LOCO: ADAPTIVE EXPLORATION IN REINFORCEMENT LEARNING VIA LOCAL ESTIMATION OF CONTRACTION COEFFICIENTS

Manfred Diaz & Liam Paull
Mila, Université de Montréal
{diazcabm, paull}@iro.umontreal.ca

Pablo Samuel Castro
Google Research, Brain Team
psc@google.com

ABSTRACT

We offer a novel approach to balance exploration and exploitation in reinforcement learning (RL). To do so, we characterize an environment's exploration difficulty via the Second Largest Eigenvalue Modulus (SLEM) of the Markov chain induced by uniform stochastic behaviour. Specifically, we investigate the connection of state-space coverage with the SLEM of this Markov chain and use the theory of contraction coefficients to derive estimates of this eigenvalue of interest. Furthermore, we introduce a method for estimating the contraction coefficients on a *local* level and leverage it to design a novel exploration algorithm. We evaluate our algorithm on a series of GridWorld tasks of varying sizes and complexity.

1 INTRODUCTION

The exploration-exploitation dilemma is a central issue in RL: how should a RL agent balance the trade-off between *exploiting* its current knowledge of the environment to maximize returns with *exploring* the environment so as to potentially discover more promising states. The difficulty of this trade-off is that it is not uniform across all environments nor even across all states in the same environment; in the well-known Arcade learning environment (Bellemare et al., 2013) there are a set of games known to be difficult for exploration. Indeed, the infamous Montezuma's Revenge has been the focal point of recent works (Bellemare et al., 2016; Ecoffet et al., 2021). Despite this, Taiga et al. (2020) demonstrated that ϵ-greedy (Sutton & Barto, 2018), one of the simplest and most popular exploration strategies, performs comparably with more sophisticated techniques. Given $\epsilon \in [0, 1]$, this strategy *exploits* its current knowledge with probability $1 - \epsilon$ and *explores* (by acting randomly) with probability ϵ; thus, one elicits a purely random behaviour by setting $\epsilon = 1$.

In this work, we characterize exploration difficulty via the *mixing time*, or time to stationarity, of the Markov Chain induced by uniformly stochastic behaviour. Although this characterization is not novel (Kearns & Singh, 2002; Liu & Brunskill, 2018), our approach differs from prior work in the use of *contraction coefficients*, a notion introduced by Wolfer (2020). Contraction coefficients prove useful as they can bound the eigenvalues of the transition matrices and their products (Seneta, 1979). We demonstrate empirically that bounding the eigenvalues of the transition matrix through contraction coefficients provides good estimates of the mixing times of these chains.

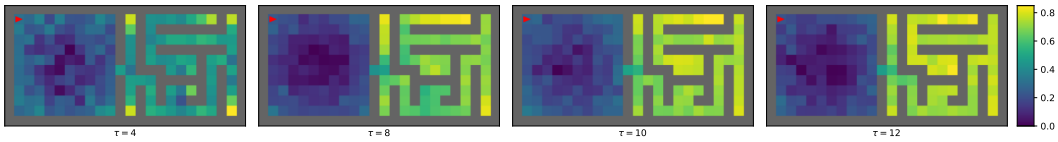


Figure 1: Complexity measures for the states of a GridWorld environment. The maze section of the environment (on the right) is clearly more complex than the open room, and correlates well with our measure of complexity. Higher numbers (colour more yellow) denotes higher complexity.

Consider the hybrid GridWorld shown in Figure 1, where we illustrate one of the local complexity measures we introduce in this paper (see subsection 3.3). It is evident that the maze section of the GridWorld proves more challenging for exploration than the open room section, which correlates well with our proposed measure.

Although the mixing time, the second largest eigenvalue modulus, or the contraction coefficient of a Markov Chain all provide a useful measure of the complexity of an Markov Decision Process (MDP), estimating any of these quantities is impractical: the number of samples required is usually quadratic in the very same quantity we are trying to estimate (Pritchard & Scott, 2004) (see Sec. subsection 3.1). This limitation may be why graph-theoretical analyses (based on the Random Walk Laplacian of the state space graph) have been predominant in the study of this problem (Mahadevan, 2009; Liu et al., 2018). The issue lies largely in the fact that the estimates are done at a *global* level (i.e. for the full MDP).

To overcome these estimation issues, we propose going from *global* to *local* estimates by evaluating the complexity in the vicinity of each state in the MDP. This *global-to-local* view connects with decomposition methods that have been used in the past to compute convergence rates of Markov Chains (Madras & Randall, 2002; Martin & Randall, 2006). We use our local complexity estimators to propose an adaptive exploration method and demonstrate empirically, on discrete MDPs of varying size and structural complexity, that it can improve the *exploitation-exploration* trade-off.

2 BACKGROUND

2.1 MARKOV PROCESSES

An MDP is a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ denotes the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(S)$ defines the transition probabilities, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. A policy $\pi(\cdot|s) : \mathcal{S} \to \Delta(A)$ is a conditional probability distribution over actions. Here, we are interested in the Markov Chains over state space of the form

$$p_\pi(s'|s) = \sum_a \pi(s|a) p(s'|s, a) \tag{1}$$

where each policy induces the transition function $p(s'|s, a)$. A Markov Chain $(S_t)_{t \geq 0}$ is a sequence of random variables whose probability distributions are given by a Markov kernel or operator M such that $S_t \sim \rho_t = \rho_0 \mathrm{M}^t$. When the state space is discrete, the transition operator of the chain will be a row stochastic matrix $\boldsymbol{M}$. We will assume that the Markov Chains we analyze are ergodic (irreducible and aperiodic) throughout. The assumption of ergodicity guarantees the existence and uniqueness of the stationary and limiting distributions (Kulkarni, 2011).

For discrete ergodic Markov Chains, we will be interested in bounding the spectrum of the transition matrix. If λ_i denotes the i-th (of n) eigenvalue (ordered by modulus) of the transition matrix, we have that:

$$1 = \lambda_1 > \lambda_2 \geq, \ldots, \lambda_n$$

From all eigenvalues, we are going to focus on the eigenvalue with the second largest absolute value. The SLEM is the value of the spectrum of the transition matrix $\boldsymbol{M}$ that controls the convergence of the chain to its limiting distribution (its mixing time), and is defined by

$$\lambda_2(\boldsymbol{M}) = \max\{\lambda_2(\boldsymbol{M}), |\lambda_n(\boldsymbol{M})|\} \tag{2}$$

This quantity is of great importance in the literature of Markov Chains as most bounds on the mixing time are monotone on the SLEM (Boyd et al., 2004). For several bounds, we will refer instead to the (absolute) spectral gap $\lambda^*(\boldsymbol{M}) = 1 - \lambda_2(\boldsymbol{M})$.

Moreover, in the discussions that will follow, we will be interested in the amount of time it will take for a chain to visit all of the state space. The state-dependent *cover time variable* $\tilde{t}_{cov}(s)$ of a Markov Chain $(S_t)_{t \geq 0}$ is the minimal value such that for every state, there exist a time $t \leq \tilde{t}_{cov}$ where the chain has visited all the states (Levin & Peres, 2017). Also, we define the worst-case *cover time* of the chain as:

$$t_{cov} = \max_{s \in S} \mathbb{E}_s[\tilde{t}_{cov}(s)] \tag{3}$$

where $s \in S$ is the initial state from which we compute the expectation.

2.2 CONTRACTION COEFFICIENTS

A *contraction* (or ergodicity) *coefficient* $\kappa(\cdot)$ is a continuous scalar function defined for stochastic matrices such that, for any stochastic matrix $\boldsymbol{M}$, $\kappa(\boldsymbol{M})$ is bounded in the $[0, 1]$ interval. From the many definitions of contraction coefficients in the literature, we focus on the one proposed by Dobrushin (Dobrushin, 1956a;b):

$$\kappa_1(\boldsymbol{M}) = \max_{i,j} \|\boldsymbol{M}(i, \cdot) - \boldsymbol{M}(j, \cdot)\|_{TV} \tag{4}$$

where $\| \cdot \|_{TV}$ denotes the Total Variation distance (Cover & Thomas, 2005) between two rows of the stochastic matrix $\boldsymbol{M}$. The Dobrushin's coefficient is simple to compute and has the following properties:

Lemma 1. *(Selee (2009), Theorem 3.5.3) For stochastic matrices $\boldsymbol{M}, \boldsymbol{M}_1, \boldsymbol{M}_2$, we have that*

1. ***(Spectral bound)*** $|\lambda(\boldsymbol{M})| \leq \kappa_1(\boldsymbol{M})$ *for all* $\lambda(\boldsymbol{M}) \neq 1$.
2. ***(Sub-multiplicative)****:* $\kappa_1(\boldsymbol{M}_1\boldsymbol{M}_2) \leq \kappa_1(\boldsymbol{M}_1)\kappa_1(\boldsymbol{M}_2)$

From the sub-mutiplicativity property of contraction coefficients in Lemma 1 it follows that for a transition matrix $\boldsymbol{M}$ of an homogeneous Markov Chain, $\kappa(\boldsymbol{M}^t) \leq \kappa(\boldsymbol{M})^t$. This fact has been useful in offering a modern perspective on mixing time analysis of ergodic Markov Chains (i.e., Wolfer (2020); Veretennikov & Veretennikova (2020)) that we will revisit in Section 3.1.

As contraction coefficients have a long history of applications to the theory of (time) non-homogeneous Markov Chains, as well as traditional and quantum information theory (Dobrushin, 1956a; Seneta, 1979; Makur & Zheng, 2019; Hiai & Ruskai, 2015), we refer the reader to (Selee, 2009) for a historical perspective on this topic.

3 LOCO: LOCAL ESTIMATION OF CONTRACTION COEFFICIENTS

We begin by looking at a notion of complexity of an MDP derived from the properties induced by the Markov Chain of a uniform policy, which we denote M_u. The use of the uniform policy is justified as the actions executed are independent of the state of the MDP and uniformly distributed. Any complexity measure constructed from this Markov Chain would reduce the analysis to the complexity of the transition dynamics of the MDP. Indeed, the transition dynamics of the induced Markov chain, p_u, are spread equally across all actions:

$$M_u(s, s') := p_u(s'|s) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} p(s'|s, a) \tag{5}$$

Additionally, exploration methods like ϵ-greedy spend a fraction of time (e.g., proportional to ϵ) executing this uniformly stochastic behaviour.

However, as illustrated in Figure 1, the amount of exploration required in a given state depends on *local* properties of the transition dynamics. Furthermore, as we show below, estimating *global* properties may be computationally impractical. This section presents a *global-local* perspective that investigates the emergence of the complexity of *global* random exploration from properties of *local* regions of the state space. We leverage Markov Chain decomposition methods to uncover this *global-to-local* connection and propose some useful decompositions that are empirically validated.

In the discussion that follows, we make no assumption on the limiting distribution $p_u(s'|s)$, but we do assume the underlying MDP is *communicating* and the Markov chain ergodic (Kallenberg, 2002).

3.1 GLOBAL COMPLEXITY

Definition 1. *The **global complexity** of a Markov Decision Process $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, R, P \rangle$ under uniform stochastic behaviour is the cover time of the Markov Chain $p_u(s'|s)$ induced by the uniform policy:*

$$\mathrm{C}_g(\mathcal{M}) = \mathbb{E}_{M_u}[t_{cov}] \tag{6}$$

Intuitively, this notion of complexity measures how hard it would be, in expectation, to cover the entire state space of an MDP following a uniformly stochastic policy, regardless of the initial state. An important result due to Broder & Karlin (1989) relates the cover time of a Markov Chain with its spectral gap.

Lemma 2. *(Broder & Karlin (1989), Corollary 8) Let $\boldsymbol{M}$ be the transition probability matrix corresponding to a random walk on a connected graph $\mathbb{G}$ [1]. The expected time to visit all states in $\mathbb{G}$ satisfies*

$$\mathbb{E}[t_{cov}] \leq \frac{1}{\lambda^*(\boldsymbol{M})} \left[|\mathcal{S}|^2 \log |\mathcal{S}|\right] (1 + o(1)) \tag{7}$$

where $|S|$ denotes the size of the state space. This lemma provides an upper bound on the complexity of the MDP under uniformly stochastic behaviour, and clarifies the sources of complexity: 1) the size of the state space as given by the $O(|\mathcal{S}|^2 \log |S|)$ term, and 2) the structural properties of the chain as indicated by the spectral gap $\lambda^*(\boldsymbol{M})$. Note that the inverse spectral gap is not upper-bounded [2], so there may exist Markov Chains where the time to cover the whole state space is controlled by the spectral gap of the chain and not by the size of its state space.

Given the online nature of RL and that the size of the state space may be unknown, it would be more practical to estimate the value of the spectral gap, or more precisely of the SLEM, from samples obtained from the environment. We make use of results from the literature on Markov Chains to compute approximations of the upper-bound on global complexity.

A natural approach is to compute the SLEM of the empirical transition matrix $\hat{\boldsymbol{M}}$ built from samples. Although simple, this estimation method is known to require a number of samples quadratic in the mixing time (Pritchard & Scott (2004), Remark 2) to converge. Furthermore, our empirical evaluations suggest that this estimator both over and underestimates the true value. We obtained more stable results with the *generalized contraction coefficient*, that we denote by κ_{gen}, proposed in Wolfer (2020). Although largely underestimating the true SLEM values, estimation via contraction coefficients is more stable (preserves ordering) independently of the number of samples (see Appendix A).

3.2 Local Complexity

Consider a Markov Chain with transition operator (matrix) $\boldsymbol{M}$ over a finite and discrete state space denoted by $\mathcal{S}$. Let $\{\mathbb{S}_i\}_{i=1}^m$ denote a partition of the state space such that the subsets form a set cover $\mathcal{S} = \cup_i \{\mathbb{S}_i\}$. For each subset, construct a *restricted* Markov Chain with transition operator $\boldsymbol{M}_i$ that models the probability of transition among elements on each subset. Also consider a *projection* Markov Chain, denoted by $\boldsymbol{M}_{\mathbb{O}}$, that models the probability of transition among the subsets in the cover, and a normalization constant $\mathbb{O} = \max_{s \in S} |\{i : s \in \mathbb{S}_i\}|$ that the measures maximum overlap among the sets on the cover. In this setting, Madras & Randall (2002) propose the following theorem that relates the spectral gap of a Markov Chain to the spectral gap of the *projection* chain $\boldsymbol{M}_{\mathbb{O}}$ and the minimum spectral gap among all of the *restricted* chains.

Lemma 3 (State Decomposition Lemma, Madras & Randall (2002), Theorem 1.1)**.** *In the preceding framework,*

$$\lambda^*(\boldsymbol{M}) \geq \frac{1}{\mathbb{O}^2} \lambda^*(\boldsymbol{M}_{\mathbb{O}}) \min_i \lambda^*(\boldsymbol{M}_i) \tag{8}$$

Moreover, using Lemma 3, we can formulate the global-local perspective of complexity under random behaviour:

Theorem 1. *The* global complexity *of a Markov Decision Process under uniformly random behaviour is bounded by*

$$\mathrm{C}_g(\mathcal{M}) \leq \mathbb{O}^2 \frac{1}{\lambda^*(\boldsymbol{M}_{\mathbb{O}})} \frac{1}{\min_i \lambda^*(\boldsymbol{M}_i)} |\mathcal{S}|^2 \log |\mathcal{S}| (1 + o(1)) \tag{9}$$

where $\langle \boldsymbol{M}_{\mathbb{O}}, \{\boldsymbol{M}_i\}_{i=1}^m \rangle$ is a decomposition of the Markov Chain $\boldsymbol{M}$.

[1] The random walk on a connected graph generates a discrete ergodic Markov Chain. (Boyd et al., 2004)

[2] For ergodic chains, the SLEM could approach values infinitesimally close to 1

Proof. See Appendix B.1 □

The dependency of this upper-bound on the inverse of the spectral gap of the restricted chains supports, in part, the idea of local complexity (in the *restricted* and *projection* chains) and how it may be related to global complexity of the original chain. However, these state decomposition theorems are not constructive in that they do not prescribe how to construct the decomposition. In the next sections, we define two state space decompositions and analyze their utility for RL agents.

3.2.1 STATIC AND MONTE CARLO LOCALITY

The first two methods proceed as follows. For each state $s \in S$, simulate the Markov Chain $\boldsymbol{M}$ for the number of steps required to ensure that each subset in the cover contains exactly τ states. We call this method the *static or fixed decomposition*. Alternatively, we can simulate the Markov Chain $\boldsymbol{M}$ for a fixed number of steps τ and construct each subset from the states visited during the simulation. We refer to this second decomposition as the *Monte Carlo decomposition* of the state space. In both, the *restricted* Markov Chains used are derived from the empirical transition matrices as described above.

We are interested to determine the properties of our fine-grained decompositions of the state space through the relative complexity of the *restricted* chains thus defined. Algorithmically, the two decompositions are quite similar (see Algorithms 4 and 5, Appendix C), the main difference stemming from the role played by τ.

3.3 LOCAL ESTIMATION THROUGH CONTRACTION COEFFICIENTS

To study the *restricted* Markov Chains we first define Theorem 1 in terms of the contraction gap $\kappa^*(\cdot) = 1 - \kappa^*(\cdot)$ by first restating Lemma 3 in terms of $\kappa^*(\cdot)$ and then providing a result analogous to Theorem 1. We show these theorems and their proofs on Appendix B.2. Given these theorems, and due to the analysis in Section 3.1, we focus the remaining discussion on contraction coefficients.

Having established our results through contraction coefficients, we use them as theoretical basis to propose LOcal estimation of COntraction coefficients (LOCO), a novel algorithm that, using either the *Static* or the *Monte Carlo* covers defined in Sec. 3.2.1, computes an approximation of the complexity of uniformly random behaviour in the τ-vicinity of a state. Algorithm 1 presents the pseudo code for LOCO that produces, for each state, the Wolfer (2020) generalized contraction coefficient (line 6) extracted from the *restricted* chain having that state as origin.

Algorithm 1 $LOCO$: Local Estimation of Contraction Coefficients

1: **procedure** $LOCO(\mathcal{M}, \tau)$ ▷ an MDP $\mathcal{M}$, τ hyper-parameter
2: $\boldsymbol{M} = \sum_a \pi_u(a|s)p(s'|s,a)$ ▷ Markov Chain for the uniform policy
3: $\mathcal{C}, \mathcal{R} = \texttt{Cover}(\boldsymbol{M}, \mathcal{S}, \tau)$ ▷ $Static$ or $MonteCarlo$ (Appendix C)
4: $\mathcal{O}(\cdot) \leftarrow [\,]$
5: **for** $\boldsymbol{M}_i \in \mathcal{R}$ **do** ▷ $\kappa_{gen}(\cdot)$ for all *restricted* chains
6: $\mathcal{O}(s_i) \leftarrow \kappa_{gen}(\boldsymbol{M}_i)$
7: **end for**
8: **return** $\mathcal{O}$ ▷ Oracle for each $s_i \in \mathcal{S}$
9: **end procedure**

Figure 2 shows the values (min-max normalized) of the Wolfer (2020) *generalized contraction coefficient* estimated for the *restricted* Markov Chains $\boldsymbol{M}_i$, started on each state, using the *static* (top row) and *Monte Carlo* (bottom row) decompositions. These results confirm that both decompositions capture the complexity on the τ-vicinity of each state, but the interpretation differs between the two. In the static decomposition, where the *restricted* Markov Chains evolve over equally-sized state spaces, higher values of the contraction coefficient indicate higher local complexity. Restricted chains with coefficient values close to one indicate that uniformly random behavior would progress less. On the other hand, the decomposition through the *Monte Carlo* construction produces higher values for the contraction coefficients in areas where random behaviour would progress more. This is an indication of how the variable size of the subsets in the decomposition affects the estimation.

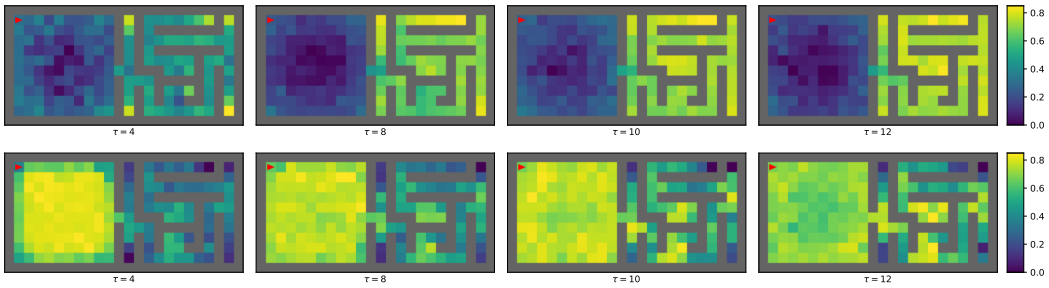


Figure 2: Static (top row) and Monte Carlo (bottom row) cover construction for an MDP, graphically depicted as GridWorld.

In the following sections, we show that both notions of local complexity in the τ-vicinity of a state can be used to balance the exploitation-exploration trade-off. Further empirical evidence is provided in Appendix E.

4 Adaptive ϵ-LOCO

One interpretation of LOCO is as a relative measure of complexity in the τ-vicinity of each state. In this section, we assume that the amount of exploration required at a state is proportional to the local complexity metric. We introduce an *adaptive* coefficient $\hat{\kappa}(s) : \mathcal{S} \to [0, 1]$ such that for each state $s \in \mathcal{S}$, an ϵ-greedy policy executes:

$$\pi_\epsilon = (1 - \hat{\kappa}(s))\, \pi_t + \hat{\kappa}(s)\pi_u \tag{10}$$

where $\epsilon(\cdot)$ is computed as a mixture between a pre-defined value of ϵ and an estimated κ_{gen} computed using LOCO. Algorithm 2 presents ϵ-$LOCO$, an algorithm that through minimal modification to conventional RL balances exploration-exploitation by trading off a (user) pre-specified ϵ coefficient and an estimate of local complexity, using LOCO.

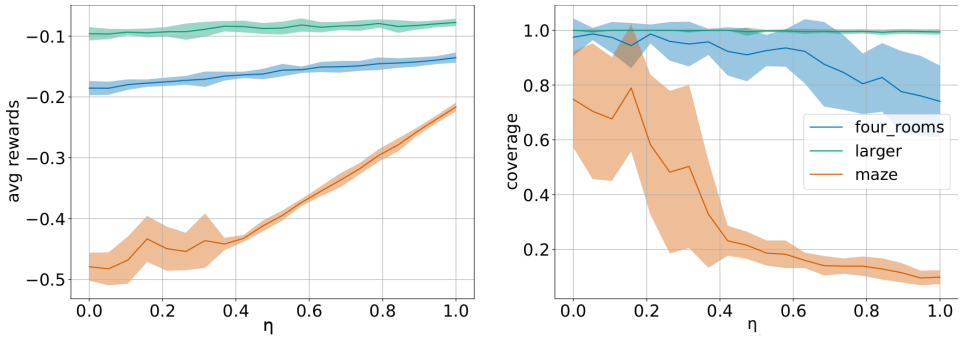
Algorithm 2 ϵ-$LOCO$

1: **procedure** ϵ-$LOCO(\mathcal{M}, \tau, \epsilon, \eta)$
2: $\quad\mathcal{O}(\cdot) \leftarrow$ LOCO$(\mathcal{M}, \tau)$ ▷ Pre-compute oracle
3: $\quad$**for** episode **in** episodes **do**
4: $\quad\quad s_0 \sim \rho$
5: $\quad\quad$**for** t **in** timesteps **do**
6: $\quad\quad\quad\hat{\kappa}(\epsilon, \eta, s_t) \leftarrow (1 - \eta)\epsilon + \eta\mathcal{O}(s_t)$ ▷ Compute an η-mixture
7: $\quad\quad\quad a_t \sim (1 - \hat{\kappa}(s_t))\pi_t + \hat{\kappa}(s_t)\pi_u$ ▷ Adaptive Exploration
8: $\quad\quad\quad s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
9: $\quad\quad\quad \pi_{t+1} = \textit{update}(\pi_t, s_t, a_t, r_t, s_{t+1})$ ▷ Any policy update.
10: $\quad\quad$**end for**
11: $\quad$**end for**
12: **end procedure**

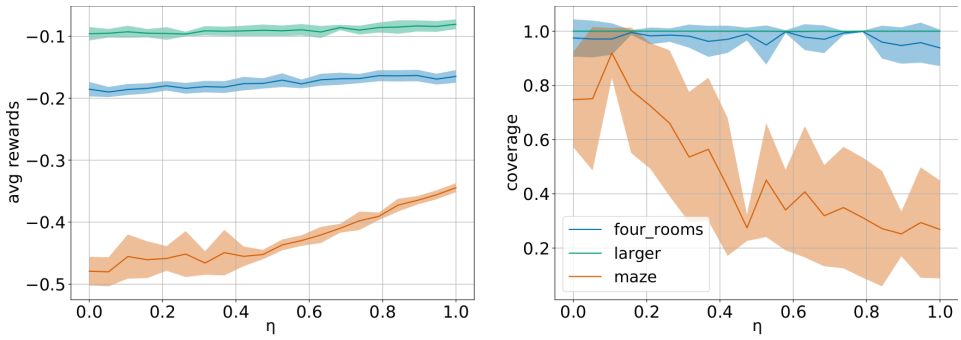
We designed two approaches for computing $\mathcal{O}(\cdot)$ (lines 2 and 6). First, we assume a pre-computed *oracle* that for each state, produces a local complexity of the MDP in the τ-vicinity of the state using LOCO. Next, we assume that the learning agent has a memory (e.g., a replay buffer) that provides access to the last τ visited states and, using these samples, can construct an *online* estimate of LOCO. In both cases, the parameter η (line 6) trades off how much the mixture explores using the preset exploration value and using estimated local complexity.

Offline Adaptive Exploration The oracle (or offline) computation pre-computes a function $\mathcal{O}(\cdot) \leftarrow$ LOCO$(\mathcal{M}, \tau)$ that holds, for each state, the amount of exploration estimated by LOCO for the τ-vicinity of the state. This approach maintains these values through the whole learning algorithm.

Memory-based Adaptive Exploration The online (or memory-based) approach assumes an agent with a memory (denoted by $\mathcal{RB}$) that stores state-to-state transitions. Thus, the amount of *local*



(a) Oracle



(b) Online

Figure 3: Average returns and state space coverage for the *offline* and *online* approaches in three GridWorlds. We used η linearly spaced in the $[0, 1]$ interval. A value $\eta = 0$ presents the results of traditional ϵ-greedy exploration with $\epsilon = 1$.

exploration at a given state (time step) is a function of the last τ transition experienced by the agent, such that $\mathcal{O}(\cdot) \leftarrow LOCO(\mathcal{RB}, \tau)$ is computed as depicted in Algorithm 3 (see Appendix C, Algorithm 6 for a full description). This algorithm allows the agent to *self-supervise* the amount of exploration in the last τ time steps as computed by our LOCO estimate.

Algorithm 3 $LOCO$ from Memory

procedure $LOCO(\mathcal{RB}, \tau)$ ▷ $\mathcal{RB}$: memory, τ: time steps
  $T \leftarrow \mathcal{RB}.last(\tau)$ ▷ Last τ state-to-state transitions.
  $\hat{M} \leftarrow$ **MLE**(T) ▷ Estimate the empirical transition matrix.
  return $\kappa_{gen}(\hat{M})$
end procedure

We empirically evaluated both approaches on a number of discrete MDPs.

4.1 Experiments

We ran ϵ-$LOCO$, and the memory-based m-$LOCO$ algorithms, as defined in Algorithms 2 and 6 in three conventional MDPs *Larger, Four Rooms*, and *Maze* (Figure 5, Appendix A). For all environments, the reward function was set such that executing an action leading to a wall yields $r_t = -1$, stepping into *free space* yielded $r_t = 0$, and reaching the goal provided a positive unit reward $r_t = 1$. The learning algorithm was tabular Q-Learning (Sutton & Barto, 2018), ran for 15 episodes of 200 time steps, with fixed learning rate $\alpha = 0.5$ and discount $\gamma = 0.99$.

Figure 3 shows average reward and state space coverage, over 10 trials, for the *oracle* and *online* approaches respectively. For each trial, we used values of η linearly spaced in the $[0, 1]$ interval, and a pre-defined value of $\epsilon = 1$. The *oracle* approach, with LOCO estimate was fixed throughout

the learning phase, showed an smaller improvement when contrasted with the *online* algorithm. In the latter case, the *self-supervision* property of the memory-based algorithm provided a better exploration-exploitation trade-off, as the *restricted* chain construction is guided by samples of the policies the agent is executing.

For an optimal value of η, we observed an increase in the average returns as well as in the amount of the state space cover, within the fixed budget of 3000 steps, across all environments, but most prominently in *Maze*. As the empirical evidence show, from all environments, making progress by uniformly sampling actions is harder in *Maze* (see Appendix D). Thus, the introduction of our measure of local complexity using LOCO estimates may be more relevant the harder the exploration problem is. In future work, we will further validate this claim on larger and more complex MDPs.

5 Conclusions and Future Work

In this work, we presented LOCO, a novel method for measuring the complexity of random exploration at a local level, using contraction coefficients. Through empirical arguments, we showed that estimating this quantity benefits exploration methods like ϵ-greedy.

The idea of measuring complexity of an MDPs and the properties of Markov Chains induced on it have been of interest in RL (Littman et al., 1995; Kakade, 2001; Kearns & Singh, 2002; Strehl et al., 2009). The study of structural properties of MDPs have been investigated either using *graph theory* (Mahadevan et al., 2006) or the mixing time of Markov Chains induced by policies (Littman et al., 1995; Kearns & Singh, 2002), known to determine the hardness of RL. In recent literature, Liu & Brunskill (2018) and Jinnai et al. (2019) present global analysis of the complexity uniformly random behaviour, though they do no offer the decomposition analysis we showed here. Also, Tarbouriech & Lazaric (2019) and Thodoroff et al. (2018) highlight the importance of looking at structural properties of MDPs in a RL.

In future work, we would like to extend our results to MDPs with continuous state and action spaces, and in particular, evaluating the effectiveness of LOCO in combination with deep networks.

References

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/afda332245e2af431fb7b672a68b659d-Paper.pdf.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, June 2013.

Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest Mixing Markov Chain on a Graph *. *Society for Industrial and Applied Mathematics*, 46(4):667–689, 2004. doi: 10.1137/S0036144503423264. URL https://epubs.siam.org/page/terms.

Andrei Z. Broder and Anna R. Karlin. Bounds on the cover time. *Journal of Theoretical Probability*, 2(1):101–120, jan 1989. ISSN 08949840. doi: 10.1007/BF01048273. URL https://link.springer.com/article/10.1007/BF01048273.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2005. ISBN 9780471241959. doi: 10.1002/047174882X.

R. L. Dobrushin. Central Limit Theorem for Nonstationary Markov Chains. II. *Theory of Probability & Its Applications*, 1(4):329–383, jan 1956a. ISSN 0040-585X. doi: 10.1137/1101029. URL https://epubs.siam.org/page/terms.

R. L. Dobrushin. Central Limit Theorem for Nonstationary Markov Chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, jan 1956b. ISSN 0040-585X. doi: 10.1137/1101006. URL https://epubs.siam.org/page/terms.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. First return, then explore. *Nature*, 2021.

Fumio Hiai and Mary Beth Ruskai. Contraction coefficients for noisy quantum channels. *Journal of Mathematical Physics*, 57(1):53–50, aug 2015. doi: 10.1063/1.4936215. URL http://arxiv.org/abs/1508.03551http://dx.doi.org/10.1063/1.4936215.

Yuu Jinnai, Jee Won Park, Marlos C Machado, Google Brain, and George Konidaris. EXPLORATION IN REINFORCEMENT LEARNING WITH DEEP COVERING OPTIONS. Technical report, sep 2019.

Sham Kakade. Optimizing average reward using discounted rewards. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 2111, pp. 605–615, 2001. ISBN 9783540423430. doi: 10.1007/3-540-44581-1_40.

L. C. M. Kallenberg. Classification Problems in MDPs. In *Markov Processes and Controlled Markov Chains*, pp. 151–165. Springer US, 2002. doi: 10.1007/978-1-4613-0265-0_9. URL https://link.springer.com/chapter/10.1007/978-1-4613-0265-0{_}9.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, nov 2002. ISSN 08856125. doi: 10.1023/A:1017984413808. URL https://link.springer.com/article/10.1023/A:1017984413808.

Vidyadhar G Kulkarni. *Introduction to Modelling and Analysis of Stochastic Systems*. Springer Science+Business Media, LLC, second edition, 2011. ISBN 9780387781884. doi: 10.1016/j.peva.2007.06.006. URL http://books.google.com/books?id=9tv0taI8l6YC.

David Levin and Yuval Peres. *Markov Chains and Mixing Times*. 2017. doi: 10.1090/mbk/107.

Michael L. Littman, Thomas L. Dean, and Leslie Pack Kaelbling. On the Complexity of Solving Markov Decision Problems. In *UAI'95: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 394–402, 1995. doi: 10.5555/2074158. URL http://arxiv.org/abs/1302.4971.

Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James M Rehg, and Le Song. Towards Black-box Iterative Machine Teaching. Technical report, 2018. URL https://arxiv.org/pdf/1710.07742.pdf.

Yao Liu and Emma Brunskill. When Simple Exploration is Sample Efficient: Identifying Sufficient Conditions for Random Exploration to Yield PAC RL Algorithms. Technical report, 2018.

Neal Madras and Dana Randall. Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, 12(2):581–606, 2002. ISSN 10505164. doi: 10.1214/aoap/1026915617.

Sridhar Mahadevan. Learning Representation and Control in Markov Decision Processes: New Frontiers. *Foundations and Trends R in Machine Learning*, 1(4):403–565, 2009. doi: 10.1561/2200000003.

Sridhar Mahadevan, Mauro Maggioni, Kimberly Ferguson, and Sarah Osentoski. Learning representation and control in continuous Markov decision processes. *Proceedings of the National Conference on Artificial Intelligence*, 2(Puterman 1994):1194–1199, 2006.

Anuran Makur and Lizhong Zheng. Information Contraction and Decomposition. Technical report, 2019.

Russell Martin and Dana Randall. Disjoint Decomposition of Markov Chains and Sampling Circuits in Cayley Graphs. *Combinatorics, Probability and Computing*, 15:411–448, 2006. doi: 10.1017/S0963548305007352. URL https://doi.org/10.1017/S0963548305007352.

Geoffrey Pritchard and David J Scott. The eigenvalues of the empirical transition matrix of a markov chain. *Journal of Applied Probability*, 41A:347–360, 2004. ISSN 00219002. doi: 10.1239/jap/1082552210. URL https://www.jstor.org/stable/3215988.

Teresa Margaret Selee. *Stochastic Matrices: Ergodicity Coefficients, and Applications to Ranking*. PhD thesis, North Carolinan State University, 2009.

E. Seneta. Coefficients of ergodicity: structure and applications. *Advances in Applied Probability*, 11(3):576–590, 1979. ISSN 0001-8678. doi: 10.2307/1426955. URL https://www.jstor.org/stable/1426955.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement Learning in Finite MDPs: PAC Analysis. Technical report, 2009.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2nd edition, 2018. URL http://incompleteideas.net/book/bookdraft2018jan1.pdf.

Adrien Ali Taiga, William Fedus, Marlos C Machado, Aaron Courville, and Marc G Bellemare. On Bonus Based Exploration Methods In The Arcade Learning Environment. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJewlyStDr.

Jean Tarbouriech and Alessandro Lazaric. Active Exploration in Markov Decision Processes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019*, Naha, Okinawa, Japan, 2019.

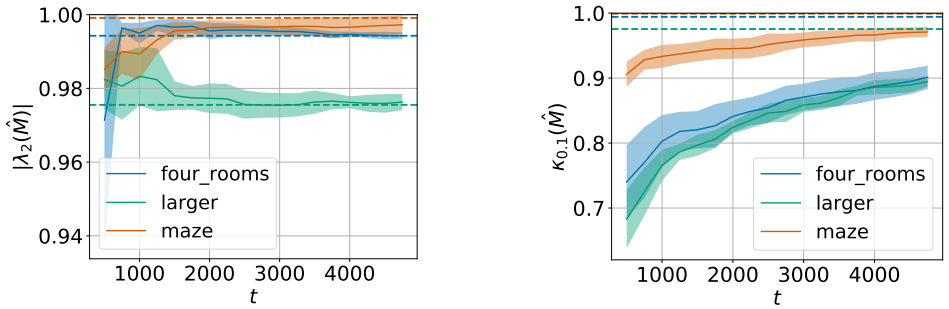
Pierre Thodoroff, Audrey Durand, Joelle Pineau, and Doina Precup. Temporal regularization in markov decision process, 2018. ISSN 23318422.

A. Yu Veretennikov and M. A. Veretennikova. On Convergence Rates for Homogeneous Markov Chains. *Doklady Mathematics*, 101(1):12–15, jan 2020. ISSN 15318362. doi: 10.1134/S1064562420010081. URL https://link.springer.com/article/10.1134/S1064562420010081.

Geoffrey Wolfer. Mixing Time Estimation in Ergodic Markov Chains from a Single Trajectory with Contraction Methods. *Proceedings of Machine Learning Research*, 117:1–15, jan 2020. ISSN 2640-3498. URL http://proceedings.mlr.press/v117/wolfer20a.html.

A SLEM ESTIMATION

We empirically verified the behaviour of the Maximum Likelihood Estimation (MLE) approximation and the *Generalized Contraction Coefficient* for three well-known MDPs, visually represented as grid worlds in Figure 5, and evaluate the effect of the sample size. As we show in Figure 4a, this estimator both over and underestimation of the true value of the SLEM. Furthermore, it would be impractical to compare the results for different chains as the number of samples used affects the estimator differently on each MDP.



(a) MLE

(b) Wolfer (2020) $\kappa_{gen}(\cdot)$

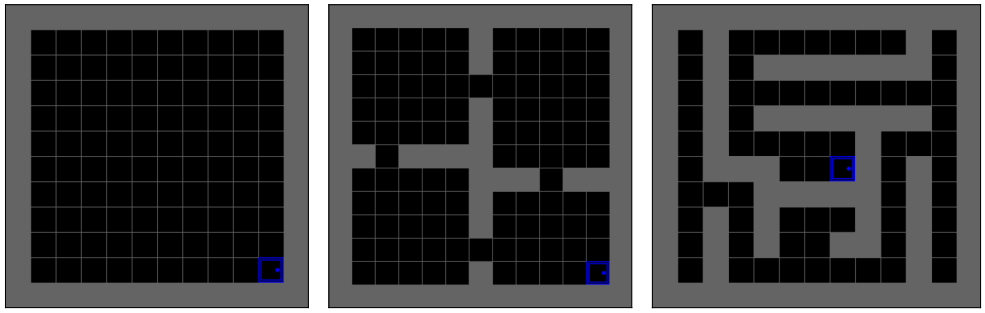
Figure 4: We investigated two approaches to estimate the SLEM, each with different level of sophistication.

Maximum Likelihood Estimation Given samples $\{s^1, s^2, \ldots, s^T\}$ obtained from a $T-$ steps simulation of the Markov Chain $\boldsymbol{M}$. Using this trajectory, one can compute the MLE of the conditional probability distribution $p(S_{t+1} = s' | S_t = s)$, such that, for a pair of states (s_i, s_j) indexed $1 \leq i, j \leq |S|$ respectively, the empirical transition matrix is estimated by:

$$\hat{\boldsymbol{M}}[i,j](T) = \frac{\sum_{t=1}^{T} \mathbb{1}_{\{s_t = s_i, s_{t+1} = s_j\}}}{\sum_{t=1}^{T} \mathbb{1}_{\{s_t = s_i\}}} \tag{11}$$

where $\mathbb{1}_{\{\}}$ denotes the indicator function. For each environment, we generated five 5000-steps rollouts of the uniform policy and averaged the results of computing the SLEM of the transition matrices estimated from samples of incremental size (every 250 steps).

Generalized Contraction Coefficient We use the *generalized contraction coefficient* proposed in Wolfer (2020), denoted by κ_{gen}. While, the estimator of this quantity does not overcome the quadratic dependency on the mixing time, we contrasted the results of computing $\kappa_{gen}(\hat{\boldsymbol{M}})$ (Wolfer (2020), Algorithm 1) with those of the MLE discussed before. As Figure 4b shows, bounding the SLEM of the chain with an estimate of the generalized contraction coefficient largely underestimates the ground truth values. However, in contrast to the MLE, the ordering of the SLEMs of the chains was preserved, independent of the number of samples.



(a) Larger Room

(b) Four Rooms

(c) Maze

Figure 5: The three gridworld environments used in the experiments of SLEM estimation.

B THEOREMS PROOFS

B.1 PROOF OF THEOREM 1

Proof.

$$\begin{aligned}
\lambda^*(\boldsymbol{M}) &\overset{\text{(i)}}{\geq} \frac{1}{\mathbb{O}^2}\lambda^*(\boldsymbol{M}_{\mathbb{O}})\lambda^*(\boldsymbol{M}_k) \\
\frac{1}{\lambda^*(\boldsymbol{M})} &\leq \mathbb{O}^2\frac{1}{\lambda^*(\boldsymbol{M}_{\mathbb{O}})}\frac{1}{\lambda^*(\boldsymbol{M}_k)} \\
\frac{1}{\lambda^*(\boldsymbol{M})}[|\mathcal{S}|^2\log|\mathcal{S}|] &\leq \mathbb{O}^2\frac{1}{\lambda^*(\boldsymbol{M}_{\mathbb{O}})}\frac{1}{\lambda^*(\boldsymbol{M}_k)}[|\mathcal{S}|^2\log|\mathcal{S}|] \\
\mathbb{E}[t_{cov}] &\overset{\text{(ii)}}{\leq} \mathbb{O}^2\frac{1}{\lambda^*(\boldsymbol{M}_{\mathbb{O}})}\frac{1}{\lambda^*(\boldsymbol{M}_k)}[|\mathcal{S}|^2\log|\mathcal{S}|]
\end{aligned}$$

where (i) comes from Lemma 3 and (ii) can be derived from Eq. 7. We omit the extra term $(1 + o(1))$ for clarity of the presentation. □

B.2 THEOREMS FOR CONTRACTION COEFFICIENTS

To study the *restricted* Markov Chains, it would prove useful, provided the estimators analysis in Section 3.1, to define Theorem 1 in terms of the contraction gap $\kappa^*(\cdot) = 1 - \kappa^*(\cdot)$. To do this, we firstly re-state Lemma 3 in terms of $\kappa^*(\cdot)$ and then, we provide an equivalent to Theorem 1.

Theorem 2 (State Space Decomposition via Contractions)**.** *In the preceding framework,*

$$\lambda^*(\boldsymbol{M}) \geq \frac{1}{\mathbb{O}^2}\kappa^*(\boldsymbol{M}_{\mathbb{O}})\min_i \kappa^*(\boldsymbol{M}_i) \tag{12}$$

Proof. First, for any Markov Chain, we have that the following result that relates the contraction gap with the spectral gap:

$$\begin{aligned}
\lambda_2(\boldsymbol{M}) &\overset{\text{(i)}}{\leq} \kappa(\boldsymbol{M}) \\
1 - \kappa(\boldsymbol{M}) &\leq 1 - \lambda_2(\boldsymbol{M}) \\
1 - \kappa(\boldsymbol{M}) &\leq \lambda^*(\boldsymbol{M}) \\
\kappa^*(\boldsymbol{M}) &\leq \lambda^*(\boldsymbol{M})
\end{aligned} \tag{13}$$

where (i) comes from Lemma 1.1. Let $\lambda^*(\boldsymbol{M}_k) = \min_i \lambda^*(\boldsymbol{M}_i)$, using (13) the rest follows:

$$\begin{aligned}
1 - \kappa(\boldsymbol{M}_{\mathbb{O}}) &\leq \lambda^*(\boldsymbol{M}_{\mathbb{O}}) \\
(1 - \kappa(\boldsymbol{M}_{\mathbb{O}}))(1 - \kappa(\boldsymbol{M}_k)) &\leq \lambda^*(\boldsymbol{M}_{\mathbb{O}})\lambda^*(\boldsymbol{M}_k) \\
\frac{1}{\mathbb{O}^2}(1 - \kappa(\boldsymbol{M}_{\mathbb{O}}))(1 - \kappa(\boldsymbol{M}_k)) &\leq \frac{1}{\mathbb{O}^2}\lambda^*(\boldsymbol{M}_{\mathbb{O}})\lambda^*(\boldsymbol{M}_k) \\
\frac{1}{\mathbb{O}^2}(1 - \kappa(\boldsymbol{M}_{\mathbb{O}}))(1 - \kappa(\boldsymbol{M}_k)) &\leq \lambda^*(\boldsymbol{M}) \\
\frac{1}{\mathbb{O}^2}\kappa^*(\boldsymbol{M}_{\mathbb{O}})\min_i \kappa^*(\boldsymbol{M}_i) &\leq \lambda^*(\boldsymbol{M})
\end{aligned}$$

□

Theorem 3. *The* global complexity *of a Markov Decision Process under uniformly random behaviour is bounded by*

$$\mathrm{C}_g(\mathcal{M}) \leq \mathbb{O}^2\frac{1}{\kappa^*(\boldsymbol{M}_{\mathbb{O}})}\frac{1}{\min_i \kappa^*(\boldsymbol{M}_i)}|\mathcal{S}|^2\log|\mathcal{S}|(1 + o(1)) \tag{14}$$

where $\langle \boldsymbol{M}_{\mathbb{O}}, \{\boldsymbol{M}_i\}_{i=1}^m \rangle$ is a decomposition of the Markov Chain $\boldsymbol{M}$.

Proof. From Theorem 2, we have that:

$$\lambda^*(\boldsymbol{M}) \geq \frac{1}{\mathbb{O}^2}\kappa^*(\boldsymbol{M}_{\mathbb{O}})\min_i \kappa^*(\boldsymbol{M}_i)$$

The rest follows as in Theorem 1. □

C ALGORITHMS

Algorithm 4 Static Cover

procedure STATICCOVER($\boldsymbol{M}, \mathcal{S}, \tau$)
  $\mathcal{C} = \{\}, \mathcal{R} = \{\}$
  for all $s \in \mathcal{S}$ **do**
    $\mathbb{S}_i \leftarrow \mathbb{S}_i \cup \{s\}$
    $T \leftarrow [s_0 = s]$
    while $|\mathbb{S}_i| < \tau$ **do**
      $s_t \sim \boldsymbol{M}$
      $\mathbb{S}_i \leftarrow \mathbb{S}_i \cup \{s_t\}$
      $T \leftarrow T + s_t$
    end while
    $\mathcal{C} \leftarrow \mathcal{C} \cup \mathbb{S}_i$
    $\mathcal{R} \leftarrow \mathcal{R} \cup \textbf{MLE}(T)$
  end for
  return $\mathcal{C}, \mathcal{R}$
end procedure

Algorithm 5 Monte Carlo Cover

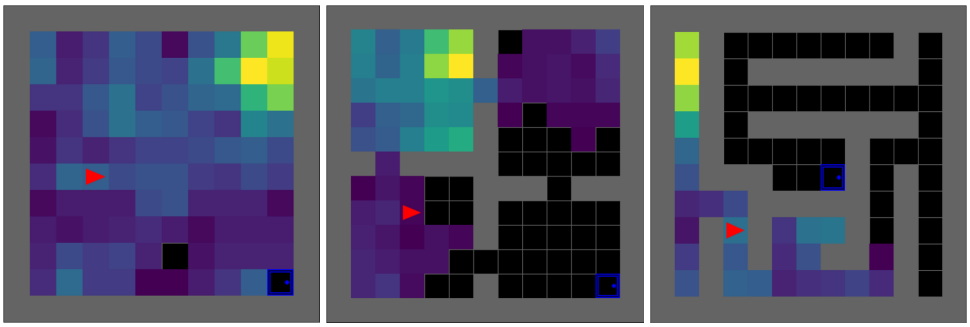
procedure MONTECARLOCOVER($\boldsymbol{M}, \mathcal{S}, \tau$)
  $\mathcal{C} = \{\}, \mathcal{R} = \{\}$
  for all $s \in \mathcal{S}$ **do**
    $\mathbb{S}_i \leftarrow \mathbb{S}_i \cup \{s\}$
    $T \leftarrow [s_0 = s]$
    for $1 \ldots \tau$ **do**
      $s_t \sim \boldsymbol{M}$
      $\mathbb{S}_i \leftarrow \mathbb{S}_i \cup \{s_t\}$
      $T \leftarrow T + s_t$
    end for
    $\mathcal{C} \leftarrow \mathcal{C} \cup \mathbb{S}_i$
    $\mathcal{R} \leftarrow \mathcal{R} \cup \textbf{MLE}(T)$
  end for
  return $\mathcal{C}, \mathcal{R}$
end procedure

Algorithm 6 ϵ-$LOCO$ from Memory

procedure m-$LOCO(\mathcal{M}, \tau, \epsilon, \eta)$
  $\mathcal{RB} \leftarrow \emptyset$ ▷ Assume a replay buffer
  for episode **in** episodes **do**
    $s_0 \sim \rho$
    for t **in** timesteps **do**
      $\mathcal{RB}.update(s_t)$
      $\epsilon(s_t) \leftarrow (1-\eta)\epsilon + \eta LOCO(\mathcal{RB}, \tau)$ ▷ Algorithm 3
      $a_t \sim (1 - \epsilon(s_t))\pi_t + \epsilon(s_t)\pi_u$ ▷ Adaptive Exploration
      $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
      $\pi_{t+1} = update(\pi_t, s_t, a_t, r_t, s_{t+1})$ ▷ Any policy update.
    end for
  end for
end procedure

D RANDOM EXPLORATION

We show the progress of uniform stochastic exploration in three MDPs, graphically depicted as Gridworlds. For a fixed budget of 3000 steps, making progress by executing each action uniformly at random, is harder on *Maze*, than in *Four Rooms*, than in *Larger Room*.



(a) Larger (b) Four Rooms (c) Maze

Figure 6: Progress of random exploration in each *GridWorld* for a fixed number of steps.

E EMPIRICAL ANALYSIS OF DECOMPOSITIONS

We provide further empirical evidence on the properties of the decomposition presented in Section 3.2.1. In this analysis, we used GridWorld with underlying MDPs of varying state space size, with deterministic transition dynamics, and common action space. For each environment, we present the raw value and a min-max per-environment normalized value that renders the visualization more amenable.

E.1 STATIC DECOMPOSITION

E.2 MONTE CARLO DECOMPOSITION

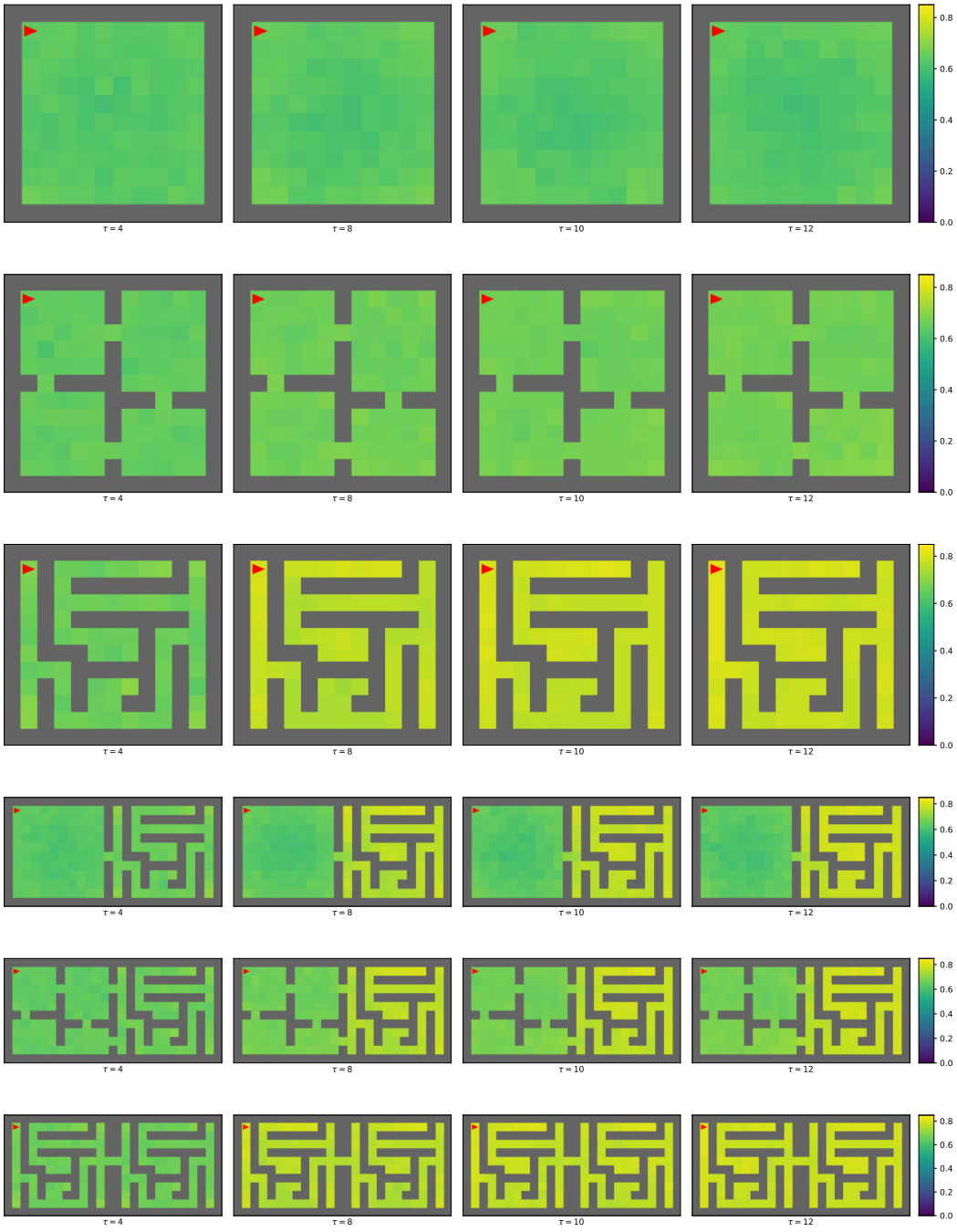


Figure 7: Raw values of the per-state local complexity $\mathrm{C}(\cdot, \tau)$ through a static cover construction.

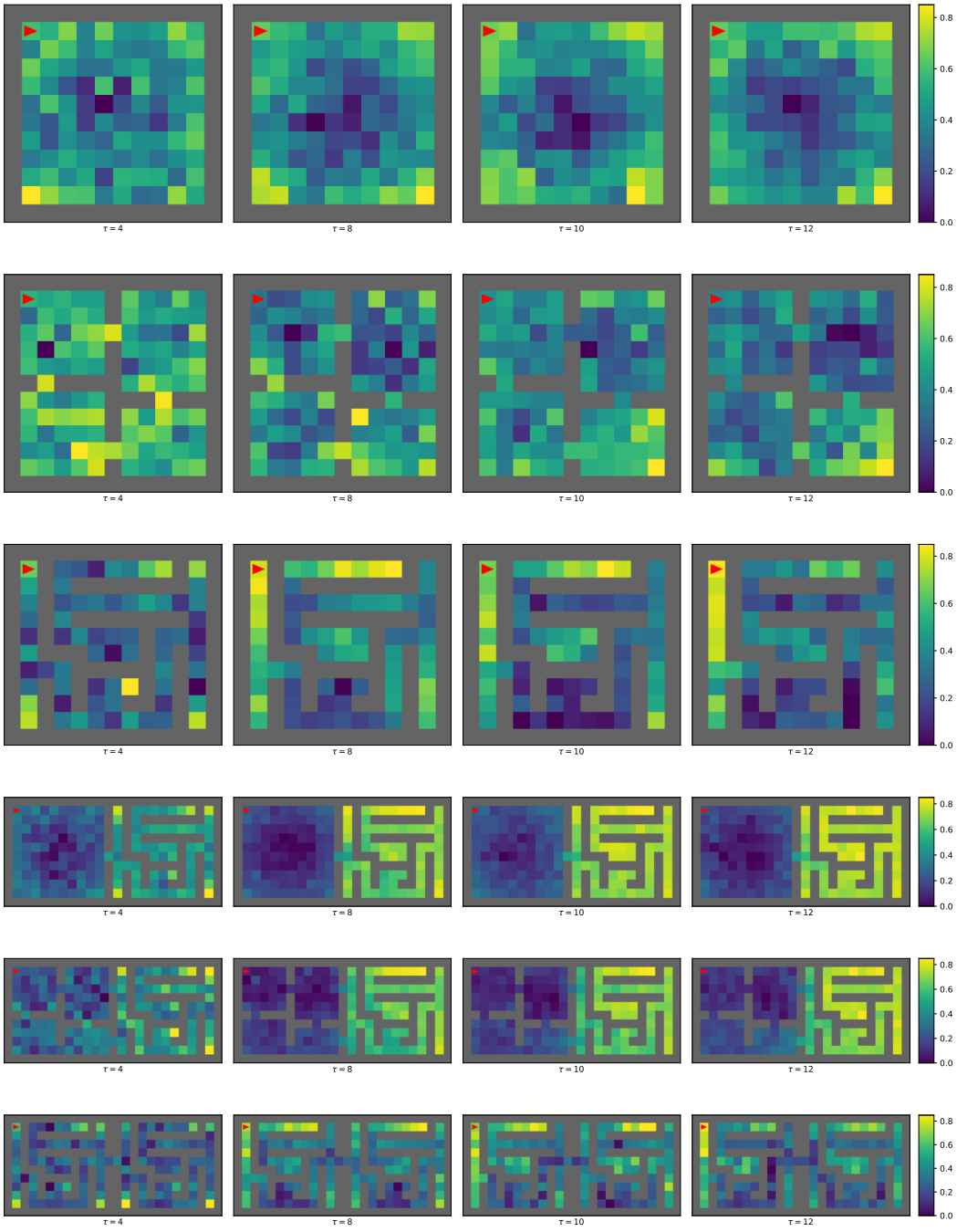


Figure 8: Min-max normalized per-state local complexity $\mathrm{C}(\cdot, \tau)$ through a static cover construction.

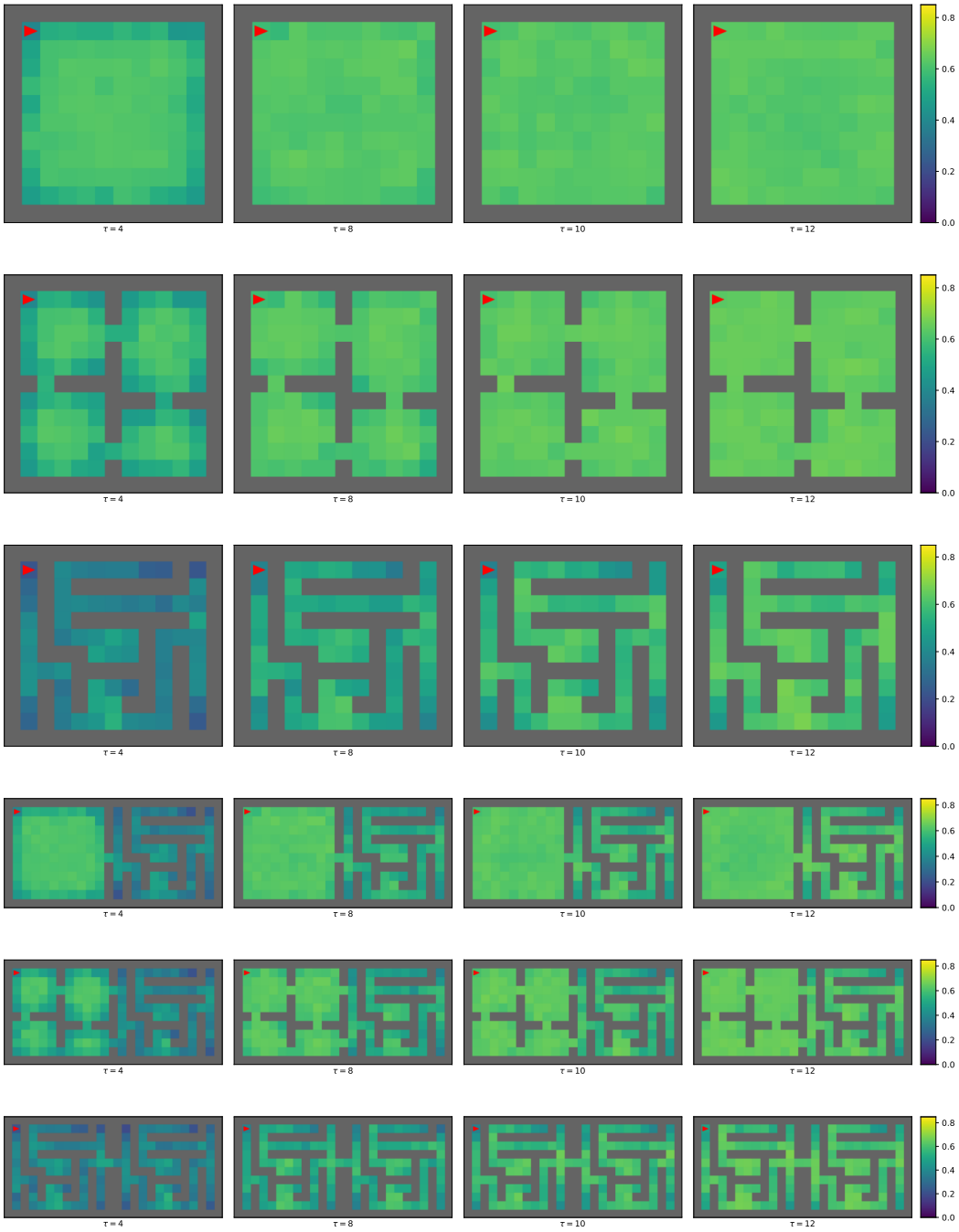


Figure 9: Absolute value of the per-state local complexity $\mathrm{C}(\cdot, \tau)$ through a static cover construction.

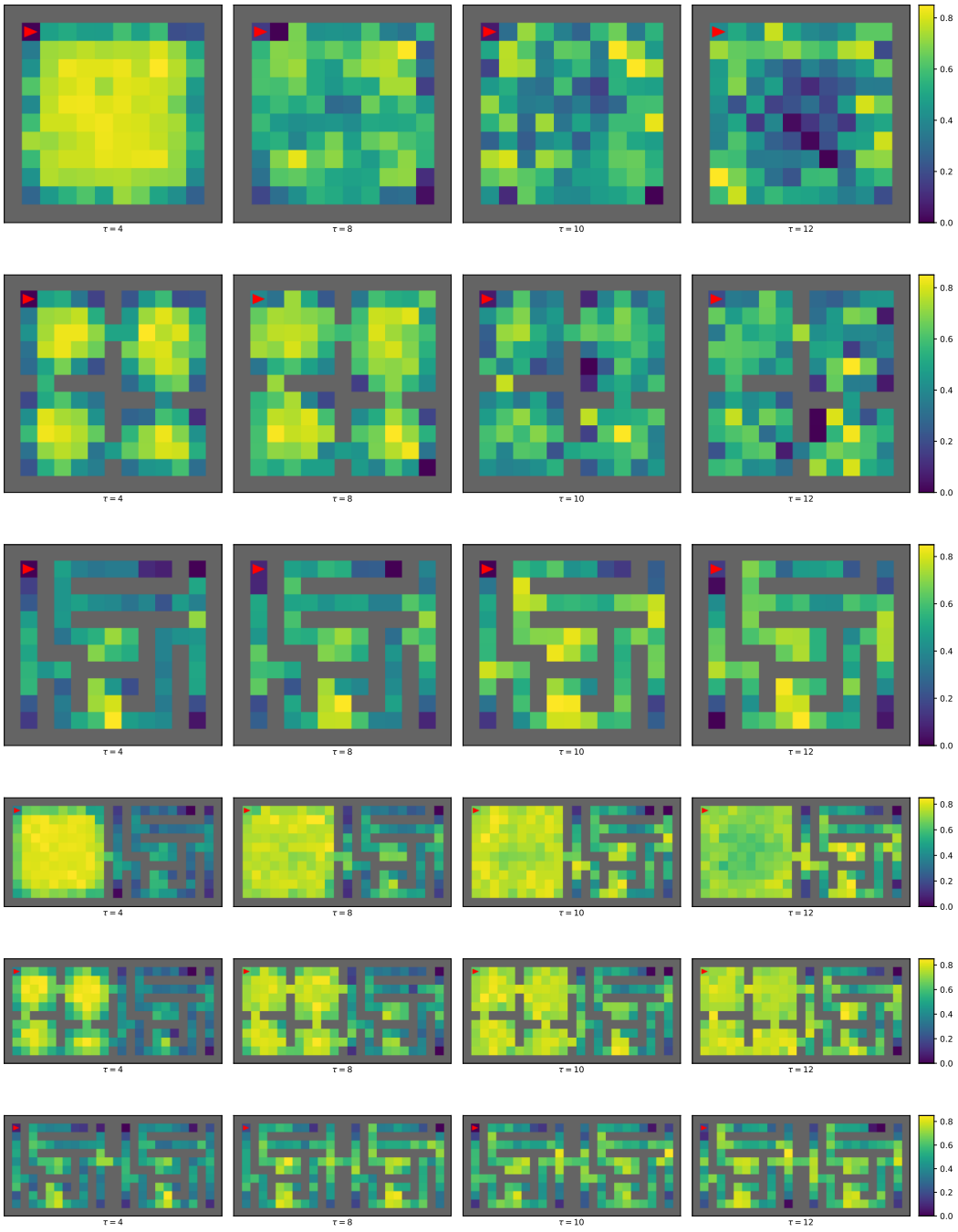


Figure 10: Min-max normalized per-state local complexity $\mathrm{C}(\cdot, \tau)$ through a Monte Carlo cover construction.