(i)\|\lambda^j(i;\boldsymbol{\theta},L'))\eta'(i)\ , && (26)
\end{aligned}
$$

where the inequality holds as $\lambda^j(i;\boldsymbol{\theta},L') := \max\{\theta(i), \theta_* - L'\cdot d(i,j)\} \geq \theta(i)$ is non-decreasing in L' and thus $\mathrm{KL}(\theta(i)\|\lambda^j(i;\boldsymbol{\theta},L'))$ is also non-decreasing. This implies $\boldsymbol{\eta}'$ verifies the constraints (23b)-(23c) and thus the constraint set is non-empty. To show the compactness, we observe that the constraint (23c) upper-bounds a weighted sum of $\boldsymbol{\eta}$ where the weights are strictly positive for $i \notin \mathcal{K}_*(\boldsymbol{\theta})$. Noting $\boldsymbol{\eta} \geq 0$, the constraint set is compact. This completes the proof of Theorem 4.

□

Appendix D
Proof of Theorem 5

Throughout the proof, we assume that $\varepsilon_\beta \leq L$ without loss of generality; one can verify that the same proof goes through when we replace ε_β with $\min\{L, \varepsilon_\beta\}$. To avoid clutter, let us use ε instead of ε_β.

Following the proof of Lemma 1, we have

$$
\begin{aligned}
&\mathbb{P}\left(\underset{m\in[M]}{m_\beta\text{-}\max}\, \hat{L}_m \leq L-\varepsilon\right)\\
&\leq \mathcal{B}\cdot\min_s\left(2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[\varepsilon-s]_+^2)\right)^{|\{m\geq m_\beta:\ \xi_m\leq s\}|}\\
&= \mathcal{B}\cdot\min_v\left(2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2[v]_+^2)\right)^{|\{m\geq m_\beta:\ \xi_m\leq \varepsilon-v\}|}\\
&\overset{(a)}{\leq} \mathcal{B}\cdot\min_{v\geq v_{\min}}\left(2K\exp(-\frac{\tau}{2}\Delta_{\boldsymbol{x}}^2 v^2)\right)^{|\{m\geq m_\beta:\ \xi_m\leq \varepsilon-v\}|}\\
&\overset{(b)}{\leq} \mathcal{B}\cdot\min_{v\geq v_{\min}}\exp\left(-\frac{\tau}{4}\Delta_{\boldsymbol{x}}^2 v^2\cdot|\{m\geq m_\beta:\ \xi_m\leq\varepsilon-v\}|\right)\\
&= \mathcal{B}\cdot\exp\left(-\frac{\tau}{4}\Delta_{\boldsymbol{x}}^2\cdot\max_{v\geq v_{\min}} v^2\cdot|\{m\geq m_\beta:\ \xi_m\leq\varepsilon-v\}|\right)\\
&\overset{(c)}{\leq} \mathcal{B}\cdot\exp\left(-\frac{\tau}{4}\Delta_{\boldsymbol{x}}^2\cdot\max_{v\in[v_{\min},\varepsilon-a]} v^2(M\cdot f(\varepsilon-v)-(m_\beta-1))\right)
\end{aligned}
$$

where

- $\mathcal{B} = \binom{M}{M-m_\beta+1}$ denotes the binomial coefficient.
- (a) is by introducing a free variable $v_{\min} \geq 0$ that we choose later.
- (b) is by assuming $\ln(2K) - \frac{\tau}{2}\Delta_{\boldsymbol{x}}^2 v^2 \leq -\frac{\tau}{4}\Delta_{\boldsymbol{x}}^2 v^2$, for all $v \geq v_{\min}$. Equivalently, we assume that

$$
\tau \geq \frac{4}{\Delta_{\boldsymbol{x}}^2 v_{\min}^2}\ln(2K)\ . \tag{27}
$$

- (c) is by $|\{m\geq m_\beta : \xi_m \leq \varepsilon - v\}| \geq M\cdot f(\varepsilon-v) - (m_\beta-1), \forall v\in[0,\varepsilon-a]$, and $\varepsilon \leq L$.

For the moment, let us focus on the optimization problem:

$$
\begin{aligned}
&\max_{v\in[v_{\min},\varepsilon-a]} v^2\left(M\cdot f(\varepsilon-v)-(m_\beta-1)\right)\\
&= \max_{v\in[v_{\min},\varepsilon-a]} v^2\left(M\left(\frac{\varepsilon-v-a}{L-a}\right)^\gamma-(m_\beta-1)\right)\\
&= \frac{M}{(L-a)^\gamma}\max_{v\in[v_{\min},\varepsilon-a]} v^2\left((\varepsilon-v-a)^\gamma-\frac{m_\beta-1}{M}(L-a)^\gamma\right)\\
&= \frac{M(\varepsilon-a)^{2+\gamma}}{(L-a)^\gamma}\\
&\quad\times\max_{v\in[v_{\min},\varepsilon-a]}\left(\frac{v}{\varepsilon-a}\right)^2\left(\left(1-\frac{v}{\varepsilon-a}\right)^\gamma-\frac{m_\beta-1}{M}\frac{(L-a)^\gamma}{(\varepsilon-a)^\gamma}\right)\\
&\overset{(a)}{=} \frac{M(\varepsilon-a)^{2+\gamma}}{(L-a)^\gamma}\max_{x\in[\frac{v_{\min}}{\varepsilon-a},1]} x^2\left((1-x)^\gamma-A\right)
\end{aligned}
$$

where (a) is by defining $A = \frac{m_\beta-1}{M}\frac{(L-a)^\gamma}{(\varepsilon-a)^\gamma}$.

Inspecting the objective function, the optimal solution x^* must satisfy $1-x^* \geq A^{1/\gamma}$. Unfortunately, the optimization problem does not have an explicit closed-form solution. One can find an integral approximation:

$$
\begin{aligned}
&\max_{x\in[\frac{v_{\min}}{\varepsilon-a},1]} x^2\left((1-x)^\gamma-A\right)\\
&\geq \frac{1}{1-A^{\frac{1}{\gamma}}-(\frac{v_{\min}}{\varepsilon-a})}\int_{x\in[\frac{v_{\min}}{\varepsilon-a},1-A^{\frac{1}{\gamma}}]} x^2\left((1-x)^\gamma-A\right)\mathrm{d}x
\end{aligned}
$$

which does have a closed form and seems to be a good approximation (numerically). This can be useful for deriving a tight confidence bound to use.

On the other hand, the equation above is hard to interpret, and we thus turn to restricting the regime of β because useful values for β are small in general. This way, we gain interpretability.

Let us assume

$$
A \leq \frac{1}{2e^2}\ , \tag{28}
$$

which becomes our requirement on β as we show later. We plugin $x = \frac{2}{2+\gamma}$ (motivation: this would be the solution of the optimization problem if A were 0). Note that this requires the following condition, which we assume hereafter:

$$
\frac{2}{2+\gamma} \geq \frac{v_{\min}}{\varepsilon-a} = \frac{1}{\varepsilon-a}\cdot\sqrt{\frac{4}{\tau\Delta_{\boldsymbol{x}}^2}\ln(2K)}\ . \tag{29}
$$

Then,

$$
\begin{aligned}
\max_{x\in[\frac{v_{\min}}{\varepsilon-a},1]} x^2\left((1-x)^\gamma-A\right) &\geq \left(\frac{2}{2+\gamma}\right)^2\left(\left(\frac{\gamma}{2+\gamma}\right)^\gamma-A\right)\\
&\overset{(a)}{\geq}\left(\frac{2}{2+\gamma}\right)^2\left(\frac{1}{e^2}-A\right)\\
&\geq\left(\frac{2}{2+\gamma}\right)^2\frac{1}{2e^2}
\end{aligned}
$$

where (a) is by $\left(\frac{\gamma}{2+\gamma}\right)^\gamma = \frac{1}{\left(1+\frac{2}{\gamma}\right)^\gamma} \geq \frac{1}{\left(\exp(2/\gamma)\right)^\gamma} = 1/e^2$.

Thus, altogether,

$$\begin{aligned}
&\mathbb{P}\left(\min_{m\in[M]}{}_{m_\beta}\text{-}\max \hat{L}_m \leq L-\varepsilon\right) \\
&\leq \binom{M}{M-m_\beta+1} \exp\left(-\frac{\tau\Delta_{\boldsymbol{x}}^2}{4}\cdot\frac{M(\varepsilon-a)^{2+\gamma}}{2e^2(L-a)^\gamma}\left(\frac{2}{2+\gamma}\right)^2\right) \\
&\overset{(a)}{\leq} \exp\left(M-\tau\Delta_{\boldsymbol{x}}^2\cdot\frac{M(\varepsilon-a)^{2+\gamma}}{2e^2(L-a)^\gamma}\frac{1}{(2+\gamma)^2}\right)
\end{aligned}$$

where (a) is from the fact $\binom{M}{M-m_\beta+1} \leq 2^M \leq \exp(M)$. Finally, we want to control the RHS above to be smaller than $1/T$.

Let us summarize the assumptions we have made:

$$\tau \geq \frac{4}{\Delta_{\boldsymbol{x}}^2 v_{\min}^2}\ln(2K) \tag{30}$$

$$\frac{2}{2+\gamma} \geq \frac{1}{\varepsilon-a}\cdot\sqrt{\frac{4}{\tau\Delta_{\boldsymbol{x}}^2}\ln(2K)} \tag{31}$$

$$\frac{1}{2e^2} \geq A = \frac{m_\beta-1}{M}\cdot\frac{(L-a)^\gamma}{(\varepsilon-a)^\gamma} \tag{32}$$

$$\exp\left(M-\tau\Delta_{\boldsymbol{x}}^2\cdot\frac{M(\varepsilon-a)^{2+\gamma}}{2e^2(L-a)^\gamma}\frac{1}{(2+\gamma)^2}\right) \leq \frac{4}{T}\ . \tag{33}$$

We take $v_{\min}^2 = 4\cdot\frac{(\varepsilon-a)^2}{(2+\gamma)^2}$ and merge (30) and (31):

$$\tau \geq \frac{(2+\gamma)^2}{(\varepsilon-a)^2}\cdot\frac{\ln(2K)}{\Delta_{\boldsymbol{x}}^2} =: \tau_0\ . \tag{34}$$

Recall that $m_\beta = \lceil\beta M\rceil$. One can verify that the following is a sufficient condition for (32):

$$\beta \leq \frac{1}{2e^2}\left(\frac{\varepsilon-a}{L-a}\right)^\gamma =: \beta_0\ .$$

Finally, using our combined condition above, the condition (33) is implied by the following:

$$\exp\left(-\tau\Delta_{\boldsymbol{x}}^2\cdot\frac{M(\varepsilon-a)^{2+\gamma}}{2e^2(L-a)^\gamma}\frac{1}{(2+\gamma)^2}\right) \leq 1/T\ . \tag{35}$$

For the lower confidence bound, from the proof of Theorem 3,

$$\mathbb{P}\left(\min_{m\in[M]}{}_{m_\beta}\text{-}\max \hat{L}_m \geq L+\varepsilon\right) \leq \exp\left(-\frac{\tau}{4}\Delta_{\boldsymbol{x}}^2\varepsilon^2\beta M\right) \tag{36}$$

which requires $\tau \geq \frac{4}{\Delta_{\boldsymbol{x}}^2\varepsilon^2}\ln(2K)$, but that is satisfied by $\tau \geq \tau_0$. We wish to control the equation above under $1/T$. For this, using $\tau \geq \tau_0$ and $\varepsilon^2 \geq (\varepsilon-a)^2$, the requirements (35) and (36) can be satisfied by the following:

$$\exp\left(-\Delta_{\boldsymbol{x}}^2\min\left\{\left(\frac{\varepsilon-a}{2+\gamma}\right)^2\frac{1}{2e^2}\left(\frac{\varepsilon-a}{L-a}\right)^\gamma, \frac{\varepsilon^2\beta}{4}\right\}\tau M\right) \leq 1/T\ .$$

Using the same argument as the proof of Theorem 3, we see that the above takes the form of $\exp(-Z^{-1}\tau M) \leq 1/T$ for some Z, and a sufficient condition for satisfying it is $\tau M \geq Z\ln(T)$. Altogether, we can conclude that there exists $Z = \Theta\left(\max\left\{\frac{1}{(\frac{\varepsilon-a}{2+\gamma})^2\beta_0}, \frac{1}{\varepsilon^2\beta}\right\}\right)$ such that when $\tau M \geq Z\ln(T)$, $\tau \geq \tau_0$, and $\beta \geq \beta_0$, we enjoy the stated regret bound. □

Appendix E
Proof of Theorem 6

Let $n_m(i)$ denote the number of playing arm $i \in [K]$ in episode $m \in [M]$ with slight abuse of notation. Let $\mathcal{F}$ be the sigma-field of observations in M episodes of length T, in which Assumption 2 holds, i.e., $n_m(i) \geq \tau$ for every episode $m \in [M]$ and arm $i \in [K]$. Then, the uniformly good estimator $\hat{L}$ verifies (10) for $(\boldsymbol{\theta}_m)_{m\in[M]}$ satisfying Assumption 1. Let $\mathbb{P}$ and $\mathbb{P}'$ be the probability measures on $\mathcal{F}$ w.r.t. $\mathcal{M}$ and $\mathcal{M}'$, respectively. Similarly, denote the expectations on $\mathcal{F}$ w.r.t. $\mathcal{M}$ and $\mathcal{M}'$ by $\mathbb{E}$ and $\mathbb{E}'$ respectively. We use a change-of-measure argument which compares two sequences of M parameters, denoted by $\mathcal{M} := (\boldsymbol{\theta}_m)_{m\in[M]}$ s.t. $\boldsymbol{\theta}_m \in \Phi(L)$ and $\mathcal{M}' := (\boldsymbol{\lambda}_m)_{m\in[M]}$ s.t. $\boldsymbol{\lambda}_m \in \Phi(L')$. We will construct $\mathcal{M}$ and $\mathcal{M}'$ to conclude the proof using the following lemma:

Lemma 2 (Lemma 19 in [35])**.** *For every event $\mathcal{E} \in \mathcal{F}$,*

$$\mathbb{E}[\mathcal{G}] \geq \mathrm{KL}(\mathbb{P}[\mathcal{E}]\|\mathbb{P}'[\mathcal{E}])\ , \tag{37}$$

where $\mathcal{G}$ is the log-likelihood ratio of $\mathcal{M}$ to $\mathcal{M}'$ defined as:

$$\mathcal{G} := \sum_{m\in[M]}\sum_{i\in[K]} n_m(i)\log\frac{\theta_m(i)}{\lambda_m(i)}\ .$$

Let i' be an arm such that $\min_{i\neq i'} d(i,i') = \Delta_{\boldsymbol{x}}$. For some $c \in (0,1)$, we consider $\mathcal{M}$ such that for each $m \in [M]$,

$$\theta_m(i) = \begin{cases} c+L\Delta_{\boldsymbol{x}} & \text{if } i = i' \\ c & \text{otherwise} \end{cases}\ .$$

Note that $L_m := \max_{i\neq j\in[K]}\frac{|\theta_m(i)-\theta_m(j)|}{d(i,j)} = L$ for all $m \in [M]$, i.e., $\boldsymbol{\theta}_m \in \Phi(L)$. In addition, $\mathcal{M}$ verifies Assumption 1 for L, ε_α and α. We now construct a perturbation $\mathcal{M}'$ which verifies Assumption 1 for $L' = L+\varepsilon$, ε_α and α. For each $m \in [\lceil\alpha M\rceil]$,

$$\lambda_m(i) = \begin{cases} \theta_m(i') + (\varepsilon-\varepsilon_\alpha)\Delta_{\boldsymbol{x}} & \text{if } i = i' \\ \theta_m(i) & \text{otherwise} \end{cases}\ , \tag{38}$$

which implies $L'_m := \max_{i\neq j}\frac{|\lambda(i)-\lambda(j)|}{d(i,j)} = L+(\varepsilon-\varepsilon_\alpha)$ due to the construction of $\boldsymbol{\theta}_m$: for $i \neq i'$,

$$\begin{aligned}
\lambda_m(i') - \lambda_m(i) &= (\theta_m(i') + (\varepsilon-\varepsilon_\alpha)\Delta_{\boldsymbol{x}}) - \theta_m(i) \\
&= \theta_m(i) + L\Delta_{\boldsymbol{x}} + (\varepsilon-\varepsilon_\alpha)\Delta_{\boldsymbol{x}} - \theta_m(i) \\
&= (L+\varepsilon-\varepsilon_\alpha)\Delta_{\boldsymbol{x}}\ .
\end{aligned}$$

For the rest, i.e., $m \in [M]\setminus[\lceil\alpha M\rceil]$, we set $\boldsymbol{\lambda}_m = \boldsymbol{\theta}_m$. Hence, $\mathcal{M}'$ verifies Assumption 1 for $L' := L+\varepsilon$, α and ε_α. Assume that for each $m \in [M]$, $n_m(i') = \tau$ and $n_m(i) \geq \tau$ if $i \neq i'$. This implies Assumption 2. Note that $\mathbb{E}[\mathcal{G}] = \sum_{m\in[M]}\sum_{i\in[K]}\mathbb{E}[n_m(i)]\mathrm{KL}(\theta_m(i)\|\lambda_m(i))$ thanks

to Markov property of bandit. With the construction of $\mathcal{M}$ and $\mathcal{M}'$, it follows that

$$\mathbb{E}[\mathcal{G}] \le \tau \left(\sum_{m=1}^{\lceil \alpha M \rceil} \mathrm{KL}(\theta_m(i') \| \lambda_m(i')) \right) \le \tau \left(\sum_{m=1}^{\lceil \alpha M \rceil} \frac{(\theta_m(i') - \lambda_m(i'))^2}{\lambda_m(i')(1-\lambda_m(i'))} \right) ,$$

where for the last inequality, we use the fact that $\mathrm{KL}(\theta\|\lambda) \le \mathcal{X}^2(\theta,\lambda) = \frac{(\theta-\lambda)^2}{\lambda(1-\lambda)}$, c.f,. Lemma 2.7 in [37]. It is not hard to select constant $c \in (0,1)$ such that

$$\tau \left(\sum_{m=1}^{\lceil \alpha M \rceil} \frac{(\theta_m(i') - \lambda_m(i'))^2}{\lambda_m(i')(1-\lambda_m(i'))} \right) = \Omega \left(\tau \left(\sum_{m=1}^{\lceil \alpha M \rceil} (\theta_m(i') - \lambda_m(i'))^2 \right) \right) .$$

With such choice of c and the construction of $\boldsymbol{\lambda}_m$ in (38), we obtain

$$\mathbb{E}[\mathcal{G}] = \Omega \left(\tau \alpha M \Delta_{\boldsymbol{x}}^2 (\varepsilon - \varepsilon_\alpha)^2 \right) . \tag{39}$$

To complete the proof using Lemma 2, we define an event $\mathcal{E} = \{\hat{L} \in [L, L+\varepsilon]\}$, and its complement $\mathcal{E}'$. Under Assumption 2 with $\tau > 0$ and the supposition that estimator $\hat{L}$ is uniformly good for $(\alpha, \varepsilon_\alpha, \tau, \varepsilon)$, we have $\mathbb{P}[\mathcal{E}'] = o\left(T^{-c}\right)$ and further

$$\mathbb{P}'[\mathcal{E}] = \mathbb{P}'[\hat{L} \in [L, L+\varepsilon]] = \mathbb{P}'[\hat{L} \in [L'-\varepsilon, L']] = o\left(T^{-c}\right) ,$$

where the last equality is from the construction of $\mathcal{M}'$ verifying Assumption 1 for $L' = L + \varepsilon$, α, and ε_α, i.e.,

$$\mathbb{P}'[\hat{L} \in [L'-\varepsilon, L']] \le \mathbb{P}'[\hat{L} < L'] + \mathbb{P}'[\hat{L} > L' + \varepsilon] = o\left(T^{-c}\right) .$$

From this, it follows that

$$\mathrm{KL}(\mathbb{P}[\mathcal{E}] \| \mathbb{P}'[\mathcal{E}]) = -\log\left(o\left(T^{-c}\right) \right) = O(\log T) . \tag{40}$$

Therefore, combining (39) and (40), Lemma 2 concludes the proof of Theorem 6. □

APPENDIX F
PROOF OF EMBEDDING DISTANCE LEARNING

A. Proof of Lemma 1′

The flow of this proof is analogous to the proof of Lemma 1. First, we analyze a concentration on $\hat{d}_m(i,j)$ to $d_m(i,j)$ for any $i, j \in [K]$. Notice that a confidence interval of $d_m(i,j)$ requires exploration only for arm i and j. Using Hoeffding's inequality with Assumption 2, for every $m \in [M]$, we have

$$\mathbb{P}\left[|\theta_m(i) - \hat{\theta}_m(i)| \ge \varepsilon \right] \le 2\exp\left(-2\varepsilon^2 \tau \right) , \tag{41}$$

which implies

$$\begin{aligned} &\mathbb{P}\left[|d_m(i,j) - \hat{d}_m(i,j)| \ge \varepsilon \right] \\ &\le \mathbb{P}\left[\{|\theta_m(i) - \hat{\theta}_m(i)| \ge \frac{\varepsilon}{2}\} \cap \{|\theta_m(j) - \hat{\theta}_m(j)| \ge \frac{\varepsilon}{2}\} \right] \\ &\le 4\exp\left(-\frac{\varepsilon^2 \tau}{2} \right) . \end{aligned} \tag{42}$$

For the following proofs, without loss of generality, we assume that $d_1(i,j) \ge \cdots \ge d_M(i,j)$ and define $\psi_m = d(i,j) - d_m(i,j)$ as the distance gaps.

Bound of $\mathbb{P}[d(i,j) > (\hat{d}(i,j))_{\beta'}]$. Recall that S_k be the k-th smallest element of the set S and $[x]_+ = \max\{0, x\}$. Define $m_{\beta'} := \lceil \beta' M \rceil$ and $\mathcal{S} := \{S \subseteq [M] : |S| = M - m_{\beta'} + 1\}$. Then,

$$\begin{aligned} &\mathbb{P}\left(\min_{m_{\beta'}\text{-}\max_{m\in[M]}} \hat{d}_m(i,j) \le d(i,j) - \varepsilon_{\beta'} \right) \\ &\le \sum_{S\in\mathcal{S}} \prod_{k=1}^{M-m_{\beta'}+1} \mathbb{P}(\hat{d}_{S_k}(i,j) \le d(i,j) - \varepsilon_{\beta'}) \\ &= \sum_{S\in\mathcal{S}} \prod_{k=1}^{M-m_{\beta'}+1} \mathbb{P}\left(\hat{d}_{S_k}(i,j) - d_{S_k}(i,j) \le -(\varepsilon_{\beta'} - \psi_{S_k}) \right) \\ &\le \sum_{S\in\mathcal{S}} \prod_{k=1}^{M-m_{\beta'}+1} 1 \wedge 4\exp(-\frac{\tau}{2}[\varepsilon_{\beta'} - \psi_{S(k)}]_+^2) . \end{aligned}$$

Using $S_k \le m_{\beta'} + k - 1 \ \forall k \in [M - m_{\beta'} + 1]$, we have $\psi_{S_k} \le \psi_{m_{\beta'}+k-1}$. Then,

$$\begin{aligned} &\sum_{S\in\mathcal{S}} \prod_{k=1}^{M-m_{\beta'}+1} 1 \wedge 4\exp(-\frac{\tau}{2}[\varepsilon_{\beta'} - \psi_{m_{\beta'}+k-1}]_+^2) \\ &= \sum_{S\in\mathcal{S}} \prod_{m=m_{\beta'}}^{M} 1 \wedge 4\exp(-\frac{\tau}{2}[\varepsilon_{\beta'} - \psi_m]_+^2) \\ &= \mathcal{B}' \prod_{m=m_{\beta'}}^{M} 1 \wedge 4\exp(-\frac{\tau}{2}[\varepsilon_{\beta'} - \psi_m]_+^2). \\ &\le \mathcal{B}' \min_s \left(1 \wedge 4\exp(-\frac{\tau}{2}[\varepsilon_{\beta'} - s]_+^2) \right)^{|\{m \ge m_{\beta'} :\ \psi_m \le s\}|} \\ &\overset{(a)}{\le} \mathcal{B}' \left(1 \wedge 4\exp(-\frac{\tau}{2}(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2) \right)^{\alpha' M - m_{\beta'} + 1} \\ &\overset{(b)}{\le} \exp\left((1 + (\alpha' - \beta')\ln(4)) M - \frac{\tau}{2}(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2 (\alpha' - \beta') M \right) \\ &\overset{(c)}{\le} \exp\left(-\frac{\tau}{4}(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2 (\alpha' - \beta') M \right) , \end{aligned}$$

where $\mathcal{B}'$ denotes $\binom{M}{M-m_{\beta'}+1}$; (a) is by Assumption 1′; (b) follows using the fact that $\binom{M}{m_{\beta'}} \le 2^M \le \exp(M)$; and (c) holds from the assumption $\tau \ge \frac{4}{(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2} \left(\ln(4) + \frac{1}{\alpha' - \beta'} \right)$.

Bound of $\mathbb{P}[(\hat{d}(i,j))_{\beta'} > d(i,j) + 2\varepsilon_{\beta'} - \varepsilon_{\alpha'}]$. From the definition of $(\hat{d}(i,j))_{\beta'} := \lceil \beta' M \rceil\text{-}\max_{m\in[M]} \hat{d}_m(i,j) + \varepsilon_{\beta'}$, it follows that

$$
\begin{aligned}
&\mathbb{P}\left(\underset{m\in[M]}{m_{\beta'}\text{-}\max} \hat{d}_m \geq d + \varepsilon_{\beta'} - \varepsilon_{\alpha'} \right) \\
&\leq \mathbb{P}(\exists H \subseteq [M] : \\
&\qquad |H| = m_{\beta'} \ \forall m \in H, \hat{d}_m(i,j) - (\varepsilon_{\beta'} - \varepsilon_{\alpha'}) \geq d(i,j)) \\
&\leq \binom{M}{m_{\beta'}} \left(4\exp\left(-\frac{\tau(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2}{2} \right) \right)^{m'_\beta} \\
&\overset{(d)}{\leq} \exp\left(\left(\frac{1}{\beta'} + \ln(4) \right) \beta' M - \frac{\tau}{2} (\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2 \beta' M \right) \\
&\overset{(e)}{\leq} \exp\left(-\frac{\tau}{4} (\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2 \beta' M \right),
\end{aligned}
$$

where (d) is by $\binom{M}{m_{\beta'}} \leq 2^M \leq \exp(M)$; and (e) holds from our assumption $\tau \geq \frac{4}{(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2} \left(\ln(4) + \frac{1}{\beta'} \right)$. □

B. Proof of Theorem 3′

Define the event $\mathcal{X} = \{ \cap_{j\in[K]\setminus\{i\}} (d(i,j) \leq (\hat{d}(i,j))_{\beta'} \leq d(i,j) + 2\varepsilon_{\beta'} - \varepsilon_{\alpha'}) \}$ to ensure that the concentration conditions of estimators are satisfied for all pairs of arms. Therefore, the regret bound of the algorithm π is derived as follows:

$$
R_T^\pi(\boldsymbol{\theta}) \leq \mathbb{E}_\pi \left[\sum_{i\in[K]} (\theta_* - \theta(i)) n_T(i) \mid \mathcal{X} \right] + \Delta_{\boldsymbol{\theta}} T \left(\mathbb{P}[\mathcal{X}^\mathcal{C}] \right) ,
$$

where $\Delta_{\boldsymbol{\theta}} := \max_{i\in[K]} \theta_* - \theta(i) \leq 1$.

Bound of $\mathbb{P}[\mathcal{X}^\mathcal{C}]$. Through the Lemma 1′, for any fixed $i \in [K]$,

$$
\begin{aligned}
\mathbb{P}[\mathcal{X}^\mathcal{C}] &\leq \mathbb{P}\left[\bigcup_{j\in[K]\setminus\{i\}} \left\{ d(i,j) > (\hat{d}(i,j))_{\beta'} \right\} \right] \\
&\quad + \mathbb{P}\left[\bigcup_{j\in[K]\setminus\{i\}} \left\{ (\hat{d}(i,j))_{\beta'} > d(i,j) + 2\varepsilon_{\beta'} - \varepsilon_{\alpha'} \right\} \right] \\
&\overset{(a)}{\leq} \sum_{j\in[K]\setminus\{i\}} \Big(\mathbb{P}\left[d(i,j) > (\hat{d}(i,j))_{\beta'} \right] \\
&\qquad + \mathbb{P}\left[(\hat{d}(i,j))_{\beta'} > d(i,j) + 2\varepsilon_{\beta'} - \varepsilon_{\alpha'} \right] \Big) \\
&\leq K \exp\left(-\frac{(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2}{4} \min\{\beta', \alpha' - \beta'\} \tau M \right) ,
\end{aligned}
$$

where for (a), we use Boole's inequality. We control

$$
K \exp\left(-\frac{(\varepsilon_{\beta'} - \varepsilon'_{\alpha})^2}{4} \min\{\beta',\ \alpha' - \beta'\} \tau M \right) \leq \frac{1}{T} .
$$

Then, a sufficient condition on τM is $\tau M \geq Z' \ln(KT)$ where $Z' = \frac{4}{(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2 \min\{\beta', \alpha' - \beta'\}}$. Therefore, when the condition for τM is satisfied, the regret bound is as follows:

$$
\begin{aligned}
&R_T^\pi(\boldsymbol{\theta}) \\
&\leq \mathbb{E}_\pi \left[\sum_{i\in[K]} (\theta_* - \theta(i)) n_T(i) \mid \mathcal{X} \right] + \Delta_{\boldsymbol{\theta}} T \left(\mathbb{P}[\mathcal{X}^\mathcal{C}] \right) \\
&\overset{(b)}{\leq} (1+\lambda) C(\boldsymbol{\theta}, \boldsymbol{d} + (2\varepsilon_{\beta'} - \varepsilon_{\alpha'})\mathbb{1}) \log T + o(\log T) \\
&\quad + \Delta_{\boldsymbol{\theta}} T \cdot K \exp\left(-\frac{(\varepsilon_{\beta'} - \varepsilon_{\alpha'})^2}{4} \min\{\beta', (\alpha - \beta')\} \tau M \right) \\
&\leq (1+\lambda) C\left(\boldsymbol{\theta}, \boldsymbol{d} + (2\varepsilon_{\beta'} - \varepsilon_{\alpha'})\mathbb{1} \right) \log T + o(\log T) ,
\end{aligned}
$$

where the inequality (b) is from Theorem 2. □

C. Proof of Theorem 6′

The proof is parallel flow with the proof of Theorem 6, but difference in $\mathcal{M}$ and $\mathcal{M}'$ for change-measure argument. Let j be an arm such that $\min_{i\neq j} d(i,j) = \Delta_{\boldsymbol{x}}$. For some $c \in (0,1)$, we consider $\mathcal{M}$ such that for each $m \in [M]$,

$$
\theta_m(i) = \begin{cases} c + \Delta_{\boldsymbol{x}} & \text{if } i = j \\ c & \text{otherwise} \end{cases} .
$$

Notice that one pair of arms $d(i,j) = |\theta_m(j) - \theta_m(i)| = \Delta_{\boldsymbol{x}}$ and the others are 0. Then $\mathcal{M}$ satisfies $\boldsymbol{\theta}_m \in \Phi(\boldsymbol{d})$ and verifies Assumption 1-1 for $d(i,j)$, $\varepsilon_{\alpha'}$ and α'. We construct a perturbation $\mathcal{M}'$: for each $m \in [\lceil \alpha' M \rceil]$,

$$
\lambda_m(i) = \begin{cases} \theta_m(j) + (\varepsilon - \varepsilon_{\alpha'}) & \text{if } i = j \\ \theta_m(i) & \text{otherwise} \end{cases} . \tag{43}
$$

For $i \neq j$, $d'_m := |\lambda_m(j) - \lambda_m(i)| = |\theta_m(j) + (\varepsilon - \varepsilon_{\alpha'}) - \theta_m(i)| = \Delta_{\boldsymbol{x}} + (\varepsilon - \varepsilon_{\alpha'})$ which verifies Assumption 1′ for $d'(i,j) = d(i,j) + \varepsilon$, $\varepsilon_{\alpha'}$ and α'. For $m \in [M] \setminus [\lceil \alpha' M \rceil]$, we set $\boldsymbol{\theta}_m = \boldsymbol{\lambda}_m$ where $\mathcal{M}'$ verifies Assumption 1′. Assume that for each $m \in [M]$, $n_m(i') = \tau$ and $n_m(i) \geq \tau$ if $i \neq j$. which implies Assumption 2. Note that $\mathbb{E}[\mathcal{G}] = \sum_{m\in[M]} \sum_{i\in[K]} \mathbb{E}[n_m(i)] \mathrm{KL}(\theta_m(i) \| \lambda_m(i))$ thanks to Markov property of bandit. With the construction of $\mathcal{M}$ and $\mathcal{M}'$, it follows that

$$
\begin{aligned}
\mathbb{E}[\mathcal{G}] &\leq \tau \left(\sum_{m=1}^{\lceil \alpha' M \rceil} \mathrm{KL}(\theta_m(j) \| \lambda_m(j)) \right) \\
&\leq \tau \left(\sum_{m=1}^{\lceil \alpha' M \rceil} \frac{(\theta_m(j) - \lambda_m(j))^2}{\lambda_m(j)(1 - \lambda_m(j))} \right) ,
\end{aligned}
$$

where for the last inequality, we use the fact that $\mathrm{KL}(\theta \| \lambda) \leq \mathcal{X}^2(\theta, \lambda) = \frac{(\theta - \lambda)^2}{\lambda(1-\lambda)}$, c.f., Lemma 2.7 in [37]. We can select constant $c \in (0,1)$ such that

$$
\begin{aligned}
&\tau \left(\sum_{m=1}^{\lceil \alpha' M \rceil} \frac{(\theta_m(j) - \lambda_m(j))^2}{\lambda_m(j)(1 - \lambda_m(j))} \right) \\
&\qquad = \Omega\left(\tau \left(\sum_{m=1}^{\lceil \alpha' M \rceil} (\theta_m(j) - \lambda_m(j))^2 \right) \right) .
\end{aligned}
$$

With such a choice of c and the construction of $\boldsymbol{\lambda}_m$ in (43), we obtain

$$\mathbb{E}[\mathcal{G}] = \Omega\left(\tau\alpha' M(\varepsilon - \varepsilon_{\alpha'})^2\right) . \tag{44}$$

To complete the proof, we define the event $\mathcal{E} = \{\cap_{i\le j\in[K]}(\hat{d}(i,j) \in [d(i,j), d(i,j)+\varepsilon])\}$ and its complement $\mathcal{E}'$. Under Assumption 2 with $\tau > 0$ and the supposition that estimator matrix $\hat{\boldsymbol{d}}$ is uniformly good for $(\alpha', \varepsilon_{\alpha'}, \tau, \varepsilon)$, we have $\mathbb{P}[\mathcal{E}'] = o\left(\frac{\log T}{T}\right)$ and further

$$\begin{aligned}
&\mathbb{P}'[\mathcal{E}] \qquad (45)\\
&= \mathbb{P}'\Big[\bigcap_{i\le j\in[K]}\Big(\hat{d}(i,j) \in [d(i,j), d(i,j)+\varepsilon]\Big)\Big]\\
&= \mathbb{P}'\Big[\bigcap_{i\le j\in[K]}\Big(\hat{d}(i,j) \in [d'(i,j)-\varepsilon, d'(i,j)]\Big)\Big]\\
&\le \mathbb{P}'\Big[\bigcap_{i\le j\in[K]}\Big((\hat{d}(i,j) < d'(i,j)) \cup (\hat{d}(i,j) > d'(i,j)+\varepsilon)\Big)\Big]\\
&= o\left(\frac{\log T}{T}\right) ,
\end{aligned}$$

where the last equality is from the construction of $\mathcal{M}'$ verifying Assumption 1′ for $d'(i,j) = d(i,j)+\varepsilon$, α', and $\varepsilon_{\alpha'}$. From this, it follows that

$$\mathrm{KL}(\mathbb{P}[\mathcal{E}]\|\mathbb{P}'[\mathcal{E}]) = -\log\left(o\left(\frac{\log T}{T}\right)\right) = O(\log T) . \tag{46}$$

Therefore, combining (44) and (46), Lemma 2 concludes the proof. □

ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)) and (No. RS-2021-II210739, Development of Distributed/Cooperative AI based 5G+ Network Data Analytics Functions and Control Technology). Kwang-Sung Jun was supported in part by the National Science Foundation under grant CCF-2327013.

REFERENCES

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematics Society*, vol. 58, pp. 527–535, 1952.

[2] S. Magureanu, R. Combes, and A. Proutiere, "Lipschitz bandits: Regret lower bounds and optimal algorithms," in *Conference on Learning Theory (COLT)*, 2014.

[3] R. Kleinberg, A. Slivkins, and E. Upfal, "Multi-armed bandits in metric spaces," in *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 2008, pp. 681–690.

[4] S. Bubeck, G. Stoltz, C. Szepesvári, and R. Munos, "Online optimization in x-armed bandits," *Advances in Neural Information Processing Systems*, vol. 21, 2008.

[5] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback." in *COLT*, vol. 2, 2008, p. 3.

[6] A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin, "Stochastic convex optimization with bandit feedback," in *Advances in Neural Information Processing Systems*, 2011, pp. 1035–1043.

[7] J. Y. Yu and S. Mannor, "Unimodal bandits," in *International Conference on Machine Learning (ICML)*, 2011, pp. 41–48.

[8] J. Ok, A. Proutiere, and D. Tranos, "Exploration in structured reinforcement learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 8874–8882.

[9] R. Combes, J. Ok, A. Proutiere, D. Yun, and Y. Yi, "Optimal rate sampling in 802.11 systems: Theory, design, and implementation," *IEEE Transactions on Mobile Computing*, vol. 18, no. 5, pp. 1145–1158, 2018.

[10] H. Qi, Z. Hu, X. Wen, and Z. Lu, "Rate adaptation with thompson sampling in 802.11 ac wlan," *IEEE Communications Letters*, vol. 23, no. 10, pp. 1888–1892, 2019.

[11] R. Combes, S. Magureanu, and A. Proutiere, "Minimal exploration in structured stochastic bandits," in *Advances in Neural Information Processing Systems*, 2017, pp. 1763–1771.

[12] S. Bubeck, G. Stoltz, and J. Y. Yu, "Lipschitz bandits without the lipschitz constant," in *International Conference on Algorithmic Learning Theory*. Springer, 2011, pp. 144–158.

[13] A. Lazaric, E. Brunskill *et al.*, "Online stochastic optimization under correlated bandit feedback," in *International Conference on Machine Learning (ICML)*. PMLR, 2014, pp. 1557–1565.

[14] M. Valko, A. Carpentier, and R. Munos, "Stochastic simultaneous optimistic optimization," in *International Conference on Machine Learning (ICML)*. PMLR, 2013, pp. 19–27.

[15] R. Munos, "Optimistic optimization of a deterministic function without the knowledge of its smoothness," *Advances in neural information processing systems*, vol. 24, 2011.

[16] J.-B. Grill, M. Valko, and R. Munos, "Black-box optimization of noisy functions with unknown smoothness," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[17] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "X-armed bandits." *Journal of Machine Learning Research*, vol. 12, no. 5, 2011.

[18] R. Combes, A. Proutière, and A. Fauquette, "Unimodal bandits with continuous arms: Order-optimal regret without smoothness," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 4, no. 1, pp. 1–28, 2020.

[19] A. Slivkins, "Multi-armed bandits on implicit metric spaces," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[20] A. Cutkosky, C. Dann, A. Das, C. Gentile, A. Pacchiano, and M. Purohit, "Dynamic balancing for model selection in bandits and rl," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 2276–2285.

[21] A. Pacchiano, M. Phan, Y. Abbasi Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari, "Model selection in contextual stochastic bandit problems," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 328–10 337, 2020.

[22] C. Podimata and A. Slivkins, "Adaptive discretization for adversarial lipschitz bandits," in *Conference on Learning Theory (COLT)*. PMLR, 2021, pp. 3788–3805.

[23] A. Slivkins, "Contextual bandits with similarity information," in *Proceedings of the 24th annual Conference On Learning Theory (COLT)*. JMLR Workshop and Conference Proceedings, 2011, pp. 679–702.

[24] R. Kleinberg, A. Slivkins, and E. Upfal, "Multi-armed bandits in metric spaces," in *ACM Symposium on Theory of Computing (STOC)*, 2008, p. 681–690.

[25] M. G. Azar, A. Lazaric, and E. Brunskill, "Sequential transfer in multi-armed bandit with finite set of models," in *Advances in Neural Information Processing Systems*, 2013, pp. 2220–2228.

[26] M. Soare, A. Lazaric, O. Alsharif, and J. Pineau, "Multi-task linear bandits," in *Advances in Neural Information Processing Systems Workshop*, 2014.

[27] L. Cella, A. Lazaric, and M. Pontil, "Meta-learning with stochastic linear bandits," in *International Conference on Machine Learning (ICML)*, 2020.

[28] M. Azizi, T. Duong, Y. Abbasi-Yadkori, A. György, C. Vernade, and M. Ghavamzadeh, "Non-stationary bandits and meta-learning with a small set of optimal arms," in *International Conference on Machine Learning Workshop on Complex feedback in online learning*, 2024.

[29] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th annual conference on learning theory (COLT)*, 2011, pp. 359–376.

[30] E. Kaufmann, "On bayesian index policies for sequential resource allocation," *The Annals of Statistics*, vol. 46, no. 2, pp. 842–865, 2018.

[31] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for markov decision processes," *Mathematics of Operations Research*, vol. 22, no. 1, pp. 222–255, 1997.

[32] E. Mammen, A. B. Tsybakov, and Others, "Smooth discrimination analysis," *The Annals of Statistics*, vol. 27, no. 6, pp. 1808–1829, 1999.

[33] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *The Annals of Statistics*, vol. 32, no. 1, pp. 135–166, 2004.

[34] A. Carpentier and M. Valko, "Simple regret for infinitely many armed bandits," in *International Conference on Machine Learning (ICML)*, 2015, pp. 1133–1141.
[35] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016.
[36] L. Besson, "SMPyBandits: an Open-Source Research Framework for Single and Multi-Players Multi-Arms Bandits (MAB) Algorithms in Python," Online at: github.com/SMPyBandits/SMPyBandits, 2018, code at https://github.com/SMPyBandits/SMPyBandits/, documentation at https://smpybandits.github.io/. [Online]. Available: https://github.com/SMPyBandits/SMPyBandits/
[37] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Hyejin Park is currently pursuing an integrated M.S./Ph.D. in the Graduate School of Artificial Intelligence at Pohang University of Science and Technology (POSTECH) in South Korea. She received the Bachelor of Engineering degree in 2020 from the School of Undergraduate Studies, College of Transdisciplinary Studies, at the Daegu Gyeongbuk Institute of Science and Technology (DGIST) in South Korea. Her research interests primarily focus on adaptive algorithms in sequential decisions, including transfer learning and test-time adaptation.

Seiyun Shin is currently pursuing a Ph.D. degree in Electrical and Computer Engineering at University of Illinois Urbana–Champaign (UIUC). He received the B.S. and M.S. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2012 and 2015, respectively. He is a recipient of Kwanjeong Educational Foundation Fellowship and Mavis Future Faculty Fellowship. His research interests lie in machine learning, algorithm design, and information theory.

Kwang-Sung Jun is an assistant professor in the Department of Computer Science at University of Arizona. His research interests lie in sequential decision-making algorithms, online learning, and concentration of measure. He obtained his Ph.D. from the University of Wisconsin-Madison in 2015. He was then a postdoctral scholar at the Wisconsin Institute for Discovery in the University of Wisconsin-Madison, followed by a postdoctral scholar at Boston University. He has won an outstanding research award from the Department of Computer Science at University of Arizona and was a recipient of the doctoral study abroad scholarship from Korea Foundation of Advanced Studies.

Jungseul Ok received B.S. degree from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), in 2011, and the Ph.D. degree under the supervision of Prof. Yung Yi and Prof. Jinwoo Shin, in 2016. He is currently an associate professor in the Department of Computer Science and Engineering and Graduate School of Artificial Intelligence at Pohang University of Science and Technology. He was a Post-doc researcher in University of Washington, University of Illinois at Urbana-Champaign, KTH Royal Institute of Technology, and Korea Advanced Institute of Science & Technology. His research focus includes machine learning with data collection, including user-friendly active learning, structured reinforcement learning, and robust crowdsourcing systems.