Congested Bandits: Optimal Routing via Short-term Resets

Pranjal Awasthi [1] **Kush Bhatia** [2] **Sreenivas Gollapudi** [1] **Kostas Kollias** [1]

Abstract

For traffic routing platforms, the choice of which route to recommend to a user depends on the congestion on these routes – indeed, an individual's utility depends on the number of people using the recommended route at that instance. Motivated by this, we introduce the problem of Congested Bandits where each arm's reward is allowed to depend on the number of times it was played in the past Δ timesteps. This dependence on past history of actions leads to a dynamical system where an algorithm's present choices also affect its future pay-offs, and requires an algorithm to plan for this. We study the congestion aware formulation in the multi-armed bandit (MAB) setup and in the contextual bandit setup with linear rewards. For the multi-armed setup, we propose a UCB style algorithm and show that its policy regret scales as $\tilde{O}(\sqrt{K\Delta T})$. For the linear contextual bandit setup, our algorithm, based on an iterative least squares planner, achieves policy regret $\tilde{O}(\sqrt{dT} + \Delta)$. From an experimental standpoint, we corroborate the no-regret properties of our algorithms via a simulation study.

1. Introduction

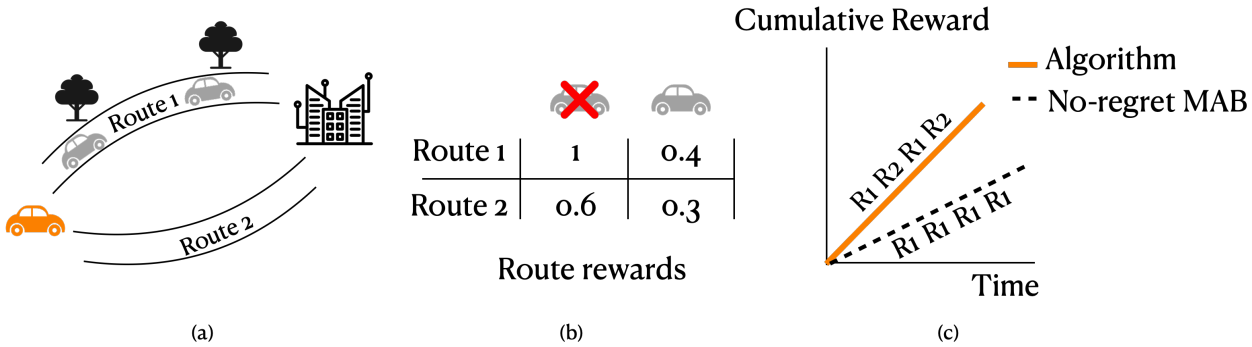
The online multi-armed bandit (MAB) problem and its extensions have been widely studied and used to model many real world scenarios (Robbins, 1952; Auer et al., 2002; 2001). In the basic MAB setup there are K arms with either stochastic or adversarially chosen reward profiles. The goal is to design an algorithm that achieves a cumulative reward that is as good as that of the best arm in hindsight. This is quantified in terms of the regret achieved by the algorithm over T time steps (see Section 2 for formal definitions). In many real world scenarios the MAB setup as described above is not suitable as the reward obtained by playing an arm/action at a given time step may depend on the algorithm's choices in the previous time steps. In particular, we are motivated by online routing problems where the reward of suggesting a particular edge to traverse along a path from source to destination often depends on the congestion on that edge. This congestion is a function of the number of times the particular edge has been recommended earlier (potentially in a time window). In such scenarios, one would desire algorithms that can compete with the best *policy*, i.e., the best sequence of actions, in hindsight as compared to a fixed best action.

Classical multi-armed bandit formulation and associated no-regret algorithms (Slivkins, 2019), or their extensions to routing problems (Kalai & Vempala, 2005; Awerbuch & Kleinberg, 2008) do not suffice for the above scenario as they only guarantee competitiveness with respect to the best fixed arm in hindsight. To overcome these limitations, we propose a new model, *viz.*, *congested bandits* which captures the above scenario. In our proposed model the reward of a given arm at each time step depends on how many times the arm has been played within a given time window of size Δ. Hence over time, an arm's expected reward may decay and reset dynamically. While our model is motivated by online routing problems, our proposed formulation is very general. As another example, consider a digital music platform that recommends artists to its end users. In order to maximize profit the recommendation algorithm may prefer to recommend popular artists, and at the same time the platform may want to promote equity and diversity by highlighting new and emerging artists as well. This scenario can be model via congested bandits where each artist is an arm and the reward for suggesting an artist is a function of how many times the artist has been recommended in the past time window of length Δ. In both the scenarios above, the ability to reset the congestion cost (by simply not playing an arm for Δ time steps) is a crucial part of the problem formulation.

Our contributions. We propose and study the *congested bandits* model with *short term resets* under a variety of settings and design no-regret algorithms. In the most basic setup we consider a K-armed stochastic bandit problem where each arm a has a mean reward μ_a, and the mean reward obtained by playing arm a at time t equals $f_{\text{cong}}(a, h_t)\mu_a$. Here h_t denotes the history of the algorithm's choices within the last Δ time steps and f_{cong} is a non-increasing congestion function. Recall that the algo-

[1]Google Research [2]UC Berkeley. Correspondence to: Kush Bhatia <kushbhatia@berkeley.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.



(a) (b) (c)

Figure 1. Our proposed Congested Bandits framework. (a) A route recommendation scenario where an algorithm can recommend one of two routes to the incoming vehicle. (b) The reward for each route depends on whether there is congestion on the route or not. (c) Traditional multi-armed bandit algorithms learn to recommend the best route, Route 1 for every incoming vehicle. This is clearly suboptimal. Our algorithm, CARMAB, adapts to the congestion and achieves better performance.

rithm's goal in this setup is to compete with the best policy in hindsight. While the above setting can be formulated as a Markov Decision Process (MDP), existing no regret algorithms for MDPs will incur a regret bound that scales exponential in the parameters (scaling as K^{Δ}) (Auer et al., 2009; Jaksch et al., 2010). Instead, we carefully exploit the problem structure to design an algorithm with near-optimal regret scaling as $\tilde{O}(\sqrt{K\Delta T})$. Next, we extend our model to the case of online routing with congested edge rewards, again presenting near optimal no-regret algorithms that avoid exponential dependence on the size of the graph in the regret bounds.

We then extend the multi-armed and the congested online routing formulations to a contextual setting where the mean reward for each arm/edge at a given time step equals $f_{\text{cong}}(a, h_t)\langle\theta_*, \phi(x_t, a)\rangle\rangle$, where $x_t \in \mathbb{R}^d$ is a context vector and θ_* is an unknown parameter. This extension is inspired from classical work in contextual bandits (Li et al., 2010; Chu et al., 2011) and captures scenarios where users may have different preferences over actions/arms (e.g., a user who wants to avoid routes with tolls). Solving the contextual case poses significant hurdles as a priori it is not even clear whether the setting can be captured via an MDP. By exploiting the structure of the problem, we present a novel epoch based algorithm that, at each time step, plays a near optimal policy by planning for the next epoch. Showing that such planning can be done when only given access to the distribution of the contexts is a key technical step in establishing the correctness of the algorithm. As a result, we obtain algorithms that achieve $\tilde{O}(\sqrt{dT} + \Delta)$ regret in the contextual MAB setting. Finally, using simulations, we perform an empirical evaluation of the effectiveness of our proposed algorithms.

Related work. Closest to our work are studies on multi-armed bandits with decaying and/or improving costs. The work of (Levine et al., 2017) proposes the rotting bandits model that has been further studied in (Seznec et al., 2019). In this model the mean reward of each arm decays in a monotonic way as a function of the number of times the arm has been played in the past. There is no notion of a short-term reset as in our setting. As a result it can be shown that a simply greedy policy is optimal in hindsight. In contrast, in our setting greedy approaches can fail miserably as highlighted in Figure 1. The work of Heidari et al. (2016) considers a setting where the mean reward of an arm can either improve or decay as a function of the number of times it has been played. However, similar to the rotting bandits setup there is no notion of a reset. Pike-Burke & Grünewälder (2019) considers a notion of reset/improvement by allowing the mean reward to depend on the number of time steps since an arm was last played. Their work considers a Bayesian setup where the mean reward function is modelled by a draw from a Gaussian process. On the other hand, in our work, the reward is a function of the number of times an arm was played in the past Δ time steps. However their regret bounds are either with respect to the instantaneous regret, or with respect to a weaker policy class of d-step look ahead policies. The notion of instantaneous regret only compares with the class of greedy policies at each timestep as compared to the globally optimal policy.

There also exist works studying the design of online learning algorithms against m-*bounded memory* adversaries. The assumption here is that the reward of an action at each time step depends only on the previous m actions. While this is also true for our setting, in general the regret bounds provided for bounded-memory adversaries are either with respect to a fixed action or a policy class containing policies that do no switch often (Arora et al., 2012; Anava et al., 2015). Our problem formulation for the contextual setup is reminiscent of the recent line of works on contextual MDPs (Azizzadenesheli et al., 2016; Krishnamurthy et al., 2016; Hallak et al., 2015; Modi & Tewari, 2020). However these works either assume that the context vector is fixed

for a given episode (allowing for easier planning), or make strong realizability assumptions on the optimal Q-function. Finally, our congested bandits formulation of the routing problem is a natural extension of classical work on the online shortest path problem (Kalai & Vempala, 2005; Awerbuch & Kleinberg, 2008; Dani et al., 2007).

2. Congested multi-armed bandits

In this section, we model the congestion problem in a multi-armed bandit (MAB) framework.Our setup for congested multi-armed bandits models the congestion phenomenon by allowing the rewards of an arm to depend on the number of times it was played in the past.

Let us consider a MAB setup with K arms, where each arm may represent a possible route. Let us denote by Δ the size of the window which affects the reward at the current time step. For any time t, let $h_t \in \mathcal{H}_\Delta := [K]^\Delta$ denote the history of the actions taken by an algorithm[1] in the past Δ time steps, that is, $h_t = [a_{t-\Delta}, \ldots, a_{t-1}]$ where a_τ is the action chosen by the algorithm at time τ. In order to model congestion arising from repeated plays of a single arm, we consider a function $f_{\mathsf{cong}} : [K] \times [\Delta]_+ \to (0, 1]$. This congestion function takes in two arguments: an arm a and the number of times this arm was played in the past Δ time steps, and outputs a value indicating the *decay* in the reward of arm a arising from congestion.

Protocol for congestion in bandits. We consider the following online learning protocol for a learner in our congested MAB framework: At each round t, the learner picks an arm $a_t \in [K]$ and observes reward

$$\tilde{r}(h_t, a_t) = f_{\mathsf{cong}}(a_t, \#(h_t, a_t)) \cdot \mu_{a_t} + \epsilon_t; \quad \epsilon_t \sim \mathcal{N}(0, 1)).$$

Finally, the history changes to $h_{t+1} = [h_{t,2:\Delta}, a_t]$ where we have used the notation $h_{t,i:j}$ to denote the vector $[h_t(i), \ldots, h_t(j)]$ and $\#(h, a)$ to denote the number of times the action a was played in history h.

Each arm a is associated with a mean reward vector μ_a and the congestion function multiplicatively decreases the reward of that arm. We assume that the learner does not know the exact form of the function f_{cong} as well as the the mean vector $\mu = \{\mu_a\}_{[K]}$. The objective of the learner is to select the actions a_t which minimizes a notion of policy regret, which we define next.

Policy regret for congested MAB. In the standard MAB setup, regret compares the cumulative reward of the algorithm to the benchmark of playing the arm with the highest mean reward at all time steps. However, as described in Section 1, this benchmark is not suitable for our setup. Indeed, the asymptotically optimal algorithm is one which maximizes the average cumulative reward

$$\rho^* := \max_{\mathsf{alg}} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(h_t, a_t),$$

where the action a_t is the one chosen by the algorithm alg and history h_t is the sequence of actions in the past Δ time steps.

This asymptotic algorithm corresponds to a stationary policy π^* whose selection a_t *only* depends on the history h_t. Accordingly, we consider the following class of stationary policies as the comparator for our regret $\Pi = \{\pi : \mathcal{H}_\Delta \to [K]\}$, with size $|\Pi| = K^{K^\Delta}$. Denote by h_t^π the history at time t by running policy π up to time t, we define the policy regret for any algorithm

$$\mathfrak{R}_T(\mathsf{alg}; \Pi, f_{\mathsf{cong}}) := \sup_{\pi\in\Pi} \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi)) - \sum_{t=1}^{T} \tilde{r}(h_t^{\mathsf{alg}}, a_t) \ , \tag{1}$$

where a_t is the action chosen by alg at time t and $r(h_t, a_t) = f_{\mathsf{cong}}(h_t,) \cdot \mu_{a_t}$. This notion of regret is called policy regret (Arora et al., 2012) because the history sequence observed by the algorithm π^* and the algorithm alg can be different from each other – this leads to a situation where choosing the same action a at time t can lead to different rewards for the algorithm and the comparator.

2.1. CARMAB: Congested MAB algorithm

We now describe our learning algorithm for the congested MAB problem. At a high level, CARMAB, detailed in Algorithm 1, is based on a reduction of this problem to a reinforcement learning problem with state space $S = \mathcal{H}_\Delta$ and action space $A = [K]$, where the underlying dynamics are known to the learner. With this reduction, CARMAB deploys an epoch-based strategy which plays a optimistic policy $\tilde{\pi}_\epsilon$ computed from optimistic estimates of the reward function.

Reduction to MDP. Our congested MAB setup can be viewed as a learning problem in a Markov decision process (MDP) with finite state and action spaces. This MDP $\mathcal{M}_{\mathsf{mab}}$ comprises state space $S = \mathcal{H}_\Delta$ and action space $A = [K]$. The reward function for this MDP is given by $r(s, a) = f_{\mathsf{cong}}(a, \#(s, a)) \cdot \mu_a$ and the deterministic state transitions are given

$$P(s'|s, a) = \begin{cases} 1 & \text{if } s' = [s_{2:\Delta}, a] \\ 0 & \text{o.w.} \end{cases} , \tag{2}$$

where we have again used the notation $s_{a:b}$ to denote the vector $[s(a), \ldots, s(b)]$.

[1] We usually suppress the dependence of this history on the algorithm, but make it explicit whenever it is not clear from context.

Algorithm 1: CARMAB: Congestion Aware Routing via Multi-Armed Bandits

Input: Congestion window Δ, confidence parameter $\delta \in (0,1)$, action set $[K]$, time horizon T.
Initialize: Set $t = 1$
for *episodes* $\mathfrak{e} = 1, \ldots, \mathcal{E}$ **do**
  Initialize episode
  Set start time of episode $t_{\mathfrak{e}} = t$.
  For all actions a and historical count j, set $n_{\mathfrak{e}}(a,j) = 0$ and $N_{\mathfrak{e}}(a,j) = \sum_{s<e} n_s(a,j)$.
  Set empirical reward estimate for each arm and historical count

$$\hat{r}_{\mathfrak{e}}(a,j) = \frac{\sum_{\tau=1}^{t_{\mathfrak{e}}-1} r_\tau \cdot \mathbb{I}[a_\tau = a, j_\tau = j]}{\max\{1, N_{\mathfrak{e}}(a,j)\}}.$$

  Compute optimistic policy
  Set the feasible rewards

$$\mathcal{R}_{\mathfrak{e}} = \Big\{ r \in [0,1]^{A\times(\Delta+1)} \mid \text{for all } (a,j), \\ |\, r(a,j) - \hat{r}_{\mathfrak{e}}(a,j)| \le 10\sqrt{\frac{\log(A\Delta t_{\mathfrak{e}}/\delta)}{\max\{1, N_{\mathfrak{e}}(a,j)\}}} \Big\}$$

  Find optimistic policy $\tilde{\pi}_{\mathfrak{e}} = \arg\max_{\pi\in\Pi, r\in\mathcal{R}_{\mathfrak{e}}} \rho(\pi, r)$
  Execute optimistic policy
  while $n_{\mathfrak{e}}(a,j) < \max\{1, N_{\mathfrak{e}}(a,j)\}$ **do**
    Select arm $a_t = \tilde{\pi}_{\mathfrak{e}}(s_t)$, obtain reward $\hat{r}_t$.
    Update $n_{\mathfrak{e}}(a, \#(s_t,a)) = n_{\mathfrak{e}}(a, \#(s_t,a)) + 1$.

Algorithm details. Our learning algorithm in this MDP is an upper confidence bound (UCB) style algorithm, adapted from the classical UCRL2 (Jaksch et al., 2010) for learning in finite MDPs. It splits the time horizon T into a total of $\mathcal{E}$ epochs, each of which can be of varying length. In each episode, for every pair (a,j) of action a and historical count $j \in [\Delta]_+$, the algorithm computes the empirical estimate of the rewards $\hat{r}_{\mathfrak{e}}(a,j)$ from observations in the past epoch and maintains a feasible set $\mathcal{R}_{\mathfrak{e}}$ of rewards. This set is constructed such that with high probability, the true reward r belong to this set for each epoch. Given this set, our algorithm computes the optimistic policy

$$\tilde{\pi}_{\mathfrak{e}} = \arg \max_{\pi\in\Pi, r\in\mathcal{R}_{\mathfrak{e}}} \rho_\pi(r) \\ \text{where} \quad \rho_\pi := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi)) \,, \tag{3}$$

that is, the policy which achieves the best average expected reward with respect to the optimistic set $\mathcal{R}_{\mathfrak{e}}$. The optimistic policy $\pi_{\mathfrak{e}}$ is then deployed in the congested MAB setup till one of the (a,j) pair doubles in the number of times it is played and this determines the size of any epoch.

Computing optimistic policy. In the MDP described above, each deterministic policy π follows a cyclical path since the transition dynamics are deterministic and the state space is finite. With this insight, this problem of finding the optimal policy with highest average reward is equivalent to finding the maximum mean cycle in a weighted directed graph. In our simulations, we use Karp's algorithm (Karp, 1978) for finding these optimistic policy. This algorithm runs in time $O(K^{\Delta+1})$.

2.2. Regret analysis for CARMAB

In this section, we obtain a bound on the policy regret of the proposed algorithm CARMAB. Our overall proof strategy is to first establish that the MDP $\mathcal{M}_{\mathsf{mab}}$ has a low diameter (the time taken to move from one state to another in the MDP), then bounding the regret in each episode e of the process and finally establishing that the total number of episodes $\mathcal{E}$ can be at most logarithmic in the time horizon T. Combining these elements, we establish in the following theorem that the regret of CARMAB scales as $\tilde{O}(\sqrt{K\Delta T})$ with high probability.[2]

Theorem 1 (Regret bound for CARMAB)**.** *For any confidence $\delta \in (0,1)$, congestion window $\Delta > 0$ and time horizon $T > \Delta K$, the policy regret (1), of* CARMAB *is*

$$\mathfrak{R}_T(\text{CARMAB}; \Pi, f_{\mathsf{cong}}) \le c\cdot\Delta^2 K \log\left(\frac{T}{\Delta K}\right) \\ + c\sqrt{K\Delta T \log\left(\frac{\Delta K T}{\delta}\right)} + c\sqrt{T\log\left(\frac{1}{\delta}\right)}$$

with probability at least $1-\delta$.

A few comments on the theorem are in order. Observe that the dominating term in the above regret bound scales as $\tilde{O}(\sqrt{K\Delta T})$ in contrast to the classical regret bounds for MAB which have a $\tilde{O}(\sqrt{KT})$ dependence. This additional factor of $\sqrt{\Delta}$ comes from the fact the stronger notion of policy regret as well as the non-stationary nature of the arm rewards. Additionally, a naïve application of the UCRL2 regret bound to the constructed MDP scales linearly with state space and would correspond to an additional factor of $O(K^\Delta)$. The CARMAB algorithm is able to avoid this exponential dependence by exploiting the underlying structure in the congested MAB problem. Our regret bound are also minimax optimal – observe that for any constant value of Δ, the lower bounds from the classical MAB setup immediately imply that the regret of any learner should scale as $\Omega(\sqrt{KT})$ (Lattimore & Szepesvári, 2020), which matches the upper bound in Theorem 1.

[2]For clarity purposes, through out the paper we denote by c an absolute constant whose value is independent of any problem parameter. We allow this value of c to change from line to line.

We defer the complete proof of this to Appendix A but provide a high-level sketch of the important arguments.

MDP $\mathcal{M}_{\mathsf{mab}}$ has bounded diameter. The diameter D of an MDP $\mathcal{M}$ measures the number of steps it takes to reach a state s' from a state s using an appropriately chosen policy. The diameter is a measure of the connectedness of the underlying MDP and is commonly studied in the literature on reinforcement learning (Puterman, 2014).

Definition 1 (Diameter of MDP.)**.** *Consider the stochastic process induced by the policy π on an MDP $\mathcal{M}$. Let $\tau(s'|s, \mathcal{M}, \pi)$ represent the first time the policy reaches state s' starting from s. The diameter D of the MDP is*

$$D := \max_{s \neq s'} \min_{\pi} \mathbb{E}[\tau(s'|s, \mathcal{M}, \pi)].$$

Recall from Section 2.1 that the MDP $\mathcal{M}_{\mathsf{mab}}$ has deterministic dynamics and state space given by the set of histories $\mathcal{H}_\Delta$. Proposition 2 in Appendix A establishes that the diameter of this MDP is at most the window size Δ. With this bound on the diameter, we then show in Lemma 4 that the total regret of the algorithm can be decomposed into a sum of regret terms, one for each episode.

$$\begin{aligned} \mathfrak{R}_T &\leq \sup_{\pi \in \Pi} \sum_{\mathfrak{e}=1}^{\mathcal{E}} \mathfrak{r}_{\mathfrak{e}} + c\sqrt{T \log\left(\frac{T}{\delta}\right)} \\ \text{with } \mathfrak{r}_{\mathfrak{e}} &:= \sum_{a,j} n_{\mathfrak{e}}(a,j)(\rho_\pi - f_{\mathsf{cong}}(a,j)\mu_a) \end{aligned}, \tag{4}$$

which holds with probability at least $1 - \delta$. We have used $n_{\mathfrak{e}}(a, j)$ to denote the number of times action a was played by the algorithm when it had a count j in the history and ρ_π the average reward of policy π. Our analysis then proceeds to bound the per-episode regret $\mathfrak{r}_{\mathfrak{e}}$.

Regret for episode $\mathfrak{e}$. In episode e, the set of feasible rewards $\mathcal{R}_{\mathfrak{e}}$ is chosen to ensure that with high probability, the true reward $r^*(a, j) = f_{\mathsf{cong}}(a, j) \cdot \mu_j$ belongs to this set. Conditioning on this event, we show that the regret in each episode is upper bounded by the window size Δ and a scaled ratio of the number of times each action-history (a, j) is played, that is,

$$\mathfrak{r}_{\mathfrak{e}} \leq \Delta + c\sqrt{\log\left(\frac{\Delta K t_{\mathfrak{e}}}{\delta}\right)} \sum_{a,j} \frac{n_{\mathfrak{e}}(a,j)}{\sqrt{\max(1, N_{\mathfrak{e}}(a,j))}} \,. \tag{5}$$

where $t_{\mathfrak{e}}$ is the time at which episode $\mathfrak{e}$ starts and $N_{\mathfrak{e}}(a, j) = \sum_{i<e} n_i(a, j)$. Theorem 1 follows from combining the above with a bound on the total number of episodes $\mathcal{E} \leq c \cdot \Delta K \log\left(\frac{T}{\Delta K}\right)$.

2.3. Routing with congested bandits

We now study an extension of the congested MAB setup where the arms correspond to edges on graph $G = (V, E)$ with a pre-defined start state s_{G} and goal state t_{G}. In this setup, at each round t the learner selects an s_{G}-t_{G} path p_t on the graph G and receives reward $\tilde{r}(h_t, e_{i,t}) = f_{\mathsf{cong}}(e_{i,t}, \#(h_t, e_{i,t})) \cdot \mu_{e_{i,t}} + \epsilon_t$ for each $e_{i,t}$ on path p_t. The history changes to $h_{t+1} = [h_{t,2:\Delta}, p_t]$.

In comparison to the multi-armed bandit protocol, the learner here selects an s_{G}-t_{G} path on the graph G, the history at any time t consists of the entire set of paths $\{p_{t-\Delta}, \ldots, p_{t-1}\}$, and we assume that the congestion function on each edge $f_{\mathsf{cong}}(h, e)$ depends on the number of times this edge has been used in the past Δ time steps. The following theorem generalizes the result from Theorem 1 and shows that a variant of CARMAB has regret $\tilde{O}(\sqrt{T})$ for the above s_{G}-t_{G} online problem.

Theorem 2 (Regret bound for CARMAB-ST)**.** *For any confidence $\delta \in (0, 1)$, congestion window $\Delta > 0$ and time horizon T, the policy regret, of* CARMAB-ST *is*

$$\begin{aligned} \mathfrak{R}_T(\textsc{Carmab-st}; \Pi, f_{\mathsf{cong}}) &\leq c \cdot \Delta^2 VE \log\left(\frac{VT}{\Delta E}\right) \\ &+ c\sqrt{VE\Delta T \log\left(\frac{VE\Delta T}{\delta}\right)} + c\sqrt{VT \log\left(\frac{1}{\delta}\right)} \end{aligned}$$

with probability at least $1 - \delta$.

The proof of the above theorem is detailed in Appendix A.3 and follows a very similar strategy to the one described in the previous section.

3. Linear contextual bandits with congestion

We now consider the contextual version of the congested bandit problem, where the reward function depends on the choice of arm a as well as an underlying context $x \in \mathcal{X}$.While most of our notation stays the same from the multi-armed bandit setup in Section 2, we introduce the modifications required to account for the context vectors x_t. We consider the linear contextual bandit problem where the reward function is parameterized as a linear function of a parameter θ_* and context-action features $\phi(x_t, a_t)$, that is,

$$r(h_t, a_t, x_t; \theta_*) := \langle \theta_*, \phi(x_t, a_t) \rangle f_{\mathsf{cong}}(a_t, \#(h_t, a_t)) \,,$$

where we assume that the context-action features $\phi(x_t, a_t) \subset \mathbb{R}^d$ satisfy $\|\phi(x_t, a_t)\|_2 \leq 1$ and the true parameter $\|\theta_*\|_2 \leq 1$. In contrast to the bandit setup, we expand our policy class to be dependent on the context as well with $\Pi_{\mathcal{X}} := \{\pi : \mathcal{H}_\Delta \times \mathcal{X} \mapsto [K]\}$.

In each round of the linear contextual bandits with congestion game, the learner observes context vectors

Algorithm 2: CARCB: Congested linear contextual bandits with known contexts

Input: Congestion window Δ, congestion function f_{cong}, action set $[K]$, time horizon T, contexts $\{x_t\}_{t=1}^T$

Initialize: Set $t = 1$, $\theta_1 \sim \text{unif}(\mathbb{B}_d)$

for *episodes* $\mathfrak{e} = 1, \dots, \mathcal{E}$ **do**

Initialize episode

Set start time of episode $t_\mathfrak{e} = t$.

Let the steps in this epoch $I_e = [t_\mathfrak{e}, \dots, t_\mathfrak{e} + 2^\mathfrak{e}\Delta]$.

Set the episode policy $\tilde{\pi}_\mathfrak{e} = \arg\max_\pi \sum_{t \in I_\mathfrak{e} \setminus \{t_\mathfrak{e} : t_\mathfrak{e} + \Delta\}} r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_\mathfrak{e})$

Execute estimated policy

for $t = t_\mathfrak{e}, \dots, t_\mathfrak{e} + 2^\mathfrak{e}\Delta$ **do**

Select arm $a_t = \tilde{\pi}_\mathfrak{e}(h_t, x_t)$ and observe reward $\hat{r}_t$.

Update $\theta_\mathfrak{e}$ via OLS update $\theta_{\mathfrak{e}+1} = \arg\min_\theta \sum_\tau (r_\tau - \langle \theta, \phi(x_\tau, a_\tau) \rangle f_{\mathsf{cong}}(a_\tau, \#(h_\tau, a_\tau))^2$

$\{\phi(x_t, a_i)\}_{[K]}$ and selects action a_t. The learner then observes reward $\tilde{r}(h_t, x_t, a_t) = r(h_t, a_t, x_t; \theta_*) + \epsilon_t$ and the history changes to $h_{t+1} = [h_{t,2:\Delta}, a_t]$. The objective of the learner in the above contextual bandit game is to output a sequence of actions which are competitive with the best policy $\pi \in \Pi_\mathcal{X}$. Formally, the regret of an algorithm alg is defined to be

$$\mathfrak{R}_T(\mathsf{alg}; \Pi_\mathcal{X}, f_{\mathsf{cong}}, \theta_*) := \sup_{\pi \in \Pi_\mathcal{X}} \sum_{t=1}^T r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_*) - \sum_{t=1}^T \tilde{r}(h_t^{\mathsf{alg}}, a_t, x_t; \theta_*) \,. \tag{6}$$

In order to provide some intuition about the algorithm, we start with a simple case where all the contexts are known to the learner in advance and later generalize the results to stochastic contexts.

3.1. Warm-up: Known contexts

In the known context setup, the learner is provided access to a set of contexts $\{x_t\}$ at the start of the online learning game. Algorithm 2 details our proposed algorithm, CARCB, for this setup.

Algorithm details. We again divide the total time T into $\mathcal{E}$ episodes, where the length of each episode $\mathfrak{e} = 2^\mathfrak{e}\Delta$. Unlike CARMAB, the algorithm does not maintain any optimistic estimate of the reward parameter θ_* but simply updates it via an ordinary least squares (OLS) procedure and executes the policy $\tilde{\pi}_\mathfrak{e}$ which maximizes this estimated reward function. The core idea underlying this algorithm is that as we observe more samples, our estimate $\theta_\mathfrak{e}$ converges to θ_* and our planner is then able to execute the optimal sequence of actions.

Regret analysis. To analyze the regret for CARCB, we study the error incurred in estimating the parameter $\theta_\mathfrak{e}$ from the reward samples. To do so, we begin by making the following assumption on the minimum eigenvalue of the sample covariance matrix obtained at any time $t_\mathfrak{e}$.

Assumption 1. *For $t > cd$ and for any sequence of actions $\{a_1, \dots, a_T\}$, we have*

$$\lambda_{\min}\left(\frac{1}{t}\sum_{\tau \le t} \phi(x_t, a_t)\phi(x_t, a_t)^\top\right) \ge \gamma,$$

for some value $\gamma > 0$.

Our bound on the regret $\mathfrak{R}_T$ will depend on this minimum eigenvalue γ. Later when we generalize our setup to the unknown setup, we will show that this assumption holds with high probability for a large class of distributions. The following theorem shows that the regret bound for CARCB scales as $\tilde{O}(\sqrt{dT} + \Delta)$ with high probability.

Proposition 1 (Regret bound; known contexts)**.** *For any confidence $\delta \in (0, 1)$, congestion window $\Delta > 0$ and time horizon $T > cd$, suppose that the sample covariance Σ_t satisfies Assumption 1. Then, the policy regret, defined in eq. (6), of* CARCB *with respect to the set Π is*

$$\mathfrak{R}_T(\text{CARCB}; \Pi_\mathcal{X}, f_{\mathsf{cong}}) \le \frac{c}{\gamma \cdot c_{\min}} \sqrt{d(T+\Delta) \cdot \log \frac{\log(T)}{\delta}} + \Delta \log(T) + c\sqrt{T \log\left(\frac{K}{\delta}\right)}$$

with probability at least $1 - \delta$ where $c_{\min} = \min_{a,j} f_{\mathsf{cong}}(a, j)$.

The proof of the above theorem is deferred to Appendix B. At a high level, the proof proceeds in two steps where first it upper bounds the error $\|\theta_\mathfrak{e} - \theta_*\|_2$ for every epoch $\mathfrak{e}$ and then uses this to bound the deviation of the policy $\tilde{\pi}_\mathfrak{e}$ from the optimal choice of policy π^*. In the next section, we generalize this result to the unknown context setup.

3.2. Unknown stochastic contexts

Our stochastic setup assumes that the context vectors $\{\phi(x_t, a_t)\}$ at each time step are sampled i.i.d. from a *known* distribution. We formally state this assumption[3] next.

[3] While our results are stated in terms of the multivariate Gaussian distribution, these can be generalized to sub-Gaussian distribution.

Assumption 2 (Stochastic contexts from known distribution.)**.** *At each time instance t, the features $\phi(x_t, a)$ for every action $a \in [K]$ are assumed to be sampled i.i.d. from the Gaussian distribution $\mathcal{N}(\bar{x}_a, \Sigma_a)$, such that $\alpha_l I \preceq \Sigma_a \preceq \alpha_u I$ and $\|\bar{x}_a\|_2 \leq 1$.*

For large-scale recommendation systems for traffic routing which interact with million of users daily, the above assumption on known distributions is not restrictive at all. Indeed these systems have a fair understanding of the demographics of the population which interact with it on a daily basis and the real uncertainty is on which person from this population will be using the system at any time.

The algorithm for this setup is similar to the known context scenario where instead of planning with the exact contexts in the optimistic policy computation step, we obtain the episode policy as

$$\tilde{\pi}_{\mathfrak{e}} = \arg\max_{\pi \in \Pi} \mathbb{E}[\sum_{t \in I_{\mathfrak{e}} \setminus \{t_{\mathfrak{e}}, \ldots, t_{\mathfrak{e}} + \Delta\}} r(h_t^{\pi}, \pi(h_t^{\pi}, x_t), x_t; \theta_{\mathfrak{e}})] \ ,$$

where the expectation is taken with respect to the sampling of context. Our regret bound for this modified algorithm depends on the mixing time of the policy set $\Pi_{\mathcal{X}}$ in an appropriately defined Markov chain.

Definition 2 (Mixing-time of Markov chain)**.** *For an ergodic discrete time Markov chain M, let d represent an arbitrary starting state distribution and let d^* denote the stationary distribution. The ϵ-mixing time $\tau_{\mathsf{mix}}(\epsilon)$ is defined as*

$$\tau_{\mathsf{mix}}(\epsilon) = \min\{t : \max_{d} \|dM^t - d^*\|_{TV} \leq \epsilon\}.$$

The mixing time of the policy set $\Pi_{\mathcal{X}}$ is given by $\tau^*_{\mathsf{mix}} := \max_{\pi \in \Pi_{\mathcal{X}}} \max_h \tau_{\mathsf{mix},\pi}(h)$. The following theorem establishes the regret bound for the modified CARCB algorithm, showing that not knowing the context can increase the regret by an additive factor of $\tilde{O}(\sqrt{\Delta \tau^*_{\mathsf{mix}} \cdot T})$.

Theorem 3 (Regret bound; unknown contexts)**.** *For any confidence $\delta \in (0, 1)$, congestion window $\Delta > 0$ and time horizon T, suppose that the context sampling distributions satisfy Assumption 2. Then, with probability at least $1 - \delta$, the policy regret of* CARCB *satisfies*

$$\begin{aligned}\mathfrak{R}_T(\textsc{Carcb}; \Pi_{\mathcal{X}}, f_{\mathsf{cong}}) &\leq c \cdot \sqrt{\alpha_u T \log\left(\frac{K}{\delta}\right)} \\ &+ c \cdot \sqrt{\Delta \tau^*_{\mathsf{mix}} T \log\left(\frac{K \log(T)}{\delta}\right)} \\ &+ \frac{c\alpha_u}{c_{\mathsf{min}} \alpha_l} \cdot \sqrt{d(T + \Delta) \cdot \log \frac{\log(T)}{\delta}} + \Delta \log(T) \ . \end{aligned} \tag{7}$$

A detailed proof of this result is deferred to Appendix B. Observe that the above bound can be seen as a sum of two terms: $\mathfrak{R}_T \lesssim \sqrt{dT} + \sqrt{\Delta \tau^*_{\mathsf{mix}} T}$. The first term is a standard regret bound in the d dimensional contextual bandit setup. The second term, particular to our setup, arises because of the interaction of the congestion window with the unknown stochastic contexts. In comparison to the bound in Theorem 1, the window size Δ interacts only additively in the regret bound surprisingly. The reason for this additive deterioration of regret is that the shared parameter θ_* allows us to use data across time steps in our estimation procedure – thus, in effect, the congestion only slows the estimation by a factor of c_{min} which shows up due to the dependence on the minimum eigenvalue.

In order to go from the regret bound in the known context case, Proposition 1, we need to address two key technical challenges: 1) bound the deviation of the reward of policy $\tilde{\pi}_{\mathfrak{e}}$ from the policy which plans with the known sampled contexts, and 2) the context vectors selected by the algorithm $\phi(x_t, a_t)$ satisfy the minimum eigenvalue condition in Assumption 1.

Deviation from known contexts. One way to get around this difficulty is to reduce the above problem to the multi-armed bandit on from Section 2. This simple reduction would lead to a regret bound which scales with the size of the context space $|\mathcal{X}|$ which is exponentially large in the dimension d. Instead of this, we show that the reward obtained by the distribution maximizer $\tilde{\pi}_{\mathfrak{e}}$ are close to those obtained by the sample maximizer via a concentration argument for random walks on the induced Markov chains. The key to our analysis is the construction of this random walk using policy $\tilde{\pi}_{\mathfrak{e}}$ and then using the following concentration bound from .

Lemma 1 (Theorem 3.1 in (Chung et al., 2012))**.** *Let $\mathcal{M}$ be an ergodic Markov chain with state space $[n]$ and stationary distribution d^*. Let $(V_1, ..., V_t)$ denote a t-step random walk on M starting from an initial distribution d on $[n]$. Let $\mu = \mathbb{E}_{V \sim d^*}[f(V)]$ denote the expected reward over the stationary distribution and $X = \sum_i f(V_i)$ denote the sum of function on the random walk. There exists a universal constant $c > 0$ such that*

$$\Pr(|X - \mu t| \geq \delta \mu t) \leq c \|d\|_{d^*} \exp\left(\frac{-\delta^2 \mu t}{72 \tau_{\mathsf{mix}}}\right) \quad \text{for } 0 \leq \delta < 1 \ , \tag{8}$$

where the norm $\|d\|^2_{d^*} := \sum_i \frac{d_i^2}{d_i^*}$.

This concentration bound is not directly applicable to our setup because of two reasons: 1) the constructed Markov chain $\mathcal{M}_{\tilde{\pi}_{\mathfrak{e}}}$ might not be ergodic, and 2) the norm $\|d\|_{d^*}$ might be unbounded in our setup. By using the fact that

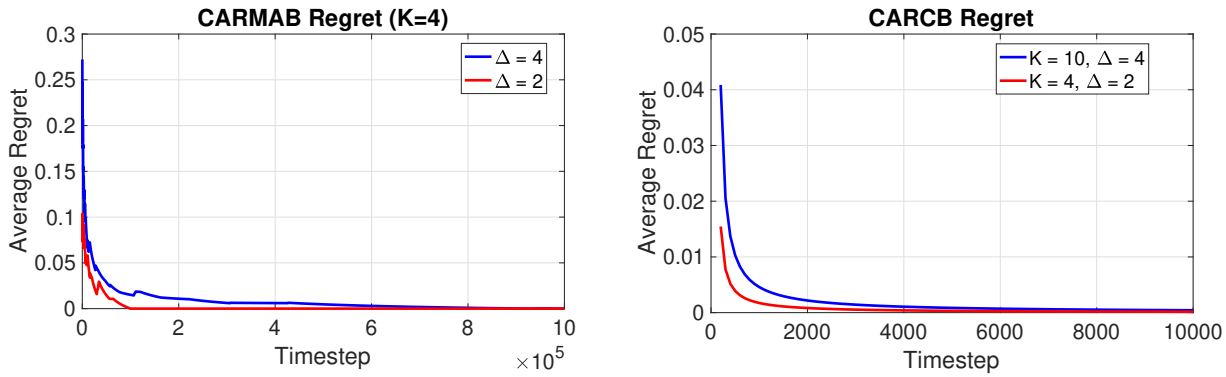


Figure 2. (Left) No-regret property of CARMAB. CARMAB is able to learn optimal sequence of arms to play and enjoys a no-regret property. Increasing the size of history window Δ makes the problem more challenging and requires larger number of time steps. (Right) No-regret property of CARCB.

the diameter of the MDP $\mathcal{M}_{\tilde{\pi}_\mathfrak{e}}$ is bounded by Δ, we use an intermediate policy in the time steps $\{t_\mathfrak{e} : t_\mathfrak{e} + \Delta\}$ in each episode reach a state starting from which the MDP is shown to be ergodic and have bounded norm $\|d\|_{d^*}$. See Appendix B for details.

Minimum eigenvalue bound. In order to obtain a regret bound for the stochastic setup, we need to establish that the covariance matrix formed by these context-action features satisfies the minimum eigenvalue assumption. The challenge here is that the the features $\phi(x_t, a_t)$ are not independent across time – they are correlated since the algorithm's choice at time t depends on the history h_t which in turn depends on the past features. In Appendix B we get around this difficulty by decoupling these dependencies and showing that even after this decoupling, the random variables still satisfy a sub-exponential moment inequality.

4. Experimental evaluation

In this section, we evaluate both our proposed algorithms, CARMAB and CARCB, in the congested bandit framework and exhibit their no-regret properties.

We generate K arms and assign a base reward of $\hat{r}_j \in (0, 1)$ to each $j \in [K]$. We draw a noise parameter $\epsilon_{t,j}$ for every action j and time step t. We set $f_{\mathsf{cong}}(a_t, \#(h_t, a_t)) = 1/\#(h_t, a_t)$. We also set the parameter Δ, which controls the length of the history h_t at time t. Then, the observed reward is $\tilde{r}(h_t, a_t) = \frac{\hat{r}_{a_t}}{\#(h_t, a_t)} + \epsilon_{t,a_t}$. We set parameter δ of algorithm CARMAB to 0.1. In terms of distributions, we draw $\hat{r}_j$ uniformly in $(0, 1)$ and $\epsilon_{t,j}$ from $\mathcal{N}(0, 0.1)$. In Figure 2, we present how the average regret of the algorithm changes as time progresses for $K = 4$ and different values of the window size Δ during which congestion occurs.

For evaluating CARCB, again we generate K arms. For each arm $i \in [K]$ and each time step $t \in [T]$, we draw a random context. A context $x_{a,t}$ is a vector of 10 numbers which are drawn uniformly in $(0, 1)$ and then normalized so that the Euclidean norm of the vector is unit. We also draw the true parameter θ_* in the same way. We assume each arm's context is available to the algorithm at each time step. We use the same noise and congestion function as in the previous section. The observed reward in this setting is:

$$\tilde{r}(h_t, a_t, x_{a,t}, \theta_*) = \frac{x_{a,t}^T \theta_*}{\#(h_t, a_t)} + \epsilon_{t,a_t}$$

In Figure 2 we present how the average regret of CARCB changes over time, in a setting with similar K and Δ ($K = \Delta = 4$) and a setting with a number of actions larger than the congestion window ($K = 10$ and $\Delta = 2$).

5. Discussion

In this work, we introduced the problem of Congested Bandits to model applications such as traffic routing and music recommendations. Our proposed framework allows the utility of any action to depend on the number of times it was played in the past few time steps. From a theoretical perspective, this leads to a rich class of non-stationary bandit problems and we propose near-optimal algorithms for the multi-armed and linear contextual bandit setups.

Our work naturally leads to several interesting open problems. In the multi-armed bandit setup, can we generalize the results to scenarios where the congestion function f_{cong} depends arbitrarily on the history h_t? How does the complexity of this unknown congestion function affect the regret bounds? In the contextual bandit setup, a natural followup to our results would be to extend these beyond linear rewards. More generally, for the non-episodic contextual MDP setup can we design no-regret algorithms which do not depend on the mixing time τ^*_{mix} of the underlying MDP?

From an application perspective, our work is a first step towards incorporating network congestion as a constraint

in the bandit formulation. Going forward, it would be interesting to view the arm choices as recommendations to the users and incorporate user choice models in the framework as a step towards personalized route recommendations.

References

Anava, O., Hazan, E., and Mannor, S. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, pp. 784–792. Citeseer, 2015.

Arora, R., Dekel, O., and Tewari, A. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.

Auer, P., Cesa-Bianchi, N., and Schapire, Y. F. R. E. The non-stochastic multi-armed bandit problem. 2001.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pp. 89–96, 2009.

Awerbuch, B. and Kleinberg, R. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74:97–114, 2008.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. Reinforcement learning of contextual mdps using spectral methods. *arXiv preprint arXiv:1611.03907*, 2016.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Chung, K.-M., Lam, H., Liu, Z., and Mitzenmacher, M. Chernoff-hoeffding bounds for markov chains: Generalized and simplified. *arXiv preprint arXiv:1201.0559*, 2012.

Dani, V., Hayes, T. P., and Kakade, S. The price of bandit information for online optimization. 2007.

Hallak, A., Di Castro, D., and Mannor, S. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Heidari, H., Kearns, M. J., and Roth, A. Tight policy regret bounds for improving and decaying bandits. In *IJCAI*, pp. 1562–1570, 2016.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Kalai, A. and Vempala, S. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Karp, R. M. A characterization of the minimum cycle mean in a digraph. *Discrete mathematics*, 23(3):309–311, 1978.

Krishnamurthy, A., Agarwal, A., and Langford, J. Contextual-mdps for pacreinforcement learning with rich observations. *arXiv preprint arXiv:1602.02722*, 2016.

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

Levine, N., Crammer, K., and Mannor, S. Rotting bandits. *arXiv preprint arXiv:1702.07274*, 2017.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Modi, A. and Tewari, A. No-regret exploration in contextual reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 829–838. PMLR, 2020.

Pike-Burke, C. and Grünewälder, S. Recovering bandits. *arXiv preprint arXiv:1910.14354*, 2019.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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

Seznec, J., Locatelli, A., Carpentier, A., Lazaric, A., and Valko, M. Rotting bandits are no harder than stochastic ones. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

Slivkins, A. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

A. Proof for Congested Multi-armed bandits

In this section, we provide the proofs for the main results from Section 2.

A.1. Diameter of $\mathcal{M}_{\mathsf{mab}}$

Proposition 2. *The diameter of the MDP $\mathcal{M}_{\mathsf{mab}}$ is at most Δ.*

Proof. In order to prove that the diameter of the MDP $\mathcal{M}_{\mathsf{mab}}$ is at most the window size Δ, we need to show that starting form any (history) state h_1, there exists a policy which will take it to another state h_2 in Δ steps. First, note the both the states h_1 and h_2 are Δ dimensional vectors.

Since the dynamics are deterministic, we will construct a deterministic policy for any pair of states h_1 and h_2. The proof begins by concatenating the two histories to form a longer sequence of length 2Δ., that is $\tilde{h} = [h_1; h_2]$. For any sequence $h_{\Delta,i}$ of length Δ in the sequence $\tilde{h}$ starting at position i, set the policy $\pi(h_{\Delta,i}) = \tilde{h}[i + \Delta]$. For any subsequence $h_{\Delta,i}$ which occurs at multiple indices, say $\{i_1, \ldots, i_h\}$, set the policy $\pi(h_{\Delta,i}) = \tilde{h}[i + h + \Delta]$, that is corresponding to the last occurrence of this subsequence.

The policy π constructed above will be definition move from state h_1 to h_2 and in the worst-case when all the subsequences $h_{\Delta,i}$ are unique, it will take Δ steps to complete. Thus, the MDP $\mathcal{M}_{\mathsf{mab}}$ has diameter at most Δ. □

A.2. Proof of Theorem 1

We now begin with our proof of the regret bound for our algorithm CARMAB (Algorithm 1) stated in Theorem 1. Recall we defined the regret of an algorithm

$$\mathfrak{R}_T(\mathsf{alg}; \Pi, f_{\mathsf{cong}}) := \sup_{\pi \in \Pi} \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi)) - \sum_{t=1}^{T} \tilde{r}(h_t^{\mathsf{alg}}, a_t) \ ,$$

and the stationary reward of any policy $\pi \in \Pi$ as

$$\rho_\pi := \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi)) \ .$$

We begin by upper bounding the comparator reward by the optimal average reward ρ^* in the following lemma.

Lemma 2. *For the MDP $\mathcal{M}_{\mathsf{mab}}$, the reward of the optimal policy is*

$$\sup_{\pi \in \Pi} \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi)) \leq T\rho^* + \Delta. \tag{9}$$

The proof of this lemma is deferred to later in the section. Taking this as given, we have that the regret

$$\mathfrak{R}_T(\textsc{CarMab}; \Pi, f_{\mathsf{cong}}) \leq T\rho^* - \sum_{t=1}^{T} \tilde{r}(h_t^{\mathsf{alg}}, a_t) + \Delta \ . \tag{10}$$

We begin by decomposing the above regret term into sub-term for each episode. the following lemma shows that with high probability we can bound the regret in terms of the mean rewrd in each episode.

Lemma 3. *With probability at least $1 - \delta$, we have that the regret*

$$\mathfrak{R}_T(\mathsf{alg}; \Pi, f_{\mathsf{cong}}) \leq \sup_{\pi \in \Pi} \sum_{\mathfrak{e}=1}^{E} \sum_{a,j} n_{\mathfrak{e}}(a, j)(\rho_\pi - f_{\mathsf{cong}}(a, j)\mu_a) + c\sqrt{T \log\left(\frac{1}{\delta}\right)} + \Delta \ . \tag{11}$$

Proof. The reward $\tilde{r}(h_t^{\mathsf{alg}}, a_t)$ is a random reward received by the algorithm at time t. For any given pair of action a and history count j, let us denote by $N_T(a, j)$ the number of times this pair was observed in the run. By Hoeffding's inequality,

we have

$$\Pr\left(\sum_{t=1}^{T}\tilde{r}(h_t^{\mathsf{alg}},a_t)\leq\sum_{a,j}N_T(a,j)f_{\mathsf{cong}}(a,j)\mu_a-\sqrt{\frac{T}{2}\log\left(\frac{1}{\delta}\right)}\mid(N_T(a,j))_{a,j}\right)\leq\delta\ ,$$

where the expectation is taken conditioning on the numbers of plays $(N_T(a,j))_{a,j}$. Noting that $\sum_{\mathfrak{e}} n_{\mathfrak{e}}(a,j)=N_T(a,j)$ completes the proof. □

As in Section 2, let us denote the regret in each episode $\mathfrak{e}$ by

$$\mathfrak{r}_{\mathfrak{e}}:=\sum_{a,j}n_{\mathfrak{e}}(a,j)(\rho_\pi-f_{\mathsf{cong}}(a,j)\mu_a).$$

Going forward, we split our analysis into two cases depending on whether the true reward r^* belongs to the set $\mathcal{R}_{\mathfrak{e}}$ or not for each episode $\mathfrak{e}$.

Case 1: $r^*\notin\mathcal{R}_{\mathfrak{e}}$**.** Let us denote by $t_{\mathfrak{e}}$ the start time of episode $\mathfrak{e}$, initialized as $t_1=1$. We will study the effect of those episodes on regret when the true rewards are not contained within the confidence bounds imposed by our algorithm. For episode e, denote by $\mathbb{I}[r\notin\mathcal{R}_{\mathfrak{e}}]$, where the set $\mathcal{R}_{\mathfrak{e}}$ is the set of possible reward values available for epoch e. Then, the sum of these terms for each episode,

$$\begin{aligned}\sum_{\mathfrak{e}=1}^{\mathcal{E}}\mathfrak{r}_{\mathfrak{e}}\mathbb{I}[r\notin\mathcal{R}_{\mathfrak{e}}]&\overset{\text{(i)}}{\leq}\sum_{\mathfrak{e}=1}^{\mathcal{E}}\sum_{a,j}n_{\mathfrak{e}}(a,j)\mathbb{I}[r\notin\mathcal{R}_{\mathfrak{e}}]\\&\overset{\text{(ii)}}{\leq}\sum_{t=1}^{T}t\mathbb{I}[r\notin\mathcal{R}_{\mathfrak{e}}]\\&\leq\sqrt{T}+\sum_{t=T^{1/4}}^{T}t\mathbb{I}[r\notin\mathcal{R}_{\mathfrak{e}}]\ ,\end{aligned}\tag{12}$$

where (i) follows from the fact that $\rho^*\leq 1$, (ii) follows from the fact that $\sum_{a,j}n_{\mathfrak{e}}(a,j)\leq\sum_{a,j}N_{\mathfrak{e}}(a,j)=t_{\mathfrak{e}}-1$, and (iii) follows from splitting the time horizon into two parts. Now for any pair (a,j) with n samples, we have from an application of Hoeffding's inequality,

$$\Pr\left(|r^*(a,j)-\hat{r}_n(a,j)|\geq\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\right)\leq\delta\ ,$$

taking a union bound over all $n=\{1,\dots,t-1\}$ and (a,j) pairs, we have

$$\Pr\left(\forall(a,j),\ |r^*(a,j)-\hat{r}_n(a,j)|\geq\sqrt{\frac{\log(\frac{2K\Delta t^3}{\delta})}{2\max(N_{\mathfrak{e}}(a,j),1)}}\right)\leq\frac{\delta}{t^2}\ ,$$

where recall $N_{\mathfrak{e}}(a,j)=\sum_{i<\mathfrak{e}}n_{\mathfrak{e}}(a,j)$. Summing the above bound over all t from $t=T^{1/4}$, we have

$$\Pr\left(\forall(a,j),t\in[T^{\frac{1}{4}},T]\ |r^*(a,j)-\hat{r}_n(a,j)|\geq\sqrt{\frac{\log(\frac{8K\Delta t^3}{\delta})}{2\max(N_{\mathfrak{e}}(a,j),1)}}\right)\leq\delta\ ,$$

Combining the above with equation (12), we have, with probability at least $1-\delta$,

$$\sum_{\mathfrak{e}=1}^{\mathcal{E}}\mathfrak{r}_{\mathfrak{e}}\mathbb{I}[r\notin\mathcal{R}_{\mathfrak{e}}]\leq\sqrt{T}\ .\tag{13}$$

Case 2: $r^* \in \mathcal{R}_{\mathfrak{e}}$**.** We now look at the regret $\mathfrak{r}_{\mathfrak{e}}$ in episodes where the the true reward function r^* belongs to the set $\mathcal{R}_{\mathfrak{e}}$. The following upper bounds this regret in terms of the diameter Δ of the MDP as well as an additional term depending on the number of times each action count pair (a, j) is played.

Lemma 4. *For episodes where the reward vector $r \in \mathcal{R}_{\mathfrak{e}}$, with probability at least $1-\delta$, we have*

$$\mathfrak{r}_{\mathfrak{e}} \le \Delta + c\sqrt{\log\left(\frac{\Delta A t_{\mathfrak{e}}}{\delta}\right)} \sum_{a,j} \frac{n_{\mathfrak{e}}(a,j)}{\sqrt{\max(1, N_{\mathfrak{e}}(a,j))}} \,. \tag{14}$$

We defer the proof this technical lemma to Appendix A.2.2. Taking the above as given, we proceed with the proof of the main theorem. We now sum over all the episodes to obtain the final bound for the episodes when the confidence interval holds,

$$\begin{aligned}
\sum_{\mathfrak{e}} \mathfrak{r}_{\mathfrak{e}} \mathbb{I}[r \in \mathcal{R}_{\mathfrak{e}}] &\le \Delta\mathcal{E} + c\sqrt{\log\left(\frac{\Delta A T}{\delta}\right)} \sum_{a,j} \sum_{\mathfrak{e}} \frac{n_{\mathfrak{e}}(a,j)}{\sqrt{\max(1, N_{\mathfrak{e}}(a,j))}} \\
&\overset{\text{(i)}}{\le} D\mathcal{E} + c\sqrt{\log\left(\frac{\Delta A T}{\delta}\right)} \sum_{a,j} \sqrt{N(a,j)} \\
&\overset{\text{(ii)}}{\le} D\mathcal{E} + c\sqrt{\log\left(\frac{\Delta A T}{\delta}\right)} \cdot \sqrt{A\Delta T} \,,
\end{aligned} \tag{15}$$

where the inequality (i) follows from using the inequality[4]

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \le (\sqrt{2}+1)\sqrt{Z_n} \,, \tag{16}$$

and finally (ii) follows from an application of Cauchy-Schwarz inequality. We next need to bound the number of total episodes $\mathcal{E}$.

Bounding number of episodes. Denote by $N_T(a,j)$ the total number of times (a,j) was played in the entire run up to time T and by $m(a,j)$ the number of episode where the termination condition was satisfied for (a,j). Then, we have

$$N_T(a,j) = \sum_{\mathfrak{e}} n_{\mathfrak{e}}(a,j) \ge 1 + \sum_{i=1}^{m(a,j)} 2^i = 2^{m(a,j)} \,.$$

Noting that $T = \sum_{a,j} N_T(a,j)$, we have,

$$T \ge K\Delta \cdot \left(\frac{1}{K\Delta} \sum_{a,j} 2^{m(a,j)}\right) \ge K\Delta \cdot 2^{\frac{1}{K\Delta}\sum_{a,j} m(a,j)} \ge K\Delta 2^{\frac{\mathcal{E}}{K\Delta}-1} \,,$$

where the last inequality follows from the fact that $\mathcal{E} \le K\Delta + \sum_{a,j} m(a,j)$ with the first factor $K\Delta$ accounting for the time steps when $n_{\mathfrak{e}}(a,j) = 0$. Taking logarithm on both sides and simplifying, we get,

$$\mathcal{E} \le K\Delta + c \cdot \Delta K \log\left(\frac{T}{\Delta K}\right) \,. \tag{17}$$

The final bound now follows from combining the bounds in equations (10), (13), (15), and (29) so that with probability at least $1-\delta$, we have,

$$\mathfrak{R}_T(\textsc{CarMab}; \Pi, f_{\mathsf{cong}}) \le c \cdot \Delta^2 K \log\left(\frac{T}{\Delta K}\right) + c\sqrt{K\Delta T \log\left(\frac{\Delta K T}{\delta}\right)} + c\sqrt{T \log\left(\frac{1}{\delta}\right)} + \Delta \tag{18}$$

[4]For a proof of this, see Jaksch et al. (2010, Lemma 19).

A.2.1. Proof of Lemma 2

Any policy $\pi \in \Pi$ will eventually form a cycle in the state space S because of the finiteness of the state space. Let τ^* denote the time at which policy π^* begins its cycle. Then, the total reward of this optimal policy can then be decomposed into two parts

$$R_{\pi^*} = R_{\pi^*,1:\tau} + R_{\pi^*,\tau+1:T}. \tag{19}$$

By the optimality of ρ^*, we have that $R_{\pi^*,\tau+1:T} \leq \rho^*(T-\tau)$. For the first term, observe that states observed in time $1:\tau$ are unique since $\tau+1$ is the first time the policy started its cycle. Since the diameter of the MDP is at most Δ, we have

$$R_{\pi^*,1:\tau} \leq \tau\rho^*(\tau+\Delta)\ , \tag{20}$$

since we can go back to the start state at time $t=1$ from $t=\tau$ in at most Δ more steps. Since the reward is bounded by 1, the desired claim follows. □

A.2.2. Proof of Lemma 4

We will introduce some notation for this proof. Let $N_{\mathfrak{e}}(a,j) = \sum_{i=1}^{\mathfrak{e}-1} n_i(a,j)$ the number of times the action a was selected when it had a history of j counts in the episodes preceding $\mathfrak{e}$. Further, denote by $P_{\mathfrak{e}}$ the transition matrix induced by the policy $\pi_{\mathfrak{e}}$ selected in episode $\mathfrak{e}$ and the vector $n = \text{vec}(n_{\mathfrak{e}}(a,j))$. Let $\tilde{\mu}_a$ denote the mean optimistic reward obtained for arm a in this episode. The regret in episode $\mathfrak{e}$ is

$$\begin{aligned}
\mathfrak{r}_{\mathfrak{e}} &\leq \sum_{a,j} n_{\mathfrak{e}}(a,j)(\tilde{\rho}_{\mathfrak{e}} - f_{\mathsf{cong}}(a,j)\tilde{\mu}_a) + \sum_{a,j} n_{\mathfrak{e}}(a,j) f_{\mathsf{cong}}(a,j)(\tilde{\mu}_a - \mu_a) \\
&\overset{\text{(i)}}{\leq} n_{\mathfrak{e}}^\top (P_{\mathfrak{e}} - I)\lambda_{\mathfrak{e}} + \sum_{a,j} n_{\mathfrak{e}}(a,j) f_{\mathsf{cong}}(a,j)(\tilde{\mu}_a - \mu_a) \\
&\overset{\text{(ii)}}{\leq} n_{\mathfrak{e}}^\top (P_{\mathfrak{e}} - I) w_{\mathfrak{e}} + \sum_{a,j} n_{\mathfrak{e}}(a,j) f_{\mathsf{cong}}(a,j)(\tilde{\mu}_a - \mu_a) \\
&= \sum_{t=t_{\mathfrak{e}}}^{t_{e+1}-1} (\mathbf{e}_{s_{t+1}} - \mathbf{e}_{s_t}) w_{\mathfrak{e}} + \sum_{a,j} n_{\mathfrak{e}}(a,j) f_{\mathsf{cong}}(a,j)(\tilde{\mu}_a - \mu_a) \\
&\leq w_{\mathfrak{e}}(s_{t_{e+1}}) - w_{\mathfrak{e}}(s_{t_{\mathfrak{e}}}) + \sum_{a,j} n_{\mathfrak{e}}(a,j) f_{\mathsf{cong}}(a,j)(\tilde{\mu}_a - \mu_a) \\
&\overset{\text{(iii)}}{\leq} \Delta + \sum_{a,j} n_{\mathfrak{e}}(a,j) f_{\mathsf{cong}}(a,j)(\tilde{\mu}_a - \mu_a)
\end{aligned} \tag{21}$$

where inequality (i) follows from the Poisson equation $\lambda = r - \rho 1 + P\lambda$ and (ii) follows from denoting by $w_{\mathfrak{e}}(s) = \lambda_{\mathfrak{e}}(s) - \frac{\min \lambda_{\mathfrak{e}}(s) + \max \lambda_{\mathfrak{e}}(s)}{2}$, and (iii) follows from noting that $\|w_{\mathfrak{e}}\|_\infty \leq \frac{\Delta}{2}$ because the value function of optimal policy has width Δ. We now focus on the second term in the expression above. Since both $\tilde{\mu}_a$ and μ_a belong to the set $\mathcal{R}_{\mathfrak{e}}$, we have

$$\begin{aligned}
\mathfrak{r}_{\mathfrak{e}} &\leq D + c\sum_{a,j} n_{\mathfrak{e}}(a,j) f_{\mathsf{cong}}(a,j) \sqrt{\frac{\log(\Delta A t_{\mathfrak{e}}/\delta)}{\max(1, N_{\mathfrak{e}}(a,j))}} \\
&\overset{\text{(i)}}{\leq} D + c\sqrt{\log\left(\frac{\Delta A t_{\mathfrak{e}}}{\delta}\right)} \sum_{a,j} \frac{n_{\mathfrak{e}}(a,j)}{\sqrt{\max(1, N_{\mathfrak{e}}(a,j))}}
\end{aligned} \tag{22}$$

where (i) follows from using the fact that $f_{\mathsf{cong}} \in [0,1]$. □

A.3. Extension to routing on graphs

In this section, we study the an extension of the congested MAB setup where the arms correspond to edges on graph $G = (V, E)$ with a pre-defined start state s_{G} and goal state t_{G}. Going froward, we use the V and E to denote both the edge and vertex sets as well as their sizes. We recap some of the notation and setup introduced in Section 2.3.

On round $t = 1, \ldots, T$,

Algorithm 3: CARMAB-ST: Congested multi-armed bandits

Input: Congestion window Δ, confidence parameter $\delta \in (0,1)$, graph $G=(V,E)$, time horizon T, start state s_{G} and goal state t_{G}.

Initialize: Set $t=1$

for *episodes* $\mathfrak{e}=1,\dots,\mathcal{E}$ **do**

Initialize episode

Set start time of episode $t_{\mathfrak{e}}=t$.

For all edges e and historical count j, set $n_{\mathfrak{e}}(e,j)=0$ and $N_{\mathfrak{e}}(e,j)=\sum_{s<e} n_s(e,j)$.

Set empirical reward estimate for each edge and historical count

$$\hat{r}_{\mathfrak{e}}(e,j)=\frac{\sum_{\tau=1}^{t_{\mathfrak{e}}-1} r_\tau \cdot \mathbb{I}[e\in p_\tau, j_\tau=j]}{\max\{1,N_{\mathfrak{e}}(e,j)\}}.$$

Compute optimistic policy

Set the feasible rewards

$$\mathcal{R}_{\mathfrak{e}}=\left\{ r\in[0,1]^{E\times(\Delta+1)} \mid \text{for all } (e,j), |r(e,j)-\hat{r}_{\mathfrak{e}}(e,j)|\le c\sqrt{\frac{\log(LE\Delta t_{\mathfrak{e}}/\delta)}{\max\{1,N_{\mathfrak{e}}(a,j)\}}}\right\}$$

Find optimistic policy $\tilde{\pi}_{\mathfrak{e}}=\arg\max_{\pi\in\Pi, r\in\mathcal{R}_{\mathfrak{e}}}\rho(\pi,r)$.

Execute optimistic policy

while $n_{\mathfrak{e}}(a,j)<\max\{1,N_{\mathfrak{e}}(a,j)\}$ **do**

Select path $p_t=\tilde{\pi}_{\mathfrak{e}}(s_t)$, obtain reward $\hat{r}_t$.

Update $n_{\mathfrak{e}}(e,\#(s_t,e))=n_{\mathfrak{e}}(e,\#(s_t,e))+1$ for each edge on path p_t.

Set $t=t+1$.

- learner selects an s_{G}-t_{G} path p_t on the graph G
- learner observes reward $\tilde{r}(h_t,e_{i,t})=f_{\mathsf{cong}}(e_{i,t},\#(h_t,e_{i,t}))\cdot\mu_{e_{i,t}}+\epsilon_t$ for each $e_{i,t}$ on path p_t
- history changes to $h_{t+1}=[h_{t,2:\Delta},p_t]$

In comparison to the multi-armed bandit protocol, the learner here selects an s_{G}-t_{G} path on the graph G, the history at any time t consists of the entire set of paths $\{p_{t-\Delta},\dots,p_{t-1}\}$, and we assume that the congestion function on each edge $f_{\mathsf{cong}}(h,e)$ depends on the number of times this edge has been used in the past Δ time steps. Our algorithm for this setup extends the CARMAB algorithm to now consider paths over the graph G instead of arms. The detailed algorithm is presented in Algorithm 3.

For this setup, we denote by N_{st} the total number of path with start state s_{G} and goal state t_{G} and assume that each path is of length L. This is without any loss of generality since we can augment the graph G with edges which have zero reward by an additional L edges. Our main result in this section is the following regret bound on the above modified algorithm.

Theorem 4 (Regret bound for CARMAB-ST)**.** *For any confidence $\delta\in(0,1)$, congestion window $\Delta>0$ and time horizon T, the policy regret, of* CARMAB-ST *is*

$$\mathfrak{R}_T(\textsc{Carmab-st};\Pi,f_{\mathsf{cong}})\le c\cdot L\Delta^2 E\log\left(\frac{LT}{\Delta K}\right)+c\sqrt{LE\Delta T\log\left(\frac{LE\Delta T}{\delta}\right)}+c\sqrt{TL\log\left(\frac{1}{\delta}\right)} \tag{23}$$

with probability at least $1-\delta$.

Proof. We begin by constructing the MDP $\mathcal{M}_{\mathsf{st}}$ whose state space comprises all possible histories of s-t paths, that is, $S=h_\Delta$ with its size $|S|=N_{\mathsf{st}}^{\Delta}$ and the action set A consists of all the s-t paths. As before, observe that the diameter of this MDP is Δ. Following a similar argument as in Lemma 2 and 3, we have,

$$\mathfrak{R}_T(\mathsf{alg};\Pi,f_{\mathsf{cong}})\le\sup_{\pi\in\Pi}\sum_{\mathfrak{e}=1}^{\mathcal{E}}\sum_{e,j}n_{\mathfrak{e}}(e,j)(\rho_{L,\pi}-f_{\mathsf{cong}}(e,j)\mu_e)+c\sqrt{TL\log\left(\frac{1}{\delta}\right)}+\Delta\,, \tag{24}$$

where the above inequality follows from noting that $TL = \sum_{\mathfrak{e}} \sum_{e,j} n_{\mathfrak{e}}(e,j)$ and the additional L factor in the second comes in because now the total reward for a path p_t can be between $[0, L]$. Note that we have denote by $\rho_{L,\pi} = \rho_\pi / L$.

let us denote the regret in each episode $\mathfrak{e}$ by

$$\mathfrak{r}_{\mathfrak{e}} := \sum_{e,j} n_{\mathfrak{e}}(e,j)(\rho_{L,\pi} - f_{\mathsf{cong}}(e,j)\mu_e).$$

As before, we split our analysis into two cases depending on whether the true reward r^* belongs to the set $\mathcal{R}_{\mathfrak{e}}$ or not for each episode $\mathfrak{e}$.

Case 1: $r^* \notin \mathcal{R}_{\mathfrak{e}}$**.** Let us denote by $t_{\mathfrak{e}}$ the start time of episode $\mathfrak{e}$, initialized as $t_1 = 1$. We will study the effect of those episodes on regret when the true rewards are not contained within the confidence bounds imposed by our algorithm. For episode e, denote by $\mathbb{I}[r \notin \mathcal{R}_{\mathfrak{e}}]$. Then, the sum of these terms for each episode,

$$\begin{aligned}
\sum_{\mathfrak{e}=1}^{\mathcal{E}} \mathfrak{r}_{\mathfrak{e}} \mathbb{I}[r \notin \mathcal{R}_{\mathfrak{e}}] &\overset{\text{(i)}}{\leq} \sum_{\mathfrak{e}=1}^{\mathcal{E}} \sum_{e,j} n_{\mathfrak{e}}(e,j) \mathbb{I}[r \notin \mathcal{R}_{\mathfrak{e}}] \\
&\overset{\text{(ii)}}{\leq} L \sum_{t=1}^{T} t \mathbb{I}[r \notin \mathcal{R}_{\mathfrak{e}}] \\
&\leq L\sqrt{T} + L \sum_{t=T^{1/4}}^{T} t \mathbb{I}[r \notin \mathcal{R}_{\mathfrak{e}}] \,, \qquad (25)
\end{aligned}$$

where (i) follows from the fact that $\rho^* \leq 1$, (ii) follows from the fact that $\sum_{e,j} n_{\mathfrak{e}}(e,j) \leq \sum_{e,j} N_{\mathfrak{e}}(e,j) = L(t_{\mathfrak{e}} - 1)$, and (iii) follows from splitting the time horizon into two parts. Now for any pair (e,j) with n samples, we have from an application of Hoeffding's inequality,

$$\Pr\left(|r^*(e,j) - \hat{r}_n(e,j)| \geq \sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \right) \leq \delta \,,$$

taking a union bound over all $n = \{1, \ldots, L(t-1))$ and (e,j) pairs, we have

$$\Pr\left(\forall (e,j),\ |r^*(a,j) - \hat{r}_n(e,j)| \geq \sqrt{\frac{\log(\frac{2LE\Delta t^3}{\delta})}{2\max(N_{\mathfrak{e}}(a,j),1)}} \right) \leq \frac{\delta}{t^2} \,,$$

where recall $N_{\mathfrak{e}}(a,j) = \sum_{i<\mathfrak{e}} n_{\mathfrak{e}}(a,j)$. Summing the above bound over all t from $t = T^{1/4}$, we have

$$\Pr\left(\forall (a,j), t \in [T^{\frac{1}{4}}, T]\ |r^*(a,j) - \hat{r}_n(a,j)| \geq \sqrt{\frac{\log(\frac{8LE\Delta t^3}{\delta})}{2\max(N_{\mathfrak{e}}(a,j),1)}} \right) \leq \delta \,,$$

Combining the above with equation (12), we have, with probability at least $1 - \delta$,

$$\sum_{\mathfrak{e}=1}^{\mathcal{E}} \mathfrak{r}_{\mathfrak{e}} \mathbb{I}[r \notin \mathcal{R}_{\mathfrak{e}}] \leq L\sqrt{T} \,. \qquad (26)$$

Case 2: $r^* \in \mathcal{R}_{\mathfrak{e}}$**.** We now look at the regret $\mathfrak{r}_{\mathfrak{e}}$ in episodes where the the true reward function r^* belongs to the set $\mathcal{R}_{\mathfrak{e}}$. The following upper bounds this regret in terms of the diameter Δ of the MDP as well as an additional term depending on the number of times each action count pair (a,j) is played.

Lemma 5. *For episodes where the reward vector $r \in \mathcal{R}_{\mathfrak{e}}$, with probability at least $1 - \delta$, we have*

$$\mathfrak{r}_{\mathfrak{e}} \leq L\Delta + c\sqrt{\log\left(\frac{LE\Delta t_{\mathfrak{e}}}{\delta}\right)} \sum_{a,j} \frac{n_{\mathfrak{e}}(a,j)}{\sqrt{\max(1, N_{\mathfrak{e}}(a,j))}} \,. \qquad (27)$$

The proof of the above lemma follows exactly the same as Lemma 4 with the only caveat that the scale of rewards can now be L. We now sum over all the episodes to obtain the final bound for the episodes when the confidence interval holds,

$$\begin{aligned}\sum_{\mathfrak{e}} \mathfrak{r}_{\mathfrak{e}} \mathbb{I}[r \in \mathcal{R}_{\mathfrak{e}}] &\leq L\Delta\mathcal{E} + c\sqrt{\log\left(\frac{LE\Delta T}{\delta}\right)} \sum_{e,j} \sum_{\mathfrak{e}} \frac{n_{\mathfrak{e}}(e,j)}{\sqrt{\max(1, N_{\mathfrak{e}}(e,j))}} \\ &\overset{\text{(i)}}{\leq} L\Delta\mathcal{E} + c\sqrt{\log\left(\frac{LE\Delta T}{\delta}\right)} \sum_{e,j} \sqrt{N(e,j)} \\ &\overset{\text{(ii)}}{\leq} L\Delta\mathcal{E} + c\sqrt{\log\left(\frac{LE\Delta T}{\delta}\right)} \cdot \sqrt{LE\Delta T}\,, \end{aligned} \tag{28}$$

where the inequality (i) follows from using the inequality

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \leq (\sqrt{2}+1)\sqrt{Z_n}\,,$$

and finally (ii) follows from an application of Cauchy-Schwarz inequality. We finally end the proof with a bound on the number of episodes $\mathcal{E}$.

Bounding number of episodes. Denote by $N_T(e,j)$ the total number of times (e,j) was played in the entire run up to time T and by $m(e,j)$ the number of episode where the termination condition was satisfied for (e,j). Then, we have

$$N_T(e,j) = \sum_{\mathfrak{e}} n_{\mathfrak{e}}(e,j) \geq 1 + \sum_{i=1}^{m(e,j)} 2^i = 2^{m(e,j)}\,.$$

Noting that $LT = \sum_{e,j} N_T(e,j)$, we have,

$$LT \geq E\Delta \cdot \left(\frac{1}{E\Delta} \sum_{e,j} 2^{m(e,j)}\right) \geq E\Delta \cdot 2^{\frac{1}{E\Delta}\sum_{e,j} m(e,j)} \geq E\Delta \cdot 2^{\frac{\mathcal{E}}{E\Delta}-1}\,,$$

where the last inequality follows from the fact that $\mathcal{E} \leq E\Delta + \sum_{a,j} m(a,j)$. Taking logarithm on both sides and simplifying, we get,

$$\mathcal{E} \leq E\Delta + c \cdot \Delta E \log\left(\frac{LT}{\Delta E}\right)\,. \tag{29}$$

Using the above with equations (24), (26), and (28), we have with probability at least $1-\delta$,

$$\mathfrak{R}_T(\mathsf{alg}; \Pi, f_{\mathsf{cong}}) \leq c \cdot L\Delta^2 E \log\left(\frac{LT}{\Delta E}\right) + c\sqrt{LE\Delta T \log\left(\frac{LE\Delta T}{\delta}\right)} + c\sqrt{TL\log\left(\frac{1}{\delta}\right)} + \Delta \tag{30}$$

This concludes the proof of the desired claim. □

B. Proofs for congested linear contextual bandits

We now focus on the proofs for our results stated in Section 3.

B.1. Proof of Proposition 1

Recall that the for the known context case, our notion of regret is defined as

$$\mathfrak{R}_T(\mathsf{alg}; \Pi_{\mathcal{X}}, f_{\mathsf{cong}}, \theta_*) := \sup_{\pi \in \Pi} \sum_{t=1}^{T} r(h_t^{\pi}, \pi(h_t^{\pi}, x_t), x_t; \theta_*) - \sum_{t=1}^{T} \tilde{r}(h_t^{\mathsf{alg}}, a_t, x_t; \theta_*)\,,$$

where the $\tilde{r}$ are the noisy reward observations received by the learner. Conditioned on the choice of actions, we have by Hoeffding's inequality with probability at least $1-\delta$,

$$\mathfrak{R}_T(\mathsf{alg}; \Pi, f_{\mathsf{cong}}, \theta_*) \leq \sup_{\pi \in \Pi} \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_*) - \sum_{t=1}^{T} r(h_t^{\mathsf{alg}}, a_t, x_t; \theta_*) + c\sqrt{T \log\left(\frac{K}{\delta}\right)} \,,$$

where we have used the assumption that each reward is bounded in $[0,1]$. Going forward, we focus on the first two terms in the regret upper bound. Note that for any policy $\pi \in \Pi$, we can decompose the regret into three terms,

$$\begin{aligned}
\tilde{\mathfrak{R}}_T(\mathsf{alg}; \Pi, f_{\mathsf{cong}}, \theta_*) = &\sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_*) - \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_t) \\
&+ \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_t) - \sum_{t=1}^{T} r(h_t^{\mathsf{alg}}, a_t, x_t; \theta_t) \\
&+ \sum_{t=1}^{T} r(h_t^{\mathsf{alg}}, a_t, x_t; \theta_t) - \sum_{t=1}^{T} r(h_t^{\mathsf{alg}}, a_t, x_t; \theta_*) \,,
\end{aligned} \tag{31}$$

where the decomposition evaluates these policies on θ_* as well as the the parameter θ_t chosen by the algorithm. Let us denote the three terms above as R_1, R_2, and R_3. We will no bound these terms separately.

Bound for terms R_1 and R_3. Observe that the parameter only affects the reward obtained at any time step t but does not affect the history h_t, which is only a function of the past actions played. Therefore, for any time t, we have

$$\begin{aligned}
r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_*) - r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_t) &= \langle \theta_* - \theta_t, \phi(x_t, \pi(h_t^\pi, x_t))\rangle f_{\mathsf{cong}}(a_t^\pi, \#(h_t^\pi, a_t^\pi))] \\
&\leq \|\theta_* - \theta_t\|_2 \cdot \|\phi(x_t, \phi(x_t, \pi(h_t^\pi, x_t)))\|_2 \\
&\leq \|\theta_* - \theta_t\| \,,
\end{aligned}$$

where $a_t^\pi = \pi(h_t^\pi, x_t)$ and the last inequality follows from our assumption $\|\phi(x,a)\|_2\| \leq 1$. Thus, for both terms R_1 and R_3, we have

$$\sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_*) - \sum_{t=1}^{T} r(h_t^\pi, \pi(h_t^\pi, x_t), x_t; \theta_t) \leq \Delta \sum_{\mathfrak{e}=1}^{\mathcal{E}} 2^{\mathfrak{e}} \|\theta_* - \theta_{\mathfrak{e}}\|_2 \,. \tag{32}$$

Recall from Algorithm 2 that the parameter $\theta_{\mathfrak{e}}$ is updated via a least-squares estimator with

$$\theta_{\mathfrak{e}} = \arg\min_{\theta} \sum_{i=1}^{t_{\mathfrak{e}}-1} (r_t - \langle \theta, \phi(x_t, a_t)\rangle f_{\mathsf{cong}}(h_t)[a_t])^2 = \Sigma_{\mathfrak{e}}^{-1} \left(\frac{1}{t_{\mathfrak{e}} - 1} \sum_{t=1}^{t_{\mathfrak{e}}-1} r_t \tilde{\phi}_t \right) \,,$$

where we have denoted by $\tilde{\phi}_t := \phi(x_t, a_t) f_{\mathsf{cong}}(h_t)$ and by $\Sigma_{\mathfrak{e}} = \frac{1}{t_{\mathfrak{e}}-1} \sum_{\tau < t_{\mathfrak{e}}} \tilde{\phi}_t \tilde{\phi}_t^\top$. The following lemma obtains a bound on the error in the estimation of the parameter θ_*.

Lemma 6. *Suppose that the covariance matrix Σ_e satisfies Assumption 1. Then, for any $\delta > 0$ and $t_e > cd$, the OLS estimate θ_e defined above satisfies*

$$\|\theta_{\mathfrak{e}} - \theta_*\|_2 \leq \frac{c}{\gamma} \sqrt{\frac{d}{t_{\mathfrak{e}} - 1} \cdot \log \frac{1}{\delta}} \,, \tag{33}$$

with probability at least $1-\delta$.

We defer the proof of this lemma to Appendix B.1.1. Taking a union bound over the number of episodes $\mathcal{E}$, we have for all episodes $\mathfrak{e}$ with probability at least $1-\delta$,

$$\|\theta_{\mathfrak{e}} - \theta_*\|_2 \leq \frac{c}{\gamma} \sqrt{\frac{d}{t_{\mathfrak{e}} - 1} \cdot \log \frac{\mathcal{E}}{\delta}} \,,$$

Substituting the above in equation (32), we get,

$$\begin{aligned}\Delta \sum_{\mathfrak{e}=1}^{\mathcal{E}} 2^{\mathfrak{e}} \|\theta_* - \theta_{\mathfrak{e}}\|_2 &\leq \Delta \cdot \frac{c}{\gamma} \sum_{\mathfrak{e}=1}^{\mathcal{E}} 2^{e} \sqrt{\frac{d}{t_e - 1} \log \frac{\mathcal{E}}{\delta}} \\ &\overset{\text{(i)}}{\leq} \frac{\sqrt{\Delta} c}{\gamma} \cdot \sum_{\mathfrak{e}=1}^{\mathcal{E}} \frac{\sqrt{d} \cdot 2^{\mathfrak{e}}}{2^{\mathfrak{e}/2}} \sqrt{\log \frac{\mathcal{E}}{\delta}} \\ &\overset{\text{(ii)}}{\leq} \frac{2\sqrt{\Delta} c}{\gamma} \cdot \frac{\sqrt{d} \cdot 2^{\frac{\mathcal{E}}{2}}}{\sqrt{2} - 1} \cdot \sqrt{\log \frac{\mathcal{E}}{\delta}} \end{aligned} \tag{34}$$

where inequality (i) follows from the fact that $t_{\mathfrak{e}} - 1 > \Delta 2^{\mathfrak{e}}$ and (ii) follows from summing up the bound over $\mathcal{E}$ episodes. In order to bound the number of episodes $\mathcal{E}$, observe that,

$$T = \sum_{\mathfrak{e}=1}^{\mathcal{E}} \Delta 2^{\mathfrak{e}} = \Delta \cdot 2(2^{\mathcal{E}+1} - 1)$$

from which it follows that $\mathcal{E} \leq \log(\frac{T}{2\Delta} + 1)$. Substituting the above bound in equation (34), we get,

$$\Delta \sum_{\mathfrak{e}=1}^{\mathcal{E}} 2^{\mathfrak{e}} \|\theta_* - \theta_{\mathfrak{e}}\|_2 \leq \frac{c}{\gamma} \cdot \sqrt{d(T + \Delta) \cdot \log \frac{\log(T)}{\delta}} \, . \tag{35}$$

Bound for term R_2. For term R_2, we note that the policy π_e is the maximizer for the parameter $\theta_{\mathfrak{e}}$ for time step steps $t \in I_{\mathfrak{e}} \setminus \{t_{\mathfrak{e}}, \ldots, t_{\mathfrak{e}} + \Delta\}$. Thus, for any epoch $\mathfrak{e}$,

$$\begin{aligned}\sum_{t=t_e}^{t_{e+1}-1} r(h_t^{\pi}, \pi(h_t^{\pi}, x_t), x_t; \theta_e) - r(h_t^{\mathsf{alg}}, a_t, x_t; \theta_e) &= \sum_{t=t_e}^{t_e+\Delta} r(h_t^{\pi}, \pi(h_t^{\pi}, x_t), x_t; \theta_e) - r(h_t^{\mathsf{alg}}, a_t, x_t; \theta_e) \\ &\quad + \sum_{t=t_e+\Delta+1}^{t_{e+1}-1} r(h_t^{\pi}, \pi(h_t^{\pi}, x_t), x_t; \theta_e) - r(h_t^{\mathsf{alg}}, a_t, x_t; \theta_e) \\ &\overset{\text{(i)}}{\leq} \Delta \, , \end{aligned} \tag{36}$$

where the inequality (i) follows by the optimality of the policy $\tilde{\pi}_e$ for θ_e as well as by the boundedness of the reward $|r| \leq 1$. Combining these bounds, we get ,

$$\mathfrak{R}_T(\mathsf{alg}; \Pi, f_{\mathsf{cong}}, \theta_*) \leq \frac{c}{\gamma} \cdot \sqrt{d(T + \Delta) \cdot \log \frac{\log(T)}{\delta}} + \Delta \log(T) + c\sqrt{T \log\left(\frac{K}{\delta}\right)} \, , \tag{37}$$

with probability at least $1 - \delta$. This concludes the proof. □

B.1.1. Proof of Lemma 6

For any episode e, the error OLS estimate θ_e from the true parameter θ_* can be bounded as

$$\begin{aligned}\|\theta_* - \theta_e\|_2 &= \|\theta_* - \Sigma_e^{-1} \sum_{t<t_e} r_t \tilde{\phi}_t\|_2 \\ &= \|\theta_* - \Sigma_e^{-1} \left(\frac{1}{t_e - 1} \sum_{t<t_e} \tilde{\phi}_t \tilde{\phi}_t^{\top} \theta_* \right) + \Sigma_e^{-1} \left(\frac{1}{t_e - 1} \sum_{t<t_e} \epsilon_t \tilde{\phi}_t \right) \|_2 \\ &= \|\Sigma_e^{-1} \left(\frac{1}{t_e - 1} \sum_{t<t_e} \epsilon_t \tilde{\phi}_t \right) \|_2 \\ &\overset{\text{(i)}}{\leq} \frac{c}{\gamma} \sqrt{\frac{\mathrm{tr}(\Sigma_e)}{t_e - 1} \cdot \log \frac{1}{\delta}} \, , \end{aligned} \tag{38}$$

where the final inequality holds with probability at least $1-\delta$ and follows from noting that the noise $\epsilon_t \sim \mathcal{N}(0,1)$ as well as an application of Chernoff's bound for Gaussian random variables. Using the fact that $\|\tilde{\phi}\|_2 \leq 1$ completes the proof of the statement. □

B.2. Proof of Theorem 3

Recall that in this setup, we assume that the learner has knowledge of the context distributions a priori and the algorithm in each epoch plans with respect to this distribution with

$$\tilde{\pi}_{\mathfrak{e}} = \arg\max_{\pi\in\Pi} \mathbb{E}[\sum_{t\in I_{\mathfrak{e}}\setminus\{t_{\mathfrak{e}},\ldots,t_{\mathfrak{e}}+\Delta\}} r(h_t^{\pi}, \pi(h_t^{\pi}, x_t), x_t; \theta_{\mathfrak{e}})] \,. \tag{39}$$

The regret analysis proceed as before in the unknown case and we need to obtain a bound on two terms corresponding to the deviation of this expectation from the observed samples. In order to do this, we will invoke concentration results for random process over Markov chains, in particular Lemma 1 state in the main text (Chung et al., 2012).

Recall that every policy $\pi \in \Pi$ is a mapping from history and contexts to a choice of action $a \in [K]$. For any fixed policy π, we can construct a Markov chain with state space comprising the all elements of the history set $\mathcal{H}_\Delta$. For any pair of states h', h in this chain, we have the transitions

$$P_\pi(h'|h) = \begin{cases} P_{\Phi(x)}(\pi(\Phi(x), h) = a) & \text{if } h' = [h_{2:\Delta}, a] \\ 0 & \text{o.w.} \end{cases} \,, \tag{40}$$

where we have used the notation $P_{\Phi(x)}$ to include the randomness in the sampling of the features $\Phi(x) = \{\phi(x,a)\}_a$. Given this representation, we can define the function

$$f_\pi(h) = \mathbb{E}_{\Phi(x)}[r(h, a, x; \theta)] \,, \tag{41}$$

where the action $a = \pi(\Phi(x), h))$. A key ingredient of our upper bound will be the mixing time of the policies π in their respective Markov chains. Formally, we define the mixing time for a Markov chain below.

Definition 3 (Mixing-time of Markov chain)**.** *For an ergodic discrete time Markov chain M, let d represent an arbitrary starting state distribution and let d^* denote the stationary distribution. The ϵ-mixing time $\tau_{\mathsf{mix}}(\epsilon)$ is defined as*

$$\tau_{\mathsf{mix}}(\epsilon) = \min\{t : \max_d \|dM^t - d^*\|_{TV} \leq \epsilon\}. \tag{42}$$

We restate the concentration bound from Lemma 1 in the main paper.

Lemma 7 (Theorem 3.1 in (Chung et al., 2012))**.** *Let M be an ergodic Markov chain with state space $[n]$ and stationary distribution d^*. Let $\tau_{\mathsf{mix}} = \tau_{\mathsf{mix}}(\epsilon)$ be its ϵ-mixing time for $\epsilon \leq \frac{1}{8}$. Let $(V_1, ..., V_t)$ denote a t-step random walk on M starting from an initial distribution d on $[n]$. Let $\mu = \mathbb{E}_{V\sim d^*}[f(V)]$ denote the expected reward over the stationary distribution and $X = \sum_i f(V_i)$ denote the sum of function on the random walk. There exists a universal constant $c > 0$ such that*

$$\Pr(|X - \mu t| \geq \delta\mu t) \leq c\|d\|_{d^*} \exp\left(\frac{-\delta^2\mu t}{72\tau_{\mathsf{mix}}}\right) \quad \textit{for } 0 \leq \delta < 1 \,, \tag{43}$$

where the norm $\|d\|_{d^*}^2 := \sum_i \frac{d_i^2}{d_i^*}$.

There are a couple of points to address here before we can directly apply this bound to our setup, namely, the bound on the norm $\|d\|_{d^*}$ as well as the ergodicity of the Markov chain M_π. We address both of them below.

Ergodicity of Markov chain M_π. There are a few possibilities because of which the Markov chain might become non-ergodic. One is the existence of a few transient states, the other being the existence of multiple components within which the chain has different stationary distributions. We address this by considering the minimal subset of states of the Markov chain on which a stationary distribution exists which has maximum average reward according to the function defined in equation (41). That is, for any policy π, we define the average reward

$$\mu_\pi := \sup_{S\subseteq\mathcal{H}_\Delta; M|_S \text{ is ergodic}} \mathbb{E}_{V\sim d_S^*}[f(V)] \,. \tag{44}$$

This alleviates the problem of ergodicity, but in this process might make the norm $\|d\|_{d^*}$ be potentially unbounded. We address this next.

Bounding the norm $\|d\|_d^*$. At the time of an epoch change when we switch to a new policy, we might land up in a state from which running the policy π_e might never reach the subset of states which maximizes the above definition. However, recall that our Markov decision process is actually communicating with a diameter of Δ; thus in Δ time steps, one can always reach a starting state which has non-zero mass under the distribution d^*_π. Also, it is easy to see that one can further select this start state, say $h_{0,\pi}$, such that $\|\mathbb{I}[h_{0,\pi}]\|^2_{d^*} \leq \frac{1}{K^\Delta}$.

With the above two modifications, we are now in a position to use the mixing theorem. We define the worst-case mixing time over policies in the set $\Pi_\mathcal{X}$ as

$$\tau^*_{\mathsf{mix}} := \max_{\pi \in \Pi_\mathcal{X}} \tau_{\mathsf{mix},\pi}(h_{0,\pi}) \ , \tag{45}$$

With this, for any epoch e with selected policy π_e, we have with probability at least $1-\delta$,

$$\begin{aligned} |\sum_{t \in I_e} \mathbb{E}_{\Phi(x_t)}[r(h_t, a_t, x_t; \theta_*] - \mathbb{E}_{h \sim d^*_{\pi_e}} \mathbb{E}_{\Phi(x)}[r(h, a_t, x_t; \theta_*]| &\leq c\sqrt{\Delta 2^e \cdot \tau^*_{\mathsf{mix}} \cdot \log\left(\frac{\|d\|_{d^*}}{\delta}\right)} \\ &\overset{\text{(i)}}{\leq} c\Delta\sqrt{2^e \tau^*_{\mathsf{mix}} \cdot \log\left(\frac{K}{\delta}\right)} \ , \end{aligned} \tag{46}$$

where the inequality (i) follows from the from the fact that there exists a starting state with $\|\mathbb{I}[h_{0,\pi}]\|^2_{d^*} \leq \frac{1}{K^\Delta}$.

With these bounds in place, we can now first upper bound the regret via an application of Hoeffding's inequality to have with probability at least $1-\delta$,

$$\mathfrak{R}_T(\mathsf{alg}; \Pi, f_{\mathsf{cong}}, \theta_*) \leq \sup_{\pi \in \Pi} \sum_{t=1}^T r(h^\pi_t, \pi(h^\pi_t, x_t), x_t; \theta_*) - \sum_{t=1}^T r(h^{\mathsf{alg}}_t, a_t, x_t; \theta_*) + c\sqrt{T \log\left(\frac{1}{\delta}\right)} \ . \tag{47}$$

Focusing on the rewards of the algorithm, we have

$$\begin{aligned} \sum_{t=1}^T r(h^{\mathsf{alg}}_t, a_t, x_t; \theta_*) = &\sum_{t=1}^T r(h^{\mathsf{alg}}_t, a_t, x_t; \theta_*) - \mathbb{E}_{\Phi(x_t)}[r(h^{\mathsf{alg}}_t, a_t, x_t; \theta_*)|\Phi(x_{1:t-1})] \\ &+ \sum_{t=1}^T \mathbb{E}_{\Phi(x_t)}[r(h^{\mathsf{alg}}_t, a_t, x_t; \theta_*)|\Phi(x_{1:t-1})] - \mathbb{E}_{h \sim d^*_\pi} \mathbb{E}_{\Phi(x)}[r(h, \pi_{\mathfrak{e}}(\Phi(x), h), x; \theta_*)] \\ &+ \sum_{t=1}^T \mathbb{E}_{h \sim d^*_{\pi_{\mathfrak{e}}}} \mathbb{E}_{\Phi(x)}[r(h, \pi_{\mathfrak{e}}(\Phi(x), h), x; \theta_*)] \end{aligned}$$

Observe that the first term can be upper bounded with probability $1-\delta$ via an application of Hoeffding's inequality on the rewards as

$$\sum_{t=1}^T r(h^{\mathsf{alg}}_t, a_t, x_t; \theta_*) - \mathbb{E}_{\Phi(x_t)}[r(h^{\mathsf{alg}}_t, a_t, x_t; \theta_*)|\Phi(x_{1:t-1})] \leq c \cdot \sqrt{\alpha_u T \log\left(\frac{K}{\delta}\right)} \ , \tag{48}$$

where α_u is an upper bound on the operator norm of the covariance and the $\log(K)$ term comes from union bound over the samplings of the K contexts. We can upper bound the second term by summing the bound from equation (46) and have with probability at least $1-\delta$,

$$\sum_{t=1}^T \mathbb{E}_{\Phi(x_t)}[r(h^{\mathsf{alg}}_t, a_t, x_t; \theta_*)|\Phi(x_{1:t-1})] - \mathbb{E}_{h \sim d^*_\pi} \mathbb{E}_{\Phi(x)}[r(h, \pi_{\mathfrak{e}}(\Phi(x), h), x; \theta_*)] \leq c \cdot \sqrt{\Delta \tau^*_{\mathsf{mix}} T \log\left(\frac{K \log(T)}{\delta}\right)} \ . \tag{49}$$

This leaves us with the final set of terms in the regret bound comparing the rewards at stationary distribution. For any policy π, we have,

$$\begin{aligned} &\mathbb{E}_{\Phi(x)}\left[\sum_{t=1}^T \mathbb{E}_{h \sim d^*_\pi}[r(h, \pi(\Phi(x), h), x; \theta_*)] - \mathbb{E}_{h \sim d^*_{\pi_{\mathfrak{e}}}}[r(h, \pi_{\mathfrak{e}}(\Phi(x), h), x; \theta_*)]\right] \\ &\quad \leq \frac{c}{\gamma} \cdot \sqrt{d(T+\Delta) \cdot \log \frac{\log(T)}{\delta}} + \Delta \log(T) \ , \end{aligned} \tag{50}$$

with probability at least $1-\delta$. The above follows from a calculation similar to the one performed for the known context case. To complete the proof, we need to obtain an upper bound on the minimum eigen value of the sample covariance matrix to show that Assumption 1 is indeed satisfied.

Bound on minimum eigenvalue. We now obtain a lower bound on the minimum eigenvalue of the sample covariance matrix $\frac{1}{n}\sum_{t=1}^{n}\phi_t\phi_t^\top$ for any sample size n. The vectors $\phi_t \in \mathbb{R}^d$ are obtained by the algorithm's choice and are used to perform the least squares update on the parameter $\theta_\mathfrak{c}$. At each time t, the algorithm observes K different vectors $\phi_t^i \sim \mathcal{N}(\mu_i, I)$ and selects one of them. What makes the problem challenging is that the selected samples ϕ_t are no longer independent – the algorithms choice at round t can depend on all previous observations.

For any unit vector $v \in \mathbb{R}^d$, we have

$$v^\top\left(\frac{1}{n}\sum_t \phi_t\phi_t^\top\right)v = \frac{1}{n}\sum_t (v^\top\phi_t)^2 \geq \frac{1}{n}\sum_t \inf_i[(v^\top\phi_t^i)^2]\,.$$

Notice that the last inequality makes the process independent across each time step since it only depends on the random samples $\{\phi_t^1,\ldots,\phi_t^K\}$. Let us denote the random variable $X_t^i = (v^\top\phi_t^i)^2$ and by $X_t = \inf_i[(v^\top\phi_t^i)^2]$. Observe that the worst-case scenario for any v is when $\langle v,\mu_i\rangle = 0$ for all $i\in[K]$. Now, each random variable X_t^i is sub-exponential with parameters $(\nu=2,\alpha=4)$ and satisfies

$$\mathbb{E}[\exp\left(\lambda X_t^i\right)] \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right) \quad \text{for all } \lambda \text{ s.t. } |\lambda| \leq \frac{1}{\alpha}\,.$$

Lets consider the moment generating function of the random variable X_t:

$$\mathbb{E}[\exp\left(\lambda X_t\right)] = \mathbb{E}[\exp(\lambda \min_i X_t^i)] \overset{\text{(i)}}{\leq} \mathbb{E}[\exp(\lambda X_t^i)] \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right) \quad \text{for all } \lambda \text{ s.t. } |\lambda| \leq \frac{1}{\alpha}\,,$$

where inequality (i) follows since from the fact that the X_t^i are all positive and the exponential is an increasing function. Thus, we have established that each random variable X_t is sub-exponential with parameters (ν,α). Using a standard concentration bound for concentration of sub-exponential variables, we have for any unit vector v,

$$\Pr\left(\frac{1}{n}\sum_t \inf_i[(v^\top\phi_t^i)^2] \geq 1 - \sqrt{\frac{8}{n}\log\frac{2}{\delta}}\right) \leq \delta\,,$$

where the above inequality holds for any $\delta\in(0,1)$ and $n > c\log\frac{1}{\delta}$. Now, using a standard covering number argument over the unit vectors v, we have

$$\inf_{v:\|v\|_2=1} v^\top\left(\frac{1}{n}\sum_t \phi_t\phi_t^\top\right)v \geq \inf_{v:\|v\|_2=1}\frac{1}{n}\sum_t\inf_i[(v^\top\phi_t^i)^2] \geq \alpha_l - c\alpha_l\sqrt{\frac{d}{n}\log\frac{1}{\delta}}\,, \tag{51}$$

with probability at least $1-\delta$ for any $\delta\in(0,1)$ and $n > c\cdot d\log\frac{1}{\delta}$.

Combining the bounds from equations (47), (48), (49), (50) and (51), we have for $T > cd\log(\frac{1}{\delta})$, with probability at least $1-\delta$,

$$\begin{aligned}\mathfrak{R}_T(\mathsf{alg};\Pi,f_{\mathsf{cong}},\theta_*) \leq c\cdot\sqrt{\alpha_u T\log\left(\frac{K}{\delta}\right)} + c\cdot\sqrt{\Delta\tau^*_{\mathsf{mix}}T\log\left(\frac{K\log(T)}{\delta}\right)} \\ + \frac{c}{\alpha_l}\cdot\sqrt{d(T+\Delta)\cdot\log\frac{\log(T)}{\delta}}\,.\end{aligned}$$

This concludes the proof of the theorem.

C. Additional experimental details

In this section we discuss further details on the experimental evaluation, specifically how we compute the episode policy of CARCB. The computation follows a dynamic program (DP) that computes the optimal policy on the θ parameter that the algorithm maintains. The DP computes the optimal reward starting from state (t, h_t), where t is the time step and h_t the previous history of length Δ. Recall that the algorithm maintains parameter θ and also has access to the contexts $x_{a,t}$. Initialization is done for the final time step and every possibly history h of length Δ as follows:

$$OPT(T, h) = \max_{a \in [K]} \frac{x_{a,T} \cdot \theta}{\#(h, a)}.$$

The recursive step is then:

$$OPT(t, \{a_{t-\Delta}, a_{t-\Delta-1}, \ldots, a_{t-1}\}) = \max_{a \in [K]} \frac{x_{a,t} \cdot \theta}{\#(h, a)} + OPT(t+1, \{a_{t-\Delta-1}, a_{t-\Delta-2}, a_{t-1}, a\}).$$

Note that running the DP with parameter θ_* and the correct noise $\epsilon_{t,a}$ as follows yields the optimal policy:

$$OPT(T, h) = \max_{a \in [K]} \frac{x_{a,T} \cdot \theta_*}{\#(h, a)} + \epsilon_{T,a}.$$

$$OPT(t, \{a_{t-\Delta}, a_{t-\Delta-1}, \ldots, a_{t-1}\}) = \max_{a \in [K]} \frac{x_{a,t} \cdot \theta_*}{\#(h, a)} + \epsilon_{t,a} + OPT(t+1, \{a_{t-\Delta-1}, a_{t-\Delta-2}, a_{t-1}, a\}).$$