

Quantification of Transfer in Reinforcement Learning via Regret Bounds for Learning Agents

Adrienne Tuynman¹ and Ronald Ortner^{2*}

¹Scool team, Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189
CRIS^tAL, F-59000 Lille, France.

²Lehrstuhl für Informationstechnologie, Montanuniversität
Leoben, Austria.

*Corresponding author(s). E-mail(s): rortner@unileoben.ac.at;
Contributing authors: adrienne.tuynman@inria.fr;

Abstract

We present an approach for the quantification of the usefulness of transfer in reinforcement learning via regret bounds for a multi-agent setting. Considering a number of \aleph agents operating in the same Markov decision process, however possibly with different reward functions, we consider the regret each agent suffers with respect to an optimal policy maximizing its average reward. We show that when the agents share their observations the mutual regret of all agents is smaller by a factor of $\sqrt{\aleph}$ compared to the case when each agent has to rely on the information collected by itself. This result demonstrates how considering the regret in multi-agent settings can provide theoretical bounds on the benefit of sharing observations in transfer learning.

Keywords: transfer, reinforcement learning, regret

1 Introduction

Transferring relevant knowledge to apply it to a variety of similar environments is not a new concept: in psychology and cognitive sciences, it is well known that humans learn better when they are already familiar with a similar task. This concept helped form *transfer learning*, a wide and important area of machine learning with a lot of empirical studies, but few theoretical guarantees available, see e.g. [16] for a survey. This holds true even more in the specific area of transfer in reinforcement learning (RL), which presents some additional challenges. In an RL problem the learner when set a task has to collect information in the underlying domain by itself with the aim of finding a successful policy fulfilling the task. The question of transfer is to what extent the collected information and the learned policy can be useful for a different task in the same or a similar domain. One of the problems is that in general the different task may demand a completely different sampling strategy from the learner so that observations for a different problem setting may turn out to be not very useful. This makes the question of transfer already difficult for a concrete problem setting and it is even harder to come up with any theoretical results that generalize beyond.

Accordingly, the two survey papers of Taylor & Stone [22] and Lazaric [14] are mainly dedicated to establishing a taxonomy of various settings as well as used performance criteria investigated in the literature. One particular problem is the quantification of the benefit of transfer in RL, for which there are hardly any general results. While Taylor & Stone [22] present different criteria for the evaluation of transfer learning methods, these are rather discussed with respect to application to empirical studies.

Among the considered criteria are the *total reward*, that is, the total amount of accumulated reward, and the *transfer ratio*, that is, the ratio of total reward

accumulated with transfer over the total reward accumulated without transfer. This latter criterion is the closest in the literature to what we suggest in this paper, that is, to consider the *regret* in order to measure the utility of re-using samples. The regret compares the accumulated rewards of a learning agent to that of an optimal policy. In what follows we aim to show that the regret of an agent that learns to solve some task becomes smaller when it uses samples collected during learning some other task.

1.1 Mutual Regret in a Multi-Agent Setting

Our results are derived in a more general multi-agent setting with several agents simultaneously learning different tasks in the same environment, represented by a Markov decision process (MDP) with a transition function common to all agents but specific reward functions for each single agent. The agents share their observations and we are interested in quantifying the usefulness of the shared information. For this we propose to compare the *mutual regret* over all agents to the regret in the standard (single-agent) RL setting when no additional information is available. The respective improvement in regret can be interpreted in terms of mutual transfer between the learning agents.

1.2 Related Work

While Taylor & Stone [22] mention multi-agent settings as an interesting application area for transfer (cf. the related survey [20]), there is also a particular survey of da Silva and Costa [6], which discusses transfer for multi-agent RL. In general, whereas we will concentrate on MDPs in which all agents share a common transition function that only depends on the chosen action by each individual agent, work on multi-agent RL (MARL) settings often rather considers the more complex *stochastic game* scenario where transitions depend on the joint actions chosen by the agents. This also gives rise to different

4 *Transfer in Reinforcement Learning via Regret Bounds for Learning Agents*

goals ranging from Nash equilibria and cooperative learning to adversarial settings in which one is interested in algorithms for a single self-interested agent. Applications for such settings can be e.g. found in game playing, fleet balance, autonomous driving, or robotics, see [4, 11, 26]. Research on transfer learning in MARL settings often provides general frameworks and architectures [25, 27], also for cross-task transfer [19], however there are hardly any general results available.

Our approach rather fits the inter-agent transfer learning setting discussed in the survey [5]: Rather than having a single agent that transfers knowledge of its previous tasks, we have several agents simultaneously learning from each other. In our setting they are solving different tasks, however do this in the same environment.

Although only loosely related to our research, we briefly point out some work on transfer in multi-agent *deep* RL. In general, transfer in (single-agent) deep RL has its own specific challenges, the survey of [28] provides a brief overview of different topics and methods. A respective survey on multi-agent deep RL is also available [7], a more recent review focuses on the cooperative setting [15]. As theoretical results are already scarce for deep RL, we are not aware of any respective guarantees in the more general multi-agent settings.

More importantly, the setting of multiple agents collaborating on a single problem has been dealt with in the literature under various different names. In *fleet learning*, the agents do not interact with each other, but instead act at the same time in similar environments. This is used in practice for various uniform machines with slight differences in manufacturing, such as wind farms [23]. Unlike that, in our setting, the agents share the environment but may have completely different tasks.

In *distributed learning* one aims to scale up learning by using multiple parallel agents on the same task, see e.g. [8]. This idea has also been used for more difficult learning settings such as POMDPs [1]. We generally consider the agents having different tasks, however also show some improvements in theory and experiments when all agents try to solve the same task. Moreover, we are not aware of any other work that considers regret in a distributed RL setting.

Another related concept is *federated learning*, which however puts more emphasis on different topics such as privacy and limiting the amount of data that is shared, see [18, 13], which is something that is not relevant in our setting.

1.3 Contributions and Outline

We consider a multi-agent setting in which all agents have the same state and action space. The transferred knowledge is the experience of each single agent (also called *instance transfer*), which is shared with all other agents. It is assumed that each agent learns a different task, however the agents act in the same environment. That is, the transition probabilities are the same, while each agent as its own reward function.

For this setting, we propose an optimistic learning algorithm that is an adaptation of a single-agent RL algorithm to the multi-agent case and for which we can show regret bounds that are able to quantify the benefit of transfer in the considered setting. Implicitly our multi-agent algorithm also contributes to an issue pointed out by Lazaric [14] who stated that “*the problem of how the exploration on one task should be adapted on the basis of the knowledge of previous related tasks is a problem which received little attention so far.*”

As already mentioned we are interested in regret bounds that measure for any time step T the difference of the accumulated reward of an agent to the

total reward an optimal policy (for the respective agent) would have achieved. We will see that while it is difficult to say how much benefit the experience of a single agent contributes to learning another task by a different agent, when considering the *mutual regret* over all agents the worst case regret bounds are smaller by a factor of \sqrt{N} for a total of N agents.

These are the first theoretical regret bounds for reinforcement learning with multiple simultaneously learning agents in MDPs we are aware of. Our theoretical findings are complemented by experiments on two toy domains, which confirm that the regret in the joint learning setting per agent is smaller than in the standard RL case. From the perspective of transfer learning, these results demonstrate the mutual utility of samples collected by agents learning different tasks in the same environment both practically and theoretically.

The paper is organized as follows: We start with some preliminaries about MDPs in Section 2. Section 3 defines the precise setting of transfer learning we are interested in and introduces the notion of *mutual regret* in a multi-agent setting. Then we present our algorithm, for which we also provide respective performance bounds in the main Section 4, followed by a detailed discussion about the interpretation of our results from the perspective of transfer learning in Section 5. Section 6 presents the results of some complementary experiments. The paper is concluded by Section 7 considering a few open questions and possible directions of future work.

2 MDP Preliminaries

Before we introduce the transfer learning setting we consider, we start with some preliminaries on Markov decision processes (MDPs), which will be used to represent single learning tasks.

Definition 1. A Markov decision process $M = (\mathcal{S}, \mathcal{A}, p, r, s_1)$ consists of a set of states \mathcal{S} of finite size S and a set of actions \mathcal{A} of finite size A available in each of the states.¹ When choosing an action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, the agent obtains a stochastic reward with mean $r(s, a)$ and moves to a new state s' that is determined by the transition probabilities $p(s' | s, a)$. The learning agent starts in an initial state s_1 .

A *policy* in an MDP defines which actions are taken when acting in the MDP. In the following, we will only consider *stationary* policies $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that pick in each state s a fixed action $\pi(s)$. For the average reward criterion we will consider, it is sufficient to consider stationary policies (cf. Section 2.2 below).

2.1 Assumptions and the Diameter

To allow the agent to learn, we make further assumptions about the underlying MDP. First, with unbounded rewards it is difficult to learn (as it is easy to miss out a single arbitrarily large reward at some step), so we assume bounded rewards. Renormalizing, in the following we make the assumption that all rewards are contained in the unit interval $[0, 1]$.

Further, we assume that it is always possible to recover when taking a wrong action once. This is possible if the MDP is *communicating*, i.e., any state is reachable from any other state with positive probability when selecting a suitable policy. That is, given a communicating MDP and two states s, s' there is always a policy π such that the expected number of steps $T_\pi(s, s')$ that an agent executing policy π needs to reach state s' when starting in state s is finite. Next, we introduce the diameter of the MDP, a related parameter we

¹It is straightforward but notationally a bit awkward to generalize our setting as well as the results to the case where different actions are available in each of the states.

will use in the following that measures the maximal distance between any two states in a communicating MDP.

Definition 2. *The diameter D in a communicating MDP M is defined as*

$$D(M) := \max_{s, s' \in \mathcal{S}} \min_{\pi} T_{\pi}(s, s').$$

2.2 The Learning Goal and the Notion of Regret

The goal of the learning agent is to maximize its accumulated reward after any T steps. For the sake of simplicity, in the following we will however consider the *average reward*

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T r(s_t, a_t),$$

where s_t and a_t are state and action at step t , respectively. Unlike the accumulated T -step reward, the average reward is known to be maximized by a stationary policy π^* [17]. The optimal average reward ρ^* is a good proxy for the (expected) optimal T -step reward, as $T\rho^*$ differs from the latter at most by a term that is of order D , as noted by [12]. Accordingly, we measure the performance of the learning agent by considering its *regret* after any T times steps defined as

$$R_T := T\rho^* - \sum_{t=1}^T r_t,$$

where r_t is the reward obtained by the agent at step t .

Literature provides bounds on the regret for various algorithms. One of the first algorithms for which regret analysis has been provided is UCRL2 [12], an algorithm that implements the idea of using optimism in the face of uncertainty and that inspired several variations based on the same idea [9, 10, 3]. For UCRL2, a regret bound of order $\tilde{O}(DS\sqrt{AT})$ has been given, complemented by lower bound of $\Omega(\sqrt{DSAT})$ showing that the upper bound is close to optimal.

In the meantime, the gap between upper and lower bound on the regret has been closed in [2] for another optimistic algorithm in the spirit of UCRL2. We will employ UCRL2 as a base for our multi-agent learning algorithm in Section 4.

3 Transfer Learning

In this section, we first suggest how regret can be used to quantify the benefit of transfer in RL in a sequential setting, before introducing the related notion of *mutual regret* in a multi-agent setting.

3.1 Sequential Transfer Learning

We are interested in the situation when a learning agent is presented two or more tasks in a row, each represented by an MDP on the same state-action space. That is, we have a finite sequence of MDPs M_1, M_2, \dots, M_N and the learner operates in each MDP M_α for T_α steps, before the task changes to the next MDP $M_{\alpha+1}$. (Unlike the tasks themselves, these change points are known to the learner.)

The transfer learning paradigm we suggest is as follows. Consider the regret of a learning algorithm that treats the tasks independently of each other and compare it to the regret of a modified algorithm that uses in each task M_α the information gathered in previous tasks $M_1, M_2, \dots, M_{\alpha-1}$. The difference in regret obviously is a quantification of the benefit of transfer.

Obtaining bounds for transfer learning in a given task sequence seems to be hard however, cf. the open questions at the end of the paper. What we managed to achieve are results for a related measure which we call *mutual regret* in a multi-agent setting, which we introduce in the following section.

3.2 Mutual Transfer and Regret in a Multi-Agent Setting

While transfer as defined in the previous section depends on a fixed task sequence $M_1, M_2, \dots, M_{\aleph}$, we rather take a look at the joint transfer over all permutations of these tasks. More precisely, we consider the following multi-agent setting, which generalizes the introduced RL setting to more than one agent. There are \aleph learning agents $\alpha = 1, 2, \dots, \aleph$ which simultaneously act in their own MDP M_α . We assume however that all agents share their experiences. That is, each agent has immediate access to all the observations made by each other agent α , in particular its history $(s_t^\alpha, a_t^\alpha, r_t^\alpha)_{t \geq 0}$ specifying the state s_t^α visited at step t , the action a_t^α chosen and the reward r_t^α obtained. (As each learning agent aims to maximize its own rewards, knowledge of the other agents' rewards is only of use in the special case when all agents have the same reward function, cf. Theorem 2 below.)

In terms of the transfer learning setting introduced before, this means that each agent learns its own task M_α , however having at any step t access to the samples collected by the other agents in their learning tasks up to the same step t . Unlike in the sequential setting where we assume that the learner has *all* the information on the previous tasks, this means a more limited access and accordingly a harder task.

In what follows we consider that all agents act in the same environment, thus having not only the same state-action space but also common transition probabilities. However, each agent α is assumed to have a different task, represented by a reward function $r^\alpha(s, a)$ specific to the agent α .

Then for each agent α we can define its *individual regret* as

$$R_T^\alpha := T\rho^{*,\alpha} - \sum_{t=1}^T r_t^\alpha,$$

where $\rho^{*,\alpha}$ is the optimal average reward under the reward function r^α . The individual regret can be considered as a proxy for the regret in the sequential transfer learning setting for which it seems difficult to obtain meaningful results. In the following we will instead consider the *mutual regret* defined as the sum of the individual regret over all agents, i.e.,

$$R_T^m := \sum_{\alpha} R_T^{\alpha}.$$

We stick to the idea we have suggested for the sequential transfer setting: Our aim is to compare the mutual regret averaged over all agents to the regret when learning one of the tasks M_{α} in isolation (i.e., without transfer). The following section shows that one can indeed obtain respective improved regret bounds for the multi-agent case, showing that one can quantify the benefit of transfer via bounds on the mutual regret.

4 Regret Bounds for Multi-Agent Learning

In this section, we propose an algorithm for the introduced multi-agent setting and present upper bounds on the mutual regret of all learning agents.

4.1 UCRL2 with Shared Information

For our approach we propose Multi-agent-UCRL (shown in Algorithm 1), a variant of the reinforcement learning algorithm UCRL2 [12] that is able to make use of shared information. That is, each agent will select its actions according to UCRL2 (briefly introduced below) however using not only information collected by itself but also by all other agents.

Algorithm 1 Multi-agent-UCRL with shared information

-
- 1: **Input:** State space \mathcal{S} , action space \mathcal{A} , number of agents \aleph , confidence parameter δ .
 - 2: **Initialization:** Observe the initial state s_1 .
 - 3: In the following, let t be the current time step.
 - 4: **for** episodes $k = 1, 2, \dots$:
 - 5: Compute estimates $\hat{r}_k^\alpha(s, a)$ for rewards of each agent α and estimates $\hat{p}_k(\cdot | s, a)$ for transition probabilities (common to all agents).
 - 6: **Compute policy $\tilde{\pi}_k^\alpha$ for each agent α :**
 - 7: Let \mathcal{M}_k^α be the set of plausible MDPs \tilde{M} with rewards $\tilde{r}(s, a)$ and transition probabilities $\tilde{p}(\cdot | s, a)$ close to the computed estimates

$$\begin{aligned} |\tilde{r}(s, a) - \hat{r}_k^\alpha(s, a)| &\leq \text{conf}_r^\alpha(s, a, S, A, \aleph, \delta, t), \\ \|\tilde{p}(\cdot | s, a) - \hat{p}_k(\cdot | s, a)\|_1 &\leq \text{conf}_p(s, a, S, A, \aleph, \delta, t). \end{aligned}$$

- 8: For each agent α compute an MDP \tilde{M}_k^α in \mathcal{M}_k^α with

$$\rho^*(\tilde{M}_k^\alpha) = \max\{\rho^*(M) \mid M \in \mathcal{M}_k^\alpha\}.$$

Further, let $\tilde{\pi}_k^\alpha$ be the respective optimal policy in \tilde{M}_k^α .²

- 9: **Let each agent α execute policy $\tilde{\pi}_k^\alpha$:**
 - 10: Let each agent α
 - 11: \triangleright choose action $a_t^\alpha = \tilde{\pi}_k^\alpha(s_t^\alpha)$,
 - 12: \triangleright obtain reward r_t^α , and
 - 13: \triangleright observe state s_{t+1}^α .
 - 14: If the number of visits of some agent in some state-action pair has been doubled since the start of the episode, terminate current episode and start new one.
-

UCRL2 is a model-based algorithm that implements the idea of optimism in the face of uncertainty. Basically, UCRL2 uses the collected rewards as well as the transition counts for each state-action pair (s, a) in order to compute estimates \hat{r} , \hat{p} for rewards and transition probabilities, respectively (cf. line 5 of Algorithm 1). While the original algorithm only considers a single agent, in the shared information setting the estimates are computed from the collected observations of all agents. Based on the estimates and respective confidence intervals a set \mathcal{M}^α of plausible MDPs is computed for each agent α (line 7).

²The computation of $\tilde{\pi}_k^\alpha$ and \tilde{M}_k^α can be done using *extended value iteration* as introduced in [12]. For the proof of the regret bound we will assume that extended value iteration is employed with accuracy $1/\sqrt{t}$.

These confidence intervals are chosen so that on the one hand they shrink with an increasing number of observations but on the other hand always contain the true values of the underlying MDP with high probability. Different choices for these confidence intervals have been suggested in the literature. Here we stick to the confidence intervals originally used for UCRL2 [12], however point out that one can e.g. use tighter Bernstein style concentration inequalities such as suggested in [3]. Instead we use simpler Hoeffding concentration inequalities and define the confidence intervals used by agent α at step t as

$$\text{conf}_r^\alpha(s, a, S, A, \aleph, \delta, t) := \sqrt{\frac{7 \log\left(\frac{2S\aleph t}{\delta}\right)}{2 \max\{1, N_t^\alpha(s, a)\}}}, \quad (1)$$

$$\text{conf}_p(s, a, S, A, \aleph, \delta, t) := \sqrt{\frac{14S \log\left(\frac{2\aleph t}{\delta}\right)}{\max\{1, N_t(s, a)\}}}, \quad (2)$$

where $N_t^\alpha(s, a)$ is the number of samples taken by agent α in (s, a) , and $N_t(s, a) := \sum_{\alpha=1}^{\aleph} N_t^\alpha(s, a)$ is the total number of samples taken by agents in (s, a) . As all agents have the same transition function, but the reward functions are individual, the confidence intervals for rewards depend only on the number of visits in (s, a) by the agent α , while the confidence intervals for the transition probabilities depend on the number of samples shared by all agents. Moreover, in order to account for the S different outcomes of a sample from the transition probability distribution, the confidence intervals in (3) have an additional factor of \sqrt{S} , cf. the bounds provided by [24], which are used in the regret analysis for UCRL2 (and implicitly also for our results).

Each agent α then optimistically picks an MDP $\tilde{M}^\alpha \in \mathcal{M}^\alpha$ and a policy $\tilde{\pi}^\alpha$ that maximize the average reward (line 8). This policy is played until the number of visits in some state-action pair has been doubled for some agent,

when a new policy is computed (lines 10–14). The phases between computation and execution of a new policy are called *episodes*.

We note that while the algorithm is described as a central algorithm controlling all \aleph agents, it could also be rewritten as a local algorithm for each single agent having access to the observations of the other agents.

4.2 Upper Bound on the Mutual Regret

The following upper bound on the mutual regret of all agents using Multi-Agent-UCRL holds.

Theorem 1. *With probability $1 - \delta$, after any T steps the mutual regret of agents controlled by Multi-agent-UCRL with confidence intervals (1) and (2) is upper bounded by*

$$R_T^m \leq 15 \left(D\sqrt{S} + \sqrt{\aleph} \right) \sqrt{S\aleph T \log \left(\frac{8\aleph T}{\delta} \right)}.$$

Corollary 1. *If $\aleph < D\sqrt{S}$, the average mutual regret per agent after T steps is $\tilde{O} \left(\frac{DS\sqrt{AT}}{\sqrt{\aleph}} \right)$.*

Corollary 1 shows that averaged over the agents the upper bound on the (mutual) regret improves over the respective bound on the individual regret of a single agent. As already mentioned, the latter is of order $\tilde{O}(DS\sqrt{AT})$ in the standard reinforcement learning setting for the UCRL2 algorithm, see [12]. Appendix A provides a sketch of the proof of Theorem 1, which is based on the analysis of [12]. The main reason for the improvement being possible is that learning the transition function in general is harder in the sense that the respective error has a larger influence on the regret. As several agents are

able to learn the common transition function faster, this results in improved bounds.

Further improvement is possible if the agents also have a joint reward function. While this setting is not relevant in the context of transfer learning, it still may be interesting in the context of distributed learning, when all agents share the same MDP. In this case it makes sense to use common confidence intervals also for the rewards that now do not depend on the individual visits $N^\alpha(s, a)$ of an agent α but on the number of total visits $N(s, a)$ of all agents. That is, confidence intervals

$$\text{conf}_r^\alpha(s, a, S, A, \aleph, \delta, t) := \sqrt{\frac{7 \log\left(\frac{2SA\aleph t}{\delta}\right)}{2 \max\{1, N_t(s, a)\}}}. \quad (3)$$

can be used for the rewards. Accordingly, also the episode termination criterion should be changed to depend not on the visits of an individual agent but on the total number of visits. That is, an episode will be terminated if the total number of visits of all agents has been doubled since the start of the episode. With these modifications, one can obtain regret bounds of order $\tilde{O}(DS\sqrt{A\aleph T})$.

Theorem 2. *In a setting where all agents share the same reward function, the mutual regret of Multi-agent-UCRL (with confidence intervals (3) and using a modified episode termination criterion) with probability $1 - \delta$ after any T steps is upper bounded by*

$$\sum_{\alpha=1}^{\aleph} R_T^\alpha \leq 34DS\sqrt{A\aleph T \log\left(\frac{8A\aleph T}{\delta}\right)}.$$

We highlight the necessary modifications in the proof of Theorem 1 to obtain Theorem 2 in Appendix B. By Theorem 2, Corollary 1 holds without any condition on the number of agents in the setting of shared rewards. As far

as we are aware of, these are the first results on the regret in a distributed RL setting.

5 Discussion

The results presented in Theorems 1 and 2 are not only the first regret bounds for simultaneously learning agents in a MARL setting. They also provide a quantification for the utility of transfer along the lines sketched in Section 3. Comparing the average mutual regret of an agent (Corollary 1) to the bounds of the single agent setting we can see an improvement due to the use of the samples collected by the other agents learning their tasks. While the bounds are derived for mutual transfer in a multi-agent setting, they also provide some information about the sequential transfer in standard (i.e., single agent) RL. That is, the mutual regret in the multi-agent setting can be considered to be a proxy for the average regret when learning a task M_α with samples from tasks $M_1, \dots, M_{\alpha-1}, M_{\alpha+1}, \dots, M_\aleph$ available, where the average is taken over these \aleph learning tasks. As already indicated, learning in the multi-agent setting is harder as less information is available: In the sequential transfer setting all samples of previous tasks are available, while in the multi-agent setting each agent only has the samples of other agents up to the current time step at its disposal.

Notably, while our results demonstrate a collective benefit over all tasks/agents, it seems difficult to obtain any results for the utility of samples collected when learning a specific task M_α for learning a different particular task $M_{\alpha'}$. Thus, while on the one hand individual transfer seems to be difficult to quantify, on the other hand a general mutual transfer is provable.

In the analysis, the role of a single agent is similar to that of a single state-action pair in UCRL2. That is, while it seems difficult to say how much the

learning process focuses on the states and actions relevant for the task to learn, the regret analysis uses that one can at least bound the sum of visits over all state-action pairs, cf. eq. (10) in the analysis.

For a start the choice of an RL standard algorithm like UCRL2 looks quite natural. More generally, we think that our approach can be easily adapted to other algorithms for which similar theoretical guarantees are available and hence can serve as a blueprint for further investigations on transfer in RL. It would be particularly interesting to see whether in the presented framework it turns out that some RL algorithms make better use of additional samples coming from other tasks than other algorithms.

6 Numerical experiments

In this section, we report some experimental results on the two domains *River-Swim* and *SixArms* taken from [21]. These results complement our theoretical findings and show that the improvement in the regret bounds (which are worst case guarantees) are not just an artifact of our theoretical analysis and that agents sharing their information indeed suffer smaller regret.

6.1 Setting and Implementation

For the implementation of our multi-agent algorithm we used for each agent the same parameter for δ set to³ $\delta = 0.5$. We investigated the same domains having a different total number of agents N in order to demonstrate that the advantage of joint learning increases with the number of agents.

The first setting is the *SixArms* environment [21] with seven states and six actions. The learner starts in a central state from which six reward states are reachable when choosing the right action. While the central state can be

³Choosing a rather large δ is equivalent to using smaller confidence intervals, which usually results in faster convergence of UCRL2 to an optimal policy. This is because the original confidence intervals are chosen quite conservatively in order to be able to show the bounds on the regret.

reached deterministically, the reward states are differently hard to reach with transition probabilities reaching from 0.01 to 1. That way, for some states the learner has to explore a lot in order to reach them. When picking the right action each of the six reward states s_i gives a different mean reward of $\frac{i}{5}$ with $i = 0, 1, \dots, 5$, and the agent needs to identify the state and action that provide the highest reward. In our experiment, the various agents had the different rewards randomly shuffled among the reward states, so that the optimal state was not fixed but dependent on the agent. That way, agents are expected to spend more time exploring their own promising states, while doing less useful exploration for other agents.

The second setting we considered is the 6-state *RiverSwim* problem also taken from [21], in which agents have to choose a direction that leads with some probability to one of two possible neighbor states. The agents have to learn to abandon safety (and a small reward of 0.05) of the initial state and go through a sequence of no-reward states in order to reach a high reward state (on the other side of the river) giving a maximal reward of 1. Here all agents shared the same reward function. While this setting is not interesting from the perspective of transfer, we performed experiments on it for the confirmation of the results of Theorem 2.

6.2 Results

For the results we report the mutual regret over time averaged over the number of agents (and all experiments), as well as the first and third quartiles of that mean. In a separate graph, we display the regret achieved by each single agent in one arbitrary run of the multi-agent UCRL algorithm.

Figure 1 shows the results for the *SixArms* environment. When there are three agents for the six states to explore, the regret is similar to that of UCRL2

in the ordinary (i.e., single agent) RL setting. However, with an increasing number of agents the regret drops considerably. Although the agents have different target states, states with low reward need less exploration as observations for estimating the transition probabilities are in particular provided by other agents for which this state is more promising. Note that while with 15 agents there will be agents sharing the same optimal arm, this information in itself cannot be exploited by any of the agents since the identity of these agents is unknown. When taking a look at a single run, it can be noticed that the regret of the single agents is quite differently distributed (unlike in the *RiverSwim* domain discussed below). This is because the agents have different tasks some of which are harder (when the state with the highest reward is difficult to reach), while others are quite easy.

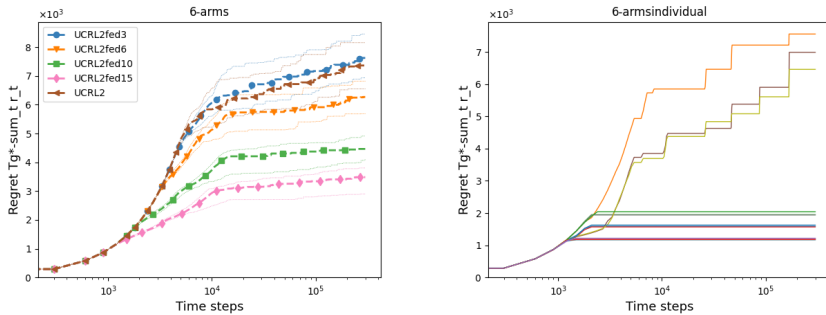


Fig. 1 An experiment in the *SixArm* environment, with (left) the average mutual regret for $N = 1$ (UCRL2) and $N = 3, 6, 10, 15$ over 64 runs, and (right) the regret of each of the 15 agents of Multi-agent-UCRL with shared information in one run.

The results for *RiverSwim* are displayed in Figure 2. We see that the average mutual regret is below that of the single agent setting and also decreases with the number of agents learning together. Moreover, when studying a single arbitrary run, we notice that the regret seems to be quite fairly divided among the various agents, though we only have theoretical guarantees for the mutual regret over all agents.

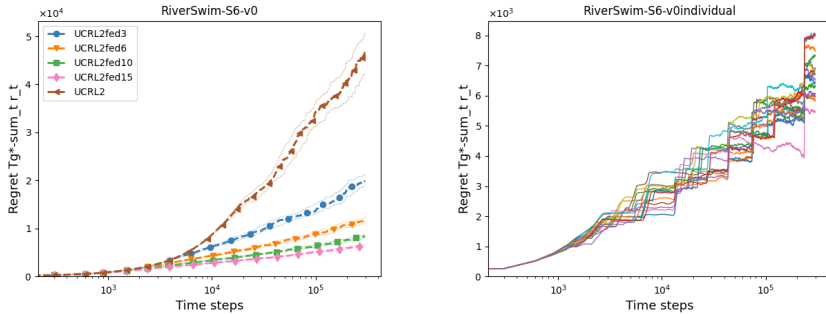


Fig. 2 An experiment in the *RiverSwim* environment, with (left) the average mutual regret for $N = 1$ (UCRL2) and $N = 3, 6, 10, 15$ over 64 runs, and (right) the regret of each of the 15 agents in one run of Multi-agent-UCRL with shared information.

Note that the experimental results give much lower regret than the upper bounds of Theorems 1 and 2. This is due to the conservative nature of the confidence intervals we use. Usually, the empirical performance of UCRL2 and similar algorithms can be improved by further reducing the size of the confidence intervals, however at the cost of losing the theoretical guarantees.

Overall, the results confirm that the regret is reduced when sharing information. This holds not only when the task is the same as in the *RiverSwim* example, but also when the agents have different goal states and only the transition probabilities are shared.

7 Open Questions and Future Work

Our aim was to provide theoretical results for the benefit of transfer in reinforcement learning. The multi-agent setting we have considered allows to illustrate that joint use of samples collected when simultaneously learning different tasks is provably useful overall. This has been formalized by our notion of *mutual regret* which shrinks with the number of agents corresponding to the number of tasks in our transfer learning setting.

While we think that our approach is of general interest, as it leads to otherwise hardly available theoretical guarantees for transfer in reinforcement

learning, the results we have achieved are only a starting point for further research and moreover leave a few open questions.

One obvious point is that mutual regret and the multi-agent setting we have considered are only proxies for the sequential setting for transfer learning introduced in Section 3.1. It is an open problem whether one can achieve any results beyond mutual transfer that are able to quantify the benefit of samples collected when learning a specific task M_α for another particular task $M_{\alpha'}$. A related question is that of obtaining bounds on the individual regret of a single agent in the considered multi-agent setting. Intuitively, if the mutual regret is smaller than the sum over the individual regret values of all agents, some agents have to profit from transfer. However, it seems to be difficult to show that such an advantage exists for a particular agent.

We have already pointed out that in principle any sensible RL algorithm with respective theoretical guarantees could be chosen instead of UCRL2 and be modified to make use of additional samples. An obvious choice would be e.g. the algorithm proposed in [2] as it guarantees optimal regret rates. One could even think of using samples from some algorithm to improve the performance of a different algorithm. In any case, it is an interesting open question whether some algorithms provably use additional samples collected during learning another task in a more efficient way than others. One can imagine that this is already determined by the regret bounds for the single-agent setting. However, this need not be the case in general.

Last but not least, while regret is a natural and in our view the most competitive measure for RL, one could try to obtain similar results using e.g. sample complexity bounds instead. In any of these directions there is a lot of room for future work and we hope that our research sparks some interest in providing more respective theoretical results for transfer in RL.

A Proof of Theorem 1

First, note that the difference between the observed rewards r_t^α when agent α chooses action a in state s at step t and the respective mean rewards $r^\alpha(s, a)$ is a martingale difference sequence and hence can be bounded by Azuma-Hoeffding (cf. e.g. Lemma 10 in [12]), so that

$$\sum_{\alpha} \sum_t (r_t^\alpha - r^\alpha(s_t^\alpha, a_t^\alpha)) \leq \sqrt{\frac{5}{2} T \aleph \log \frac{8 T \aleph}{\delta}}$$

with probability at least $1 - \left(\frac{\delta}{8 \aleph T}\right)^{5/4} \geq 1 - \frac{\delta}{12(\aleph T)^{5/4}}$, similarly to eq. (7) of [12]. Accordingly, writing $v_k^\alpha(s, a)$ for the number of visits of agent α in (s, a) during episode k and defining $\Delta_k^\alpha := v_k^\alpha(s, a)(\rho^{*, \alpha} - r^\alpha(s, a))$ to be the regret of agent α in episode k , we can bound

$$R_T^m = \sum_{\alpha} R_T^\alpha \leq \sum_{\alpha} \sum_k \Delta_k^\alpha + \sqrt{\frac{5}{2} \aleph T \log \frac{8 \aleph T}{\delta}} \quad (4)$$

with probability $1 - \frac{\delta}{12(\aleph T)^{5/4}}$.

Analogously to App. B.1 of [12] using our slightly adapted confidence intervals (3) for the rewards, it can be shown that at each step t the probability that there is an agent α whose true MDP M^α is not contained in its set of plausible MDPs $\mathcal{M}^\alpha(t)$ is bounded by $\frac{\delta}{15t^6}$. As shown in Sec. 4.2 of [12], this implies that the respective regret caused by failing confidence intervals is bounded by

$$\sum_{\alpha} \sum_k \Delta_k^\alpha \mathbb{1}\{M^\alpha \notin \mathcal{M}_k^\alpha\} \leq \aleph \sqrt{T} \quad (5)$$

with probability $1 - \frac{\delta}{12(\aleph T)^{5/4}}$.

For the episodes k in which $M^\alpha \in \mathcal{M}_k^\alpha$, one has analogously to eqs. (14)–(18) of [12] that

$$\begin{aligned}
 \sum_{\alpha} \sum_k \Delta_k^\alpha \mathbb{1}\{M^\alpha \notin \mathcal{M}_k^\alpha\} &\leq \\
 &\left(\sqrt{14 \log \left(\frac{2S\aleph T}{\delta} \right)} + 2 \right) \sum_{\alpha} \sum_k \sum_{s,a} \frac{v_k^\alpha(s,a)}{\sqrt{\max\{1, N_{t_k}^\alpha(s,a)\}}} \\
 &+ D \sqrt{14S \log \left(\frac{2AT}{\delta} \right)} \sum_{\alpha} \sum_k \sum_{s,a} \frac{v_k^\alpha(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}} \\
 &+ \sum_{\alpha} \sum_k \sum_{t=t_k}^{t_{k+1}-1} (p(\cdot \mid s_t^\alpha, a_t^\alpha) - \mathbf{e}_{s_t^\alpha}) \mathbf{w}_k^\alpha,
 \end{aligned} \tag{6}$$

$$\tag{7}$$

where t_k denotes the initial time step of episode k , \mathbf{e}_i is the unit vector with i -th coordinate 1 and all other $S-1$ coordinates 0, and \mathbf{w}_k^α is a modified value vector (computed by extended value iteration for episode k , cf. footnote 2) for agent α with $\|\mathbf{w}_k^\alpha\|_\infty \leq D/2$.

The last term of (7) is a martingale difference sequence and another application of Azuma-Hoeffding gives (similar to eq. 19 of [12] but now for a sequence of length $\aleph T$)

$$\begin{aligned}
 \sum_{\alpha} \sum_k \sum_{t=t_k}^{t_{k+1}-1} (p(\cdot \mid s_t^\alpha, a_t^\alpha) - \mathbf{e}_{s_t^\alpha}) \mathbf{w}_k^\alpha &\leq \\
 D \sqrt{\frac{5}{2} \aleph T \log \left(\frac{8\aleph T}{\delta} \right)} + DS\aleph \log_2 \left(\frac{8T}{SA} \right)
 \end{aligned} \tag{8}$$

with probability at least $1 - \frac{\delta}{12T^{5/4}}$, where the last term comes from a bound over the number of episodes, similar to Appendix B.2 of [12] (but now counted extra for each agent).

Finally, for the two triple sums in (7), we have as shown in [12] that

$$\sum_k \sum_{s,a} \frac{v_k^\alpha(s,a)}{\sqrt{\max\{1, N_{t_k}^\alpha(s,a)\}}} \leq (1 + \sqrt{2}) \sum_{(s,a)} \sqrt{N_T^\alpha(s,a)} \quad (9)$$

and similarly

$$\sum_\alpha \sum_k \sum_{s,a} \frac{v_k^\alpha(s,a)}{\sqrt{\max\{1, N_{t_k}^\alpha(s,a)\}}} \leq (1 + \sqrt{2}) \sum_{(s,a)} \sqrt{N_T(s,a)}.$$

Observing that $\sum_{s,a} N_T^\alpha(s,a) = T$ for each α and $\sum_{s,a} N_T(s,a) = \aleph T$ it follows by Jensen's inequality that

$$\sum_\alpha \sum_k \sum_{s,a} \frac{v_k^\alpha(s,a)}{\sqrt{\max\{1, N_{t_k}^\alpha(s,a)\}}} \leq (1 + \sqrt{2}) \aleph \sqrt{SAT}. \quad (10)$$

and

$$\begin{aligned} \sum_\alpha \sum_k \sum_{s,a} \frac{v_k^\alpha(s,a)}{\sqrt{\max\{1, N_{t_k}^\alpha(s,a)\}}} &= \sum_k \sum_{s,a} \frac{\sum_\alpha v_k^\alpha(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}} \\ &\leq (1 + \sqrt{2}) \sqrt{S \aleph T}. \end{aligned} \quad (11)$$

In summary, we obtain from eqs. (4)–(11) that the mutual regret is bounded by

$$\begin{aligned} R_T^m = \sum_\alpha R_T^\alpha &\leq (D + 1) \sqrt{\frac{5}{2} \aleph T \log \frac{8 \aleph T}{\delta}} + \aleph \sqrt{T} + D S \aleph \log_2 \left(\frac{8T}{SA} \right) \\ &\quad + (1 + \sqrt{2}) \left(\sqrt{14 \log \left(\frac{2 \aleph T}{\delta} \right)} + 2 \right) \aleph \sqrt{SAT} \\ &\quad + (1 + \sqrt{2}) D \sqrt{14 S \log \left(\frac{2 \aleph T}{\delta} \right)} \sqrt{S \aleph T} \end{aligned}$$

with probability $1 - \frac{\delta}{4T^{5/4}}$. Summing over all $T = 2, \dots$ shows that this bound holds simultaneously for all $T \geq 2$ with probability at least $1 - \delta$.

It remains to simplify the bound. Summarizing terms, noting that $DS\aleph \log_2\left(\frac{8T}{SA}\right) \leq \frac{2}{34}DS\sqrt{\aleph AT \log \frac{\aleph T}{\delta}}$ if $T \leq 34^2 A\aleph \log \frac{\aleph T}{\delta}$ (otherwise the theorem holds trivially, cf. App. B of [12]) we obtain

$$R_T^m = \sum_{\alpha} R_T^{\alpha} \leq 14\aleph \sqrt{SAT \log\left(\frac{2A\aleph T}{\delta}\right)} + 15DS\sqrt{S\aleph NT \log\left(\frac{8A\aleph T}{\delta}\right)},$$

which completes the proof of Theorem 1.

B Proof of Theorem 2

The proof of Theorem 2 follows the same line as the proof of Theorem 1 with a few minor adaptations due to the modified confidence intervals (3) for the rewards, which however only causes slight changes in the constants. The main difference is that due to the modified confidence intervals, all terms linear in \aleph can be avoided:

First, all agents now share the same true MDP M as well as the set of plausible MDPs \mathcal{M}_k , so that the probability that M is not in \mathcal{M}_k^{α} neither depends on the single agents nor on their number \aleph . The respective regret over all agents then is bounded by $\sqrt{\aleph T}$ instead of $\aleph \sqrt{T}$ as in (5), however with the same error probability.

Further, using the modified confidence intervals instead of eq. (10) one obtains

$$\begin{aligned} \sum_{\alpha} \sum_k \sum_{s,a} \frac{v_k^{\alpha}(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}} &= \sum_k \sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}} \\ &\leq (1 + \sqrt{2})\sqrt{S\aleph NT}, \end{aligned}$$

which also replaces the linear dependence on \aleph by $\sqrt{\aleph}$ and allows to summarize terms similar to the proof of the original regret bounds of [12]. Details of the derivation are straightforward and are therefore skipped.

Acknowledgments

We would like to thank the anonymous reviewers whose valuable comments greatly helped to improve the presentation of our results. The first author conducted work on of this paper during an internship for Université Paris-Saclay, ENS Paris-Saclay, France.

Declarations Concerning Funding and Competing interests

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

References

- [1] B. Banerjee, J. Lyle, L. Kraemer, and R. Yellamraju. Sample bounded distributed reinforcement learning for decentralized pomdps. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1256–1262. AAAI Press, 2012.
- [2] V. Boone and Z. Zhang. Achieving tractable minimax optimal regret in average reward MDPs. In *Advances in Neural Information Processing Systems 38, NeurIPS 2024*, 2024.
- [3] H. Bourel, O.-A. Maillard, and M. S. Talebi. Tightening exploration in upper confidence reinforcement learning. In *Proc. 37th International Conference on Machine Learning, ICML 2020*, 2020.

- [4] L. Buşoniu, R. Babuška, and B. De Schutter. Multi-agent Reinforcement Learning: An Overview. In J. Kacprzyk, D. Srinivasan, and L. C. Jain, editors, *Innovations in Multi-Agent Systems and Applications - 1*, volume 310 of *Studies in Computational Intelligence*, pages 183–221. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
- [5] F. Da Silva, G. Warnell, A. Costa, and P. Stone. Agents teaching agents: a survey on inter-agent transfer learning. *Auton. Agent Multi-Agent Syst.*, 34(9), 2020.
- [6] F. L. da Silva and A. H. R. Costa. A survey on transfer learning for multiagent reinforcement learning systems. *J. Artif. Intell. Res.*, 64:645–703, 2019.
- [7] D. S. Du, W. A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications. *Artif. Intell. Rev.*, 54(1):3215—3238, 2021.
- [8] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1407–1416. PMLR, July 2018.
- [9] S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122, Sept. 2010.
- [10] R. Fruit, M. Pirota, and A. Lazaric. Improved analysis of UCRL2 with empirical Bernstein inequality. *CoRR*, abs/2007.05456, 2020.
- [11] S. Gronauer and K. Diepold. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 55(2):895–943, Feb. 2022.

28 REFERENCES

- [12] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.
- [13] S. Khodadadian, P. Sharma, G. Joshi, and S. T. Maguluri. Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling. In *PMLR*, volume 162, 2022.
- [14] A. Lazaric. Transfer in reinforcement learning: A framework and a survey. In M. A. Wiering and M. van Otterlo, editors, *Reinforcement Learning*, volume 12 of *Adaptation, Learning, and Optimization*, pages 143–173. Springer, 2012.
- [15] R. Ortner. A review of cooperative multi-agent deep reinforcement learning. *Appl. Intell.*, 53:13677–13722, 2012.
- [16] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.
- [17] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.
- [18] J. Qi, Q. Zhou, L. Lei, and K. Zheng. Federated Reinforcement Learning: Techniques, Applications, and Open Challenges. *Intelligence & Robotics*, 2021.
- [19] H. Shi, J. Li, J. Mao, and K.-S. Hwang. Lateral transfer learning for multiagent reinforcement learning. *IEEE Transactions on Cybernetics*, 53(3):1699–1711, 2023.
- [20] P. Stone and M. M. Veloso. Multiagent systems: A survey from a machine learning perspective. *Auton. Robots*, 8(3):345–383, 2000.
- [21] A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for Markov decision processes. *J. Comput. System Sci.*, 74(8):1309–1331, 2008.

- [22] M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *J. Mach. Learn. Res.*, 10:1633–1685, 2009.
- [23] T. Verstraeten, P. J. K. Libin, and A. Nowé. Fleet Control Using Coregionalized Gaussian Process Policy Iteration. In *24th European Conference on Artificial Intelligence (ECAI 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1571–1578, 2020.
- [24] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. L. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical Report HPL-2003-97, HP Laboratories Palo Alto, 2003.
- [25] T. Yang, W. Wang, H. Tang, J. Hao, Z. Meng, H. Mao, D. Li, W. Liu, Y. Chen, Y. Hu, C. Fan, and C. Zhang. An Efficient Transfer Learning Framework for Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 17037–17048. Curran Associates, Inc., 2021.
- [26] C. Yu, A. Velu, E. Vinitzky, J. Gao, Y. Wang, A. Bayen, and Y. Wu. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Advances in Neural Information Processing Systems 35*, pages 24611–24624, Dec. 2022.
- [27] T. Zhou, F. Zhang, K. Shao, K. Li, W. Huang, J. Luo, W. Wang, Y. Yang, H. Mao, B. Wang, D. Li, W. Liu, and J. Hao. Cooperative Multi-Agent Transfer Learning with Level-Adaptive Credit Assignment, June 2021. 10.48550/arXiv.2106.00517.
- [28] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13344–13362, 2023.