Remote Contextual Bandits

Francesco Pase
University of Padova
Padova, Italy
pasefrance@dei.unipd.it

Deniz Gündüz
Imperial College London
London, UK
d.gunduz@imperial.ac.uk

Michele Zorzi
University of Padova
Padova, Italy
zorzi@dei.unipd.it

***Abstract*—We consider a remote contextual multi-armed bandit (CMAB) problem, in which the decision-maker observes the context and the reward, but must communicate the actions to be taken by the agents over a rate-limited communication channel. This can model, for example, a personalized ad placement application, where the content owner observes the individual visitors to its website, and hence has the context information, but must convey the ads that must be shown to each visitor to a separate entity that manages the marketing content. In this remote CMAB (R-CMAB) problem, the constraint on the communication rate between the decision-maker and the agents imposes a trade-off between the number of bits sent per agent and the acquired average reward. We are particularly interested in characterizing the rate required to achieve sub-linear regret. Consequently, this can be considered as a policy compression problem, where the distortion metric is induced by the learning objectives. We first study the fundamental information theoretic limits of this problem by letting the number of agents go to infinity, and study the regret achieved when Thompson sampling strategy is adopted. In particular, we identify two distinct rate regions resulting in linear and sub-linear regret behavior, respectively. Then, we provide upper bounds for the achievable regret when the decision-maker can reliably transmit the policy without distortion.**

***Index Terms*—Multi-Armed Bandit, Rate-Distortion Theory, Regret Bound.**

I. Introduction

In the last few years, synergies between machine learning (ML) and communication networks have attracted a lot of interest in the research community, thanks to the fruitful interplay of the two fields in emerging applications, from Internet of Things (IoT) to autonomous vehicles, and other edge services. In most of these applications, both the generated data and the processing power are distributed across a network of physically distant devices, thus a reliable communication infrastructure is pivotal to run ML algorithms that can leverage the collected distributed knowledge [1], [2]. To this end, many recent works have tried to redesign networks and to efficiently represent information to support distributed ML applications, where the activities of data collection, processing, learning and inference are performed in different geographical locations; and therefore, the corresponding learning algorithms must take into account limited communication, memory, and processing resources, while addressing privacy issues.

In contrast to our desire to gather more data and intelligence, available communication resources (bandwidth and power, in particular) are highly limited, and must be shared among many different devices and applications. This requires the design of highly communication-efficient distributed learning algorithms. Information theory, and in particular rate-distortion theory, have laid the fundamental limits of efficient data compression, with the aim to reconstruct the source signal with the highest fidelity [3]. However, in the aforementioned applications, the goal is often not to reconstruct the source signal, but to make inferences based on it. This requires *task-oriented compression*, filtering out the unnecessary information for the target application, and thus decreasing the number of bits that have to be transmitted over the communication channels. This approach should target the questions of *what* is the most useful information that has to be sent, and *how* to represent it, in order to meet the application requirements while consuming the minimum amount of network resources [4], [5].

Our goal in this paper is to theoretically investigate a contextual multi-armed bandit (CMAB) problem, in which the context information is available to a remote *decision-maker*, whereas the actions are taken by a remote entity, called the *controller*, controlling a multitude of agents, each with an independent context realization. We can assume that a limited communication link is available between the decision-maker and the controller at each round to communicate the intended actions. The controller must decide on the action to be taken by each agent based on the message received over the channel, while the decision-maker observes the rewards at each round, and updates its policy accordingly. This framework is described in Fig. 1.

This scenario can model, for example, a personalized ad placement application, where the content owner observes the individual visitors to its website, and hence has the context information, but must convey the ads that must be shown to each visitor to a separate entity that manages the marketing content. This will require communicating hundreds or thousands of ads to be placed at each round, chosen from a large set of possible ads, within the resource and delay constraints of the underlying communication channel, which is quantified as the number of bits available per agent, i.e., per visitor. Other practical examples include the training of a single policy with N parallel agents, speeding up convergence while collecting more data per unit time [6]. In this case a server, i.e., the decision-maker, stores a centralized policy that is updated at every round, while N parallel environment instances run locally at each agent. The link between the server and the parallel agents is the one analyzed in this work, whereas possible constraints on the state observation process will be considered in future work.

II. Related Work

Given the amount of data that is generated by machines, sensors and mobile devices, the design of distributed learning algorithms is a hot topic in the ML literature. These algorithms often impose communication constraints among agents, requiring the design of methods to allow efficient representation of messages to be exchanged. While rate-distortion theory deals with efficient lossy transmission of signals [3], in ML applications we typically do not need to reconstruct the underlying signal, but wish to make some inference based on it. These applications can be modeled through distributed hypothesis testing [7]–[9] and estimation [10], [11] problems under rate constraints.

There is a growing literature on multi-agent reinforcement learning (RL) with communication links [12]–[15]. These papers consider a multi-agent partially observable Markov decision process (POMDP), where the agents collaborate to resolve a specific task. In addition to the usual reward signals, agents can also benefit from the available communication links to better cooperate and coordinate their actions. It is shown that communication can help overcome the inherent non-stationarity of the multi-agent environment. Our problem can be considered as a special case of this general RL formulation, where the state (context) at each time is independent of the past states and actions. Moreover, we focus on a particular setting in which the communication is one-way, from the decision-maker that observes the context and the reward, towards the controller that takes the actions. This formulation is different from the existing results in the literature involving multi-agent multi-armed bandit (MAB). In [16], each agent can pull an arm and communicate with others. They do not consider the contextual case, and focus on a particular communication scheme, where each agent shares the index of the best arm according to its own experience. Another related formulation is proposed in [17], where a pool of agents collaborate to solve a common MAB problem with a rate-constrained communication channel from the agents to the server. In this case, agents observe their rewards and upload them to the server, which in turn updates the policy used to instruct them. In [18], the authors consider a partially observable CMAB scenario, where the agent has only partial information about the context. However, this paper does not consider any communication constraint, and the partial/noisy view of the context is generated by nature. Differently from the existing literature, our goal is to identify the fundamental information theoretic limits of learning with communication constraints in this particular scenario.

III. Problem Formulation

A. The Contextual Multi-Armed Bandit (CMAB) Problem

We consider N agents, which experience independent realizations of the same CMAB problem. The CMAB is a sequential decision game in which the environment imposes a probability distribution P_S over a set of contexts, or states, $\mathcal{S}$, which is finite in our case. The game proceeds in rounds, and at each round $h = 1, \dots, H$, a realization of the state $s_h^n \in \mathcal{S}$ is sampled from distribution P_S for each agent $n \in \mathcal{N} = \{1, \dots, N\}$, independently across time and agents. The decision-maker observes the states $\{s_h^n\}_{n=1}^N$, and chooses an action (or arm) $a_h^n \in \mathcal{A} = \{1, \dots, K\}$, for each agent, where K is the total number of available actions, with probability $\pi_h(a_h^n|s_h^n)$. Once the actions have been taken, the environment returns rewards for all the agents following independent realizations of the same reward process, $r_h^n = r(s_h^n, a_h^n) \sim P_R(r|s_h^n, a_h^n), \quad \forall n \in \mathcal{N}$, which depends on the state and the action of the corresponding agent. The policy $\pi_h(a_h^n|s_h^n)$ used to sample the actions is a mapping $\pi_h : \mathcal{H}^{h-1} \times \mathcal{S} \to \Delta_K$. The set $\mathcal{H}^{h-1}$ contains all possible observations of the decision-maker, and $H(h-1) \in \mathcal{H}^{h-1}$ represents the knowledge accumulated by all the agents up to round $h-1$, i.e., $H(h-1) = \left(\left(s_1^1, a_1^1, r_1^1\right), \dots, \left(s_{h-1}^N, a_{h-1}^N, r_{h-1}^N\right)\right) \in \mathcal{H}^{(h-1)}$. The set Δ_K is the K-dimensional simplex, containing all possible distributions over the set of actions. Based on the history of rewards up to round $h-1$, the decision-maker can optimize its policy to minimize the Bayesian system regret, that is defined as

$$\mathrm{BR}(H, \pi) = \mathbb{E}\left[\sum_{h=1}^{H} \sum_{n \in \mathcal{N}} \mu(s_h^n, a^*(s_h^n)) - \mu(s_h^n, a_h^n)\right], \quad (1)$$

where a_h^n is the action taken by agent n at round h using policy $\pi_h(a|s)$, which does not depend on n, i.e., the decision-maker adopts the same policy for all the agents, $\mu(s, a) = \mathbb{E}[r(s, a)]$ is the average reward of action a in state s, and $a^*(s) = \arg\max_{a \in \mathcal{A}} \mu(s, a)$ is the optimal action for state s, i.e., the action with the highest expected reward, which is unknown at the beginning. The expectation is taken with respect to the state, action, and problem instance distributions.

B. Remote CMAB

In our scenario, the process of observing the system states is spatially separated from the process of taking actions. The environment states, $\{s_h^n\}_{n=1}^N$, are observed by a central entity, i.e., the decision-maker, that has to communicate to the controller over a rate-limited communication channel, at each round h, the information about the actions $\{a_h^n\}_{n=1}^N$ the agents should take. Consequently, the problem is to communicate the action distribution, i.e., the policy $\pi_h(a|s)$, which depends on the specific state realizations, to the controller within the available communication resources.

Specifically, the decision-maker employs a function $f_h^{(N)} : \mathcal{H}^{h-1} \times \mathcal{S}^N \to \{1, 2, \dots, B\}$ to map the observed history and the N states at time h to a message index to be transmitted over the channel. The controller, on the other hand, employs a function $g_h^{(N)} : \{1, 2, \dots, B\} \to \mathcal{A}^N$ to map the received message to a set of actions for the agents. In general, both functions $f_h^{(N)}$ and $g_h^{(N)}$ can be stochastic. The Bayesian regret achieved by sequences $\left\{f_h^{(N)}, g_h^{(N)}\right\}_{h=1}^H$ is given by

$$\mathrm{BR}(H, (f, g)) = \mathbb{E}\left[\sum_{h=1}^{H} \sum_{n \in \mathcal{N}} r(s_h^n, a^*(s_h^n) - r(s_h^n, g_h^n(m_h))\right], \quad (2)$$

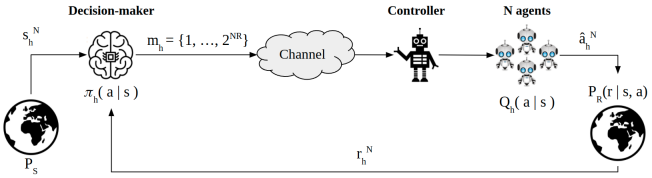

Fig. 1: The R-CMAB problem with a rate-limited communication channel.

where $g_h^n(m_h)$ is the action taken by agent n based on message $m_h = f_h^{(N)}\left(H(h-1), s_h^N\right)$ transmitted in round h, and $s_h^N \in \mathcal{S}^N$ is the vector containing the states of all the agents. We say that, for a given problem with N agents, a rate R is *achievable* if there exist functions $\left\{f_h^{(N)}, g_h^{(N)}\right\}_{h=1}^{H}$ as defined above with rate $\frac{1}{N}\log_2 B \leq R$ and regret

$$\lim_{H\to\infty} \frac{\text{BR}\left(H, \left\{f_h^{(N)}, g_h^{(N)}\right\}\right)}{H} = 0, \tag{3}$$

i.e., sub-linear in rounds.

If a rate $R \geq \log K$ is available, then the intended action for each agent can be easily conveyed to the controller, and a policy π_h that achieves sub-linear regret in the classical problem can achieve the same regret in the remote version (Thm. IV.6). However, in general, it may not be possible to convey the decision-maker's policy perfectly to the controller, and it is not clear whether distorted versions of the policy π can obtain sub-linear regret. If this is the case, it would be possible to reduce the necessary communication rate, while still solving the underlying learning problem, by *compressing* the policy π.

IV. Solution

We first split the problem of learning a policy π at the decision-maker, and of characterizing the required rate to convey it, when a fixed distortion between π and the policy adopted by the agents Q is allowed. We then study the problem exploiting Thompson sampling (TS), which is a popular strategy to efficiently solve MAB problems, and characterize the required asymptotic rate to solve the problem. We also provide an upper bound on the Bayesian system regret when the TS policy can be perfectly conveyed to the controller.

A. The Asymptotic Policy Rate

We model the environment as a discrete memoryless source (DMS), which generates at each round states from a finite alphabet $\mathcal{S}$ with probability P_S, emitting sequences of N symbols $s^N = (s_1, \ldots, s_N)$, one per agent. We then denote with $\hat{Q}_{s^N}(s)$ the empirical probability of state $s \in \mathcal{S}$ in s^N. We also consider the sequence of actions a^N, and denote with $\hat{Q}_{s^N a^N}(s,a)$ the empirical joint probability of the pair (s,a) in $((s_1,a_1),\ldots,(s_N,a_N))$. The whole picture can be seen in Fig. 1, where the actions taken by the agents are denoted by $\hat{a}$ to indicate that they can differ from a dictated by policy π. We assume that the distribution P_S is known (or accurately estimated).

The decision-maker can observe the realization s^N of the contexts, and its task is to transmit an index $m \in \{1,\ldots,B\}$ over the channel so that the controller can generate from m the actions a^N, where $\hat{Q}_{s^N a^N}$ is as close to $P_{SA}(s,a) = P_S(s)\pi(a|s)$ as possible, where closeness is defined in terms of a distortion measure $\mathbb{E}[d(\hat{Q}_{S^N A^N}, P_{SA})]$, which in general is not an average of a per-letter distortion measure. The problem is a compression task in which the decision-maker has knowledge of the states s^N, and wants to transmit a conditional probability distribution $\pi_{A|S}$ to the agents, consuming the minimum amount of bits, in such a way that the empirical distribution $\hat{Q}_{s^N a^N}$ is close to the joint distribution P_{SA} induced by the policy. For a distortion function $d(Q_{SA}, P_{SA})$ that is 1) nonnegative, 2) upper bounded by a constant D_{max}, 3) continuous in Q_{SA}, and 4) convex in Q_{SA}, in [19] the authors provide the rate-distortion function $R(D)$, i.e., the minimum rate $R = \frac{\log_2 B}{N}$ bits per symbol such that $\mathbb{E}[d(\hat{Q}_{S^N A^N}, P_{SA})] \leq D$, in the limit when N is arbitrarily large.

Theorem IV.1 ([19], Theorem 1)**.** *The rate-distortion function for the problem of communicating policies is*

$$R(D) = \min_{Q_{A|S}: d(Q_{SA}, P_{SA}) \leq D} I(S;A) \tag{4}$$

assuming the set of $Q_{A|S}$ satisfying $d(Q_{SA}, P_{SA}) \leq D$ is not empty.

Here $Q_{SA} = P_S Q_{A|S}$ is the joint probability induced by the environment distribution P_S and by policy $Q_{A|S}$, which depends on the information sent by the decision-maker. As we can see from Eq. (4), in the asymptotic limit of N agents, the problem admits a single-letter solution, which also serves as a lower bound on the finite agent scenario. When imposing $D = 0$, the needed rate is the mutual information between the states and actions, which are related by the policy π. Furthermore, if we allow $D > 0$, Eq. (4) characterizes the minimum rate needed to convey the actions to the controller. However, finding a closed form solution for the rate-distortion function is not a trivial task in general.

B. Thompson Sampling (TS)

In the proposed solution, the decision-maker adopts the TS strategy [20] to learn a policy. The reason why TS is adopted is because, among the state-of-the-art MAB solutions, it relies on posterior sampling [21], that can be exploited within one round to sample different actions in parallel across the N agents. If upper confidence bound (UCB)-based algorithms are used, they should be adapted to perform exploration within one round, given that the policy is deterministic. Consequently, the action probability distribution induced by TS is exploited in the R-CMAB problem to perform exploration in parallel, and to further compress the original policy using Eq. (4).

In particular, the decision-maker implements one TS instance for each state $s \in \mathcal{S}$. Indeed, in our general formulation, there is no known structure between the states and rewards to be exploited. Consequently, the decision-maker maintains estimates of the distributions $p_h^{s,a}(\mu)$ of the mean reward

$\mu(s,a) \in \mathbb{R}$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$. To take a decision in round h and state s_h, the decision-maker samples $\hat{\mu}_h(s_h,a) \sim p_h^{s_h,a}$, $\forall a \in \mathcal{A}$, and takes the action $a^* = \arg\max_{a\in\mathcal{A}}\{\hat{\mu}_h(s_h,a)\}$. This procedure is repeated for each agent $n \in \mathcal{N}$. After receiving the rewards $\{r_h^n\}_{n=1}^N$, the decision-maker can update its belief on $\mu(s,a)$, i.e., the probabilities $p_h^{s,a}(\mu)$, in order to minimize the regret. We notice that this strategy induces a probability distribution $\pi_h(a|s)$ over the actions that is $\pi_h(a|s) = \int_{\mathcal{D}} p_h^{s,a}(\mu) \prod_{j=1,j\neq a}^K P_h^{s,j}(\mu) d\mu$, where $P_h^{s,j}(\mu)$ is the cumulative distribution function (CDF) of $\mu(s,j)$, and the random variables $\mu(s,a)$ are considered independently distributed.

However, in our scenario, the constraint on the rate imposed by the communication channel can make it infeasible for the controller to sample the actions directly from the true distribution $\pi_h(a|s)$. The agents have to use a proxy $Q_h(a|s)$, which is the one obtained from the message received over the channel. This problem is similar to approximate TS, where a proxy distribution is used to sample the actions, or the reward means, given that the true one is too complex to sample from. In that case, the bottleneck is due to the complexity of sampling from the true mean reward distribution, whereas in this work it is imposed by the limited-rate communication channel.

C. Asymptotic Limit for the Achievable Rate

We adopt Assumption 1 in [22], that considers rewards to be distributed following canonical exponential families, and the priors used by TS to be bounded away from zero $\forall(s,a)$.

In the following, we provide the minimum rate needed to achieve sub-linear regret in all the states, $s \in \mathcal{S}$, exploiting the TS scheme explained above. We define $H(A^*)$ as the entropy of the optimal arm, which we assume unique, or uniquely determined within a set of optimal arms, and computed based on the marginal $\pi^*(a) = \sum_s P_S(s)\pi^*(a|s)$, where π^* is the optimal policy, and we prove that it is the minimum rate required.

We will use the following result from [23].

Theorem IV.2 ([23], Theorem 2)**.** *Suppose that the TS policy $\pi(a|s)$ achieves sub-linear regret in each state $s \in \mathcal{S}$, then*

$$\lim_{h\to\infty} \pi_h(a^*(s)|s) = 1 \quad a.s. \tag{5}$$

where

$$a^*(s) = \arg\max_{a\in\mathcal{A}} \mu(s,a).$$

We now provide the following lemma.

Lemma IV.3. *Assuming that Thompson Sampling policy $\pi_h(a|s)$ achieves sub-linear Bayesian system regret, then*

$$\lim_{h\to\infty} \boldsymbol{I}_{\pi_h}(S;A) = \lim_{h\to\infty} H_{\pi_h}(A) = H(A^*). \tag{6}$$

Sketch of the Proof. The proof follows from Theorem IV.2, whose consequence is that, in the limit, the entropy of the TS policy conditioned on state s is zero. By using this with the definition of the rate provided in Eq. (4), it is possible to conclude the proof. □

Theorem IV.2 and Lemma IV.3 are useful to prove the following results. Here the available rate R is considered fixed in all rounds $h = \{1,\dots,H\}$.

Lemma IV.4. *If $R < H(A^*)$, then it is not possible to convey a policy $Q(a|s)$ that achieves sub-linear Bayesian system regret.*

Sketch of the Proof. If $R < H(A^*)$, from Eq. (4), the policy Q conveyed to the controller will always have non-zero distortion $d(Q_{SA}, \pi^*_{SA}) = D > 0$ with respect to π^*. If we take, for example, the total variation as the distortion measure, in each round h, Q would sample a sub-optimal arm with constant probability of at least D. Consequently, sub-linear regret cannot be achieved. □

The following Lemma provides the achievability part.

Lemma IV.5. *If $R > H(A^*)$, then achieving sub-linear regret is possible in all states $s \in \mathcal{S}$, as $N \to \infty$.*

Sketch of the Proof. The intuition is that, even though during training the required rate R_h to convey the current policy may exceed R, exploration is never penalized (actually it is enforced by the system). Consequently, TS will converge to the optimal policy [24], that can be eventually perfectly transmitted to the controller, given that $R > H(A^*)$, which is the rate required in the limit as $N \to \infty$. This, together with the fact that TS achieves sub-linear regret in this parallel multi-agent version of the problem (Theorem IV.6), concludes the proof. □

The consequence of Lemma IV.5 is that, even if the exact TS policy π_h cannot be transmitted $\forall h$, as long as $R > H(A^*)$, it is still possible to achieve sub-linear regret. According to the definition in Eq. (3), this implies that, as $N \to \infty$, any rate $R > H(A^*)$ is achievable, while any rate $R < H(A^*)$ is not achievable.

D. Regret of the Optimal Policy

In this section, we present both finite-time and asymptotic upper bounds on the regret obtained by the TS strategy, when the policy π_h can be perfectly transmitted at each round h. We further provide the per-agent regret, defined as the one obtained by a single agent. However, to fairly compare the obtained regret with TS applied to the standard CMAB problem, we write it as a function of the virtual time-steps $t \in \{1,\dots,T\}$, with $T = NH$, i.e., it represents the total number of interactions the system has with the environment through the agents. Indeed, the problem is mathematically equivalent to a single-agent CMAB, in which the parallel interactions of the N agents are mapped onto a one-dimensional time line, with the additional constraint that the policy π can be updated only every N time-steps, i.e., at time-steps $t = Nh$.

Theorem IV.6 (Bayesian System Regret)**.** *The Bayesian system regret of TS is upper bounded by*

$$\boldsymbol{BR}(\pi,T) \leq 2K|S|N + 4\sqrt{(2+6\log T)\,KN|S|T}, \tag{7}$$

and the asymptotic regret is

$$BR(\pi, T) \in \mathcal{O}\left(\sqrt{KT|S|\log T}\right). \tag{8}$$

Sketch of the Proof. The proof follows similar arguments to those in [21], Section 6, with the difference that during each round h, the policy adopted by the N parallel agents is not sequentially optimized, but can be updated only at the end of the round. Consequently, a penalty of $\sqrt{N}$ appears on the upper bound of finite-time regret, as when T is small, playing with a sub-optimal policy N times in parallel amplifies the regret. The result follows from bounding the gap between the counter of the number of times a particular action has been sampled until t, and the counter at the end of the previous round, which is the value used to update the policy, and to construct the confidence bounds [21]. In the asymptotic case, i.e., $T >> N$, this effect vanishes, as the gap is almost N. □

Theorem IV.7 (Bayesian Agent Regret)**.** *The Bayesian per-agent regret of TS is upper bounded by*

$$BR(\pi, T) \leq 2K|S| + 4\sqrt{\frac{(2 + 6\log T)\, K|S|T}{N}}, \tag{9}$$

and the asymptotic regret is

$$BR(\pi, T) \in \mathcal{O}\left(\frac{1}{N}\sqrt{KT|S|\log T}\right). \tag{10}$$

Sketch of the Proof. The proof relies on Theorem IV.6, and on the observation that the per-agent regret is equal to $\mathrm{BR}(\pi, H, n) = \frac{\mathrm{BR}(\pi, H)}{N}$, due to the symmetry of the problem. Indeed, each agent interacts with an independent and identically distributed (i.i.d.) copy of the environment and, at each round h, adopts the policy $\pi_h(a|s)$ known by the decision-maker, and equal for all the agents $n \in \mathcal{N}$. □

V. Numerical results

In this experiment we analyze the asymptotic rate required by the TS policy to be conveyed, that serves as a lower bound for practical scenarios with finite N, in three different environments. In all the scenarios, there are 16 actions per state, and 16 states that are sampled uniformly by the environment. The first one is called *16 Groups*, and the reward behind arm a_j in state s_i is a Bernoulli random variable with parameter $\mu(s_i, a_j) = 0.8$ if $i = j$, whereas $\mu(s_i, a_j) \sim \mathrm{Unif}_{[0,0.75]}$ if $i \neq j$, with $i, j \in \{0, \ldots, 15\}$. The best action is thus strongly correlated with the state, and a sufficiently high rate is required to sample from the optimal policy π^*. In the second experiment, the setting is similar to the one presented above, but $\mu(s_i, a_j) = 0.8$ if $\lfloor \frac{j}{2} \rfloor = i$, and sampled uniformly in $[0, 0.75]$ otherwise. Consequently, the best actions can be grouped into 8 different classes. This scenario is indicated as the *8 Groups* experiment. The same procedure is applied to generate the last environment, except that the best responses are grouped into just 2 different classes.

Fig. 2 shows the theoretical rate needed to convey the TS policy in the three described scenarios, as a function of the number of rounds. It is possible to observe that the policy rates are converging to $4, 3, 1$ bits, respectively, which

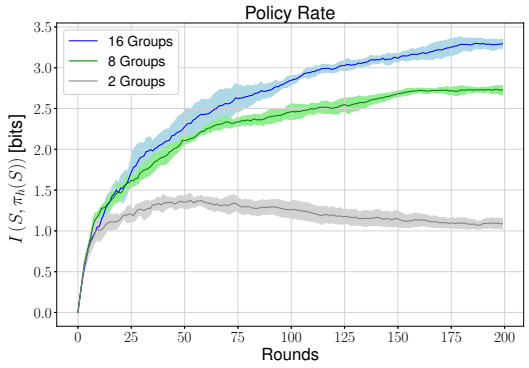

Fig. 2: Asymptotic rate needed to reliably transmit the decision-maker's policy. Curves are average values $\pm\sigma$, computed over 5 independent runs per scenario.

are the mutual information values between the states and optimal actions, i.e., the entropy of uniform distributions over the problem classes. We can also notice that, during the exploration phase at the beginning of the training process, very limited information has to be sent, whereas the required rate gradually increases as the decision-maker learns to map states to optimal responses. Interestingly in the *2 Groups* case, we can observe that during the training phase the algorithm requires a rate $R^\pi > H(A^*)$ to convey the policy π with zero error, before converging to the optimal π^*. However, by Lemma IV.5, we know we could potentially restrict the available rate to $H(A^*)$, and communicate the compressed policy during those rounds in which $R^\pi > H(A^*)$, while still achieving sub-linear regret.

VI. Conclusion

We have introduced and studied the R-CMAB problem, in which an intelligent entity, i.e., the decision-maker, observes the contexts of N parallel CMAB processes, and has to decide on the actions depending on the current contexts and the past actions and rewards. However, the actions are implemented by a controller that is connected to the decision-maker through a communication link. First, we cast the problem into the proper information-theoretic framework, and provided the needed rate to convey a policy, when admitting a maximum distortion between a compressed policy adopted by the controller and the one of the decision-maker. We then analyzed the problem when the TS algorithm is used, and characterized the minimum achievable rate to obtain sub-linear regret. In the end, we provided finite-time and asymptotic upper bounds on the system regret, when the policy can be conveyed to the controller. Ongoing work includes the formulation of the problem with specific distortion functions, which can be derived from the underlying learning objectives, and an analysis of the behavior when non-zero distortion is allowed.

VII. Acknowledgments

This work was supported by the European Research Council (ERC) Starting Grant BEACON (grant agreement no. 677854).

REFERENCES

[1] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[2] D. Gündüz, D. B. Kurka, M. Jankowski, M. M. Amiri, E. Ozfatura, and S. Sreekumar, "Communicate to learn at the edge," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 14–19, Jan. 2020.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

[4] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 89–100, Nov. 2021.

[5] T.-Y. Tung, S. Kobus, J. P. Roig, and D. Gündüz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2590–2603, Aug. 2021.

[6] A. V. Clemente, H. N. Castejón, and A. Chandra, "Efficient Parallel Methods for Deep Reinforcement Learning," *ArXiv e-prints*, May 2017.

[7] T. Berger, "Decentralized estimation and decision theory," in *IEEE 7th. Spring Workshop on Inf. Theory*, Mt. Kisco, NY, Sep. 1979.

[8] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. 32, no. 4, pp. 533–542, Jul. 1986.

[9] S. Sreekumar and D. Gündüz, "Distributed hypothesis testing over discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2044–2066, Apr. 2020.

[10] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems*, vol. 26, Dec. 2013.

[11] A. Xu and M. Raginsky, "Information-theoretic lower bounds on Bayes risk in decentralized estimation," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1580–1600, Mar. 2017.

[12] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," *arXiv:1605.06676 [cs]*, May 2016, arXiv: 1605.06676.

[13] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," in *Proc. of 30th Int'l Conf. on Neural Information Proc. Systems*, ser. NIPS'16, Red Hook, NY, Dec. 2016, pp. 2252–2260.

[14] S. Havrylov and I. Titov, "Emergence of language with multi-agent games: Learning to communicate with sequences of symbols," in *Advances in Neural Information Processing Systems*, Dec. 2017.

[15] A. Lazaridou, A. Peysakhovich, and M. Baroni, "Multi-agent cooperation and the emergence of (natural) language," *arXiv:1612.07182 [cs]*, Mar. 2017, arXiv: 1612.07182.

[16] M. Agarwal, V. Aggarwal, and K. Azizzadenesheli, "Multi-agent multi-armed bandits with limited communication," in *arXiv:2102.08462 [cs]*, 2021.

[17] O. A. Hanna, L. F. Yang, and C. Fragouli, "Solving multi-arm bandit using a few bits of communication," in *38th International Conference on Machine Learning*, 2021.

[18] H. Park and M. K. S. Faradonbeh, "Analysis of Thompson sampling for partially observable contextual multi-armed bandits," *arXiv:2110.12175 [stat.ML]*, 2021.

[19] G. Kramer and S. A. Savari, "Communicating probability distributions," *IEEE Transactions on Information Theory*, vol. 53, no. 2, pp. 518–525, Feb. 2007.

[20] W. R. Thompson, "On the theory of apportionment," *American Journal of Mathematics*, vol. 57, no. 2, pp. 450–456, Apr. 1935.

[21] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operation research*, vol. 39, no. 4, pp. 1221–1243, Nov. 2014.

[22] D. Russo, "Simple Bayesian algorithms for best arm identification," in *29th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 49. Columbia University, New York, New York, USA: PMLR, 23–26 Jun 2016, pp. 1417–1418.

[23] C. Kalkanli and A. Ozgur, "Asymptotic convergence of Thompson sampling," in *arXiv:2011.03917v1*, 2020.

[24] M. Phan, Y. Abbasi Yadkori, and J. Domke, "Thompson sampling with approximate inference," in *Advances in Neural Information Processing Systems*, Dec. 2019.