Markov Chain Estimation, Approximation, and Aggregation for Average Reward Markov Decision Processes and Reinforcement Learning

Ronald Ortner

[a]*Montanuniversität Leoben, Franz-Joseph-Strasse 18, Leoben, 8700, Austria*

Abstract

Markov chains naturally appear in Markov decision processes (MDPs), as in an MDP each policy induces a Markov chain. In the context of reinforcement learning the underlying MDP is unknown and its transition probabilities have to be estimated. The deviation of the corresponding estimates from the true values will cause an error that can be bounded by perturbation bounds for the induced Markov chains. This paper investigates respective questions of Markov chain estimation, approximation, and aggregation which are of interest in MDP and reinforcement learning applications. The focus will be put on error bounds involving transition parameters such as mean first passage times, Kemeny's constant, or the diameter.

Keywords: Markov decision process, Markov reward process, Markov chain perturbation, Markov chain approximation, MDP aggregation, reinforcement learning

1. Introduction

Markov decision processes (MDPs) are generalizations of Markov chains that introduce a decision making component which predestine them as model for various applications. MDPs have been introduced by Richard Bellmann in the 1950s, however special cases have been considered in the context of game theory before, cf. Kallenberg (2016). In the meantime, MDPs have been used in numerous applications as suitable representation for various problem settings. Since beginning of the 1980s, MDPs have also become the standard framework for modeling an unknown environment in reinforcement learning, where a learning agent aims to learn some particular behavior. The learning process by trial and error is considered to be special as the learner has to generate the data from which is learned by herself. Thus, particular attention has been turned to the balance of *exploitation* (that is, playing the option one considers best so far) and *exploration* (looking for a better option), cf. e.g. Fruit (2019).

In the following, we consider various natural questions on MDPs and reinforcement learning that can be dealt with by investigation of suitable underlying simpler Markov chain processes. Thus, we will consider basic properties of Markov reward processes (i.e., Markov chains with an additional reward function) as well as questions of estimation, approximation, and aggregation in Markov chains and MDPs. On the other hand, we will also show how results derived for the more general MDP setting can provide new insights for Markov chains.

Following basic definitions for Markov chains in this section, we will first study Markov reward processes in Section 2. Section 3 introduces the MDP model. Sections 4 and 5 then will consider questions of approximation, while Section 6 deals with aggregation. The final Section 7 illustrates how the concepts and results of the preceding sections can be applied to the restless bandit problem and also introduces a more general notion of approximation.

1.1. Markov Chain Preliminaries

Basic Definitions. We start with introducing some notation, fixing the used terminology, and recalling some basic facts about discrete-time Markov chains.[1] We will usually consider time homogeneous Markov chains, which we write as triples $(\mathcal{S}, \mathbf{P}, \nu)$ of state space $\mathcal{S}$, transition matrix $\mathbf{P}$, and initial distribution ν over $\mathcal{S}$. Sometimes we skip the initial distribution if it is clear from context or does not matter, often we replace it with a fixed initial state. For the entries in row s and column s' of the matrix $\mathbf{P}$, that is, the *transition probability* from s to s', we usually write[2] $p(s'|s)$. That is, writing $S_1, S_2, \ldots$ for the sequence of random states generated by the Markov chain, we have for all $t > 1$

$$\mathbb{P}[S_t = s' \,|\, S_{t-1} = s, S_{t-2} = s'', \ldots, S_1 = s'''] = \mathbb{P}[S_t = s' \,|\, S_{t-1} = s] =: p(s'|s).$$

[1] As reference any textbook on Markov chains such as (Norris, 1998) will suffice.

[2] The notation $p(s, s')$ is more common, however $p(s'|s)$ is more consistent with the MDP notation used later on.

Similarly, we write $p^t(s'|s)$ for the t-step transition probability, which is the respective entry in the power matrix $\mathbf{P}^t$.

Transition Structure. Two states s, s' in a Markov chain are *communicating* if s can be reached from s' with positive probability after a finite number of $t \geq 0$ steps and vice versa. Note that any state communicates with itself so that we can consider the equivalence classes with respect to the relation of being communicating, which we call the *(irreducible) classes* of the Markov chain. A Markov chain is *irreducible* if its state space constitutes an irreducible class.

A state s is *recurrent* if it will be visited again with probability 1 when starting in s. Other states are called *transient*. A class of recurrent states is also called *recurrent*. In the following we will assume the state space to be finite, so that any Markov chain with a single irreducible class is also recurrent.

A Markov chain is called *unichain* if it consists of a single irreducible class and a (possibly empty) set of transient states. When there is more than one irreducible class the Markov chain is called *multi-chain*.

For general multi-chain Markov chains with state space $\mathcal{S}$ we sometimes refer to their *transition structure* meaning the graph with the vertex set $\mathcal{S}$ and directed edges between states s, s' for which $p(s'|s) > 0$. This graph sometimes is called the *graph of the Markov chain*, cf. Woess (2009).

Periodicity. The *period* of a state s is the greatest common divisor of all t for which $p^t(s|s) > 0$. States in the same irreducible class have the same period. We call such a class *periodic* if it has period > 1 and otherwise *aperiodic* and use this terminology accordingly also for irreducible Markov chains.

Stationary Distribution. The Markov chains we will deal with may be periodic so that the powers $\mathbf{P}^t$ of the transition matrix will not converge. Thus, we consider the Cesaro limit $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbf{P}^t$ instead. For an irreducible Markov chain this sum converges to a limiting matrix with identical rows $\boldsymbol{\mu}$, which correspond to the unique *stationary distribution*[3] μ satisfying $\boldsymbol{\mu}^\top \mathbf{P} = \boldsymbol{\mu}$ and $\sum_{s\in\mathcal{S}} \mu(s) = 1$.

[3] We write the stationary distribution μ and other functions $\mathcal{S} \to \mathbb{R}$ sometimes as (column) vector $\boldsymbol{\mu} = (\mu(s))_{s\in\mathcal{S}}$ using bold variables.

Measuring Distance. In the following we will also be interested in the distance of the stationary distributions of two Markov chains over the same state space. For that purpose we introduce various distance measures for vectors, distributions, and matrices. First we recall the definition of the ∞-norm, that is, $\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_\infty := \max_s |\mu(s)-\mu'(s)|$ for vectors $\boldsymbol{\mu}, \boldsymbol{\mu}'$ and $\|\mathbf{P}-\mathbf{P}'\|_\infty := \max_s \sum_{s'} |p(s'|s) - p'(s'|s)|$ for (transition) matrices $\mathbf{P}, \mathbf{P}'$. Further, for vectors $\boldsymbol{\mu}, \boldsymbol{\mu}'$ the 1-norm is defined as $\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_1 := \sum_s |\mu(s)-\mu'(s)|$. Finally, we will also consider the total variational distance, which for the discrete domains we will consider is simply given by $d_{TV}(\boldsymbol{\mu}, \boldsymbol{\mu}') = \frac{1}{2}\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_1$.

Mean First Passage Times. First passage times (sometimes also called *hitting times*) are stopping times for first visits in a given state (or more generally a set of states). The *mean first passage time* $\tau(s, s')$ for two different states s, s' is defined as the expected number of steps it takes to reach s' for the first time when starting in s. Further, we let $\tau(s, s)$ be the *mean return time* for state s, that is, the expected number of steps it takes to return to state s for the first time when starting in s.

Mixing Times. Mixing times measure the time to stationarity of a Markov chain. Accordingly, we define the *ε-mixing time* of a Markov chain to be

$$T_{\mathrm{mix}}(\varepsilon) := \min\big\{t \in \mathbb{N} \,\big|\, \max_s \|\boldsymbol{\mu}_s^t - \boldsymbol{\mu}\|_1 \leq \varepsilon\big\},$$

where $\boldsymbol{\mu}_s^t$ is the distribution after t steps when starting in state s. For defining *the* mixing time of a Markov chain one chooses an arbitrary ε, e.g.[4] $T_{\mathrm{mix}} := T_{\mathrm{mix}}(\frac{2}{\mathrm{e}})$. It can be shown (eq. 4.36 of Levin et al., 2006) that

$$T_{\mathrm{mix}}(\varepsilon) \leq \left\lceil \log_2\left(\tfrac{1}{\varepsilon}\right)\right\rceil \cdot T_{\mathrm{mix}}. \tag{1}$$

2. Prologue: Markov Reward Processes

Before we introduce the general MDP framework, we start with the simpler notion of *Markov reward process*, which on one hand is a simple extension of a Markov chain, while being a special case of an MDP on the other hand.

[4] With this choice our mixing time coincides with that of Paulin (2015).

2.1. Definition and Average Reward

A Markov reward process is obtained by adding random rewards to each state of a discrete-time, time-homogeneous Markov chain.

Definition 1. *A Markov reward process (MRP)* $(\mathcal{S}, r, \mathbf{P}, \nu)$ *consists of a Markov chain* $(\mathcal{S}, \mathbf{P}, \nu)$ *and reward distributions in each state* s *in* $\mathcal{S}$ *with mean* $r(s)$*.*

In the following we assume that rewards are bounded and more precisely consider for the sake of simplicity that they are taken from the interval $[0,1]$. For bounded rewards this can always be achieved by renormalization.

Given an MRP, one is usually interested in the random rewards R_t accumulated in steps $t = 1, 2, \ldots$. If R_t is collected at step t this means that it is drawn from the reward distribution of state S_t and the respective expected value is $r(S_t)$. The S_t are created by the underlying Markov chain and we will consider the expectation $\mathbb{E}[r(S_t)]$ with respect to this random process. The respective *expected finite horizon reward* after T steps is given by

$$v_T(s) := \sum_{t=1}^{T} \mathbb{E}[R_t \,|\, S_1 = s] = \sum_{t=1}^{T} \mathbb{E}\big[r(S_t) \,\big|\, S_1 = s\big],$$

while for $T \to \infty$ one can compute the *average reward* defined as

$$\rho(s) := \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\big[r(S_t) \,\big|\, S_1 = s\big], \tag{2}$$

taking into account that the average reward may be different for different initial states s. However, if the Markov chain underlying the MRP is irreducible then the average reward $\rho = \rho(s)$ is independent of the initial state s and can also be computed via the stationary distribution μ of the underlying Markov chain. More precisely, in this case it holds by the *ergodic theorem* (see e.g. Theorem 1.10.2 of Norris, 1998) that

$$\rho \,=\, \boldsymbol{\mu}^\top \mathbf{r} \,=\, \sum_{s \in S} \mu(s)\, r(s). \tag{3}$$

That is, the average reward ρ is what one obtains in the stationary distribution.

2.2. The Bias

According to (3) the time to stationarity of the underlying Markov chain is related to the question of how long it takes until the average reward is reached, which is obviously an important question in many applications. However, the structure of the reward functions also plays a crucial role here. For example, if all rewards are the same, convergence to the stationary distribution is not important for reaching the average reward. This is made more precise by the notion of *bias*. In MRPs in which the underlying Markov chain is aperiodic the bias $\lambda(s)$ of a state s is defined as

$$\lambda(s) = \mathbb{E}\Big[\sum_{t=1}^{\infty} \big(r(S_t) - \rho(S_t)\big) \,\Big|\, S_1 = s\Big]. \tag{4}$$

Thus, the bias $\lambda(s)$ measures the difference between the mean accumulated reward and the average reward ρ when starting in state s. For periodic chains one has to consider the Cesaro limit and can define the bias more generally as

$$\lambda(s) = \lim_{T\to\infty} \frac{1}{T} \sum_{\tau=1}^{T} \mathbb{E}\Big[\sum_{t=1}^{\tau} \big(r(S_t) - \rho(S_t)\big) \,\Big|\, S_1 = s\Big]. \tag{5}$$

The difference $\lambda(s) - \lambda(s')$ for two states s, s' intuitively quantifies the (dis-)advantage in accumulated reward of starting in state s over starting in state s'. This is exemplified in the following example, which is a slightly modified version of Example 8.2.1 of Puterman (1994).

Example 2. *Consider the periodic two state MRP with* $\mathcal{S} = \{0, 1\}$*, transition matrix*

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and deterministic rewards $r(0) = 0$*,* $r(1) = 1$*. The average reward obviously is* $\frac{1}{2}$*. However, depending on the initial state the observed reward sequences will be different:*

$$\begin{aligned} &\textit{initial state 0:} \quad 0, 1, 0, 1, 0, 1, 0, 1, \ldots \\ &\textit{initial state 1:} \quad 1, 0, 1, 0, 1, 0, 1, 0, \ldots \end{aligned}$$

Accordingly, the corresponding sequences of accumulated rewards are

$$\textit{initial state 0:} \quad 0, 1, 1, 2, 2, 3, 3, 4, \ldots \tag{6}$$

$$\textit{initial state 1:} \quad 1, 1, 2, 2, 3, 3, 4, 4, \ldots \tag{7}$$

Starting in state 1 thus yields an advantage that can be quantified by the bias. For computing it according to (5) *we have to compare the actual rewards to the average reward* $\frac{1}{2}$*. More precisely, we consider the sequence of accumulated average rewards, which is*

$$\tfrac{1}{2}, 1, \tfrac{3}{2}, 2, \tfrac{5}{2}, 3, \tfrac{7}{2}, 4, \ldots.$$

Subtracting this sequence from the sequences of accumulated rewards in (6) *and* (7) *gives the following sequences:*

$$\begin{array}{ll} \textit{initial state 0:} & -\tfrac{1}{2}, 0, -\tfrac{1}{2}, 0, -\tfrac{1}{2}, 0, -\tfrac{1}{2}, 0, \ldots \\ \textit{initial state 1:} & +\tfrac{1}{2}, 0, +\tfrac{1}{2}, 0, +\tfrac{1}{2}, 0, +\tfrac{1}{2}, 0, \ldots \end{array}$$

Computing the average value of these sequences gives the bias in each state, that is, $\lambda(0) = -\frac{1}{4}$ *and* $\lambda(1) = \frac{1}{4}$*. The difference* $\lambda(1) - \lambda(0) = \frac{1}{2}$ *is the average difference of the sequences of accumulated reward in* (6) *and* (7)*.*

According to the definition of the bias, it can also be used to bound the difference between the average reward and the actual accumulated reward over a finite number of T steps.

Proposition 3. *In an irreducible and aperiodic MRP it holds that*

$$|v_T(s) - T\rho| \le \operatorname{span}(\lambda) := \max_{s'} \lambda(s') - \min_{s'} \lambda(s').$$

Proof. As the underlying chain is irreducible, ρ is constant and we can rewrite (4) as

$$\begin{aligned} \lambda(s) &= \mathbb{E}\Big[\sum_{t=1}^{T} \big(r(S_t) - \rho\big) \,\Big|\, S_1 = s\Big] + \mathbb{E}\Big[\sum_{t=T+1}^{\infty} \big(r(S_t) - \rho\big) \,\Big|\, S_1 = s\Big] \\ &= v_T(s) - T\rho + \mathbb{E}\Big[\sum_{t=T+1}^{\infty} \big(r(S_t) - \rho\big) \,\Big|\, S_1 = s\Big]. \end{aligned}$$

Since

$$\min_{s'} \lambda(s') \le \mathbb{E}\Big[\sum_{t=T+1}^{\infty} \big(r(S_t) - \rho\big) \,\Big|\, S_1 = s\Big] \le \max_{s'} \lambda(s'),$$

the proposition follows. □

Proposition 3 holds more generally also for MRPs that are just unichain and may be periodic, cf. Exercise 38.17 of Lattimore and Szepesvári (2020), p. 479.

2.3. Bounding the Bias Span

Interestingly, the term $\mathrm{span}(\lambda)$, which we will refer to as the *bias span* in the following, can be upper bounded in terms of the mean first passage times $\tau(s,s')$. Intuitively, state s cannot have an accumulated advantage of more than $\tau(s,s')$ over state s': This is the maximal expected time to reach s' from s and per step one cannot lose more than the maximal reward of 1.

Proposition 4. *In an irreducible and aperiodic MRP it holds that*

$$\lambda(s) - \lambda(s') \;\leq\; \tau(s,s').$$

Proof. Let $T_{s,s'}$ be the random time of the first visit to state s' when starting in state s. Then we can rewrite the definition of the bias $\lambda(s)$ in (4) as

$$\lambda(s) = \mathbb{E}\Bigg[\sum_{t=1}^{T_{s,s'}} \big(r(S_t)-\rho\big)\,\Big|\, S_1 = s\Bigg] + \mathbb{E}\Bigg[\sum_{t=T_{s,s'}+1}^{\infty} \big(r(S_t)-\rho\big)\,\Big|\, S_1 = s\Bigg].$$

The second term is $\lambda(s')$ and since the rewards $r(S_t)$ are assumed to be in $[0,1]$, it follows that

$$\lambda(s) - \lambda(s') \;=\; \mathbb{E}\Bigg[\sum_{t=1}^{T_{s,s'}} \big(r(S_t)-\rho\big)\,\Big|\, S_1 = s\Bigg] \;\leq\; \tau(s,s'), \qquad \square$$

As Proposition 3, also Proposition 4 holds more generally for periodic MRPs, cf. Exercise 38.13 of Lattimore and Szepesvári (2020), p. 478.

2.4. The Poisson Equation

The *Poisson equation* (cf. Theorem 8.2.6 of Puterman, 1994)

$$\boldsymbol{\rho} + \boldsymbol{\lambda} \;=\; \mathbf{r} + \mathbf{P}\boldsymbol{\lambda} \tag{8}$$

formalizes the connection between the average reward, the individual rewards in the single states, and the bias. However, the bias λ is determined by (8) only up to an additive constant. In addition, λ as defined in (5) has to satisfy $\boldsymbol{\mu}^\top\boldsymbol{\lambda} = 0$, cf. eq. 8.2.2 and Theorem 8.2.6 c of Puterman (1994).

3. Markov Decision Processes

The notion of *Markov decision process (MDP)* generalizes MRPs as follows. At each time step $t = 0, 1, 2, \dots$ one observes the current state S_t and chooses an *action* A_t from a given set of available actions $\mathcal{A}$. After choosing A_t one obtains a random reward R_t and observes the transition to the next state S_{t+1}. The transition probabilities as well as the reward distributions now depend not only on the state S_t but also on the chosen action A_t. The precise definition of Markov decision process we are going to work with is the following.

Definition 5. *A Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, r, p, S_1)$ consists of*

(i) a set of states $\mathcal{S}$,

(ii) a set of actions $\mathcal{A}$ available in each state s in $\mathcal{S}$,

(iii) the mean reward functions $r(s, a)$ that specify the expected value of the random reward for choosing action a in state s,

(iv) the transition probability distributions $p(\cdot|s, a)$ that for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$ specify the probability for a transition to a next state. That is, when choosing action a in state s at step t the next state at step $t + 1$ is s' with probability $p(s'|s, a)$,

(v) an initial state $S_1 \in \mathcal{S}$.

There are a few variations considered in the literature. Thus, in many applications it makes sense to have different sets of actions available in each state. As this complicates notation, one usually tries to convert this setting to the one given above by adding dummy actions and renaming actions if necessary. Rewards are sometimes considered simpler only to be state dependent, or more generally to also depend on the next state. Further, just as for Markov chains, the initial state often more generally can be considered to be an initial distribution over $\mathcal{S}$. None of these variations makes a big difference however and results usually can be easily adapted.

Although MDPs with continuous state-action space are important in practice, in the following we will consider only the setting in which there are finitely many states and actions. As for Markov reward processes we are further going to assume that the support of the reward distributions is contained in the unit interval $[0, 1]$. Note that an MRP can be considered to be an MDP whose action space consists of a single action.

3.1. Stochastic Process and Policies

In the following, we write S_t, A_t, R_t for the state, the action, and the reward obtained at step t, respectively. The *history* H_t at time step t is the sequence of states, actions, and rewards up to and including step t, that is, $H_t = (S_1, A_1, R_1, \dots, S_{t-1}, A_{t-1}, R_{t-1}, S_t)$. According to the definition we have for $t = 1, 2, \dots$

$$\begin{aligned} \mathbb{P}[S_{t+1} = s' | H_t, A_t] &= \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a] = p(s'|s, a), \\ \mathbb{E}[R_t | S_t = s, A_t = a] &= r(s, a). \end{aligned}$$

Thus, ignoring the rewards one obtains for a given sequence of actions $A_1, A_2, \dots$ a time-inhomogeneous Markov chain. Obviously there are many ways to obtain an action sequence, which are subsumed under the term *policy*. Thus, a *policy* is a function that maps histories to (probability distributions over) actions. Usually however simpler sub-classes of policies are considered. A policy π which is independent of time t is called *stationary* and it is *deterministic* if it maps states to actions, that is, $\pi : \mathcal{S} \to \mathcal{A}$. As we will see, for the usual purposes it is sufficient to consider stationary deterministic policies, which choose the same action whenever in the same state. Note that any stationary deterministic policy induces an MRP.

3.2. Transition Structure and Diameter

Before turning our focus to the rewards and the respective optimization criteria, we consider the transition structure of an MDP. In general, just like a Markov chain an MDP can be *multichain*, that is, at least one policy induces a Markov chain that is multichain. In reinforcement learning a learner has to find an optimal policy (with respect to criteria that will be introduced below) in an environment modeled by an MDP, however unknown to her. When learning *online*, that is, on the fly, she inevitably has to explore different actions and states and will make mistakes, that is, choose suboptimal actions. Such mistakes can be nonrecoverable when the learner ends up in a part of the state space that gives suboptimal reward and cannot be left anymore. While the real world is full of such examples that obviously cannot be learned by trial and error without any further knowledge, here we will focus on MDPs with transition structures which allow a recovery from choosing a bad action.

Definition 6. *An MDP is communicating if for any two states s, s' there is a policy π such that s' is reachable from s with positive probability when following π.*

For a given MDP it can be checked in polynomial time whether it is communicating (Kallenberg, 2002). A common generalization of communicating MDPs are *weakly communicating* MDPs, which may have some transient states, that are visited only a finite number of times under any policy. While it is possible to consider this more general setting, transient states complicate some matters so that in the following we will assume MDPs to be communicating.

In a communicating MDP any state can be reached from any other state. The maximal expected time it takes to connect any two states is the *diameter*, which is a generalization of the respective concept in graphs.

Definition 7. *Given an MDP M, let $\tau_\pi(s, s')$ be the expected time it takes to reach s' from s when following π. The diameter D of M is defined as the maximal expected time it takes to connect any two states in M, that is,*

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

Obviously, the diameter will be finite if and only if the MDP is communicating. Note that the policy to reach a state s' from another state s will in general be specific to the states s, s'. That is, there need not be a single policy connecting any two states, except it is allowed to be randomized. Accordingly, an MDP with finite diameter in general will be multichain, i.e., some policies will induce multichain Markov chains where certain states are separated from each other.

A much stronger assumption on an MDP than having finite diameter is that each (stationary deterministic) policy induces an irreducible Markov chain. Such an MDP is called *ergodic* (or *recurrent*).

3.3. Optimization Criteria

For a given MDP one is usually interested in maximizing reward.[5] There are different criteria one may assume. The *discounted reward*

$$\sum_{t=1}^{\infty} \gamma^t \, \mathbb{E}[r(S_t, A_t)] \tag{9}$$

for $0 \leq \gamma < 1$ is commonly used, as discounting is the easiest way to take care of convergence of the infinite sum. The downside is that rewards in the far future do not have much influence anymore. Accordingly, one can alternatively consider the *finite horizon reward*

$$\sum_{t=1}^{T} \mathbb{E}[r(S_t, A_t)]$$

for some T. Here we consider the *average reward*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r(S_t, A_t)], \tag{10}$$

which we think is the most natural criterion and also the one where Markov chain theory is most useful.[6] However, optimization with respect to average reward is also quite subtle for various reasons. First of all, the limit in (10) need not exist for all policies, cf. Example 8.1.1 of Puterman (1994) for a case of an MDP and a history-dependent policy, where the $\liminf$ and the $\limsup$ counterparts of (10) do not coincide. Luckily, the limit exists at least for stationary policies (Proposition 8.1.1 of Puterman, 1994) so that for an initial state s the average reward of a stationary policy π can be defined as

$$\rho(\pi, s) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r(S_t, \pi(S_t)) | S_1 = s].$$

[5] In the operations research community it is quite common to equivalently consider costs to be minimized instead.

[6] Actually, by modifying the MDP adding a state with specific transition probabilities one can also simulate the discounted reward criterion in terms of an average reward MDP, cf. Sec. 5.4 of Norris (1998). Also note that Proposition 3 shows that the difference between finite horizon and average reward can be bounded.

Note that the average reward coincides with that of the MRP induced by policy π, cf. eq. (2). An optimal stationary deterministic policy π^* has *optimal average reward*

$$\rho^* = \max_{\pi:\mathcal{S}\to\mathcal{A}} \rho(\pi, s),$$

which in a communicating MDP is independent of the initial state s and just like (3) can be written in terms of the stationary distribution μ_{π^*} induced by π^* as

$$\boldsymbol{\mu}_{\pi^*}^\top \mathbf{r}_{\pi^*} = \sum_{s\in S} \mu_\pi(s)\, r(s, \pi^*(s)).$$

It turns out that the optimal average reward given by stationary deterministic policies cannot be increased when considering more general (i.e., randomized or history-dependent) policies.

Theorem 8 (Theorem 9.1.8 of Puterman, 1994)**.** *In any communicating MDP with finite state and action space there is a stationary deterministic optimal policy.*

Accordingly, in the following if not mentioned otherwise, by a policy we shall mean a stationary and deterministic policy. We note that in general the optimal policy need not be unique even when only considering stationary deterministic policies. To complicate matters, in terms of total collected reward one optimal policy may perform better than others as the following example demonstrates.

Example 9. *Consider a two state MDP with states* 0 *and* 1 *giving independent of the chosen action deterministic reward 0 and 1, respectively. We assume that state* 0 *is transient, while state* 1 *is absorbing, that is, it cannot be left anymore. More precisely, there are two actions a and a', which coincide in state* 1 *and have different transition probabilities in state 0. The respective transition matrices under a and a' are given by*

$$\mathbf{P}_a = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{P}_{a'} = \begin{pmatrix} 0.99 & 0.01 \\ 0 & 1 \end{pmatrix}.$$

For the average reward only the reward obtained in the absorbing state matters, so that both policies π, π' with $\pi(0) = a$ and $\pi(0) = a'$ give optimal average reward 1. However, obviously policy π is preferable as the absorbing high reward state 1 is reached much faster.

Examples like this have led to more refined optimality criteria such as *bias optimality* (Lewis and Puterman, 2002) and *Blackwell optimality* (Hordijk and Yushkevich, 2002). The latter implies the former and also presents another connection between average and discounted reward: A Blackwell optimal policy gives optimal discounted reward as defined in (9) for all values of $\gamma > \gamma_0$ for some $0 < \gamma_0 < 1$.

While here we assume only finite MDPs, when state and action space are infinite the existence of an optimal policy usually depends on further properties of the MDP. Respective necessary and sufficient conditions are the subject of ongoing research in operations research. In reinforcement learning it is usually simply assumed that such an optimal policy exists.

3.4. Optimality Equations and Algorithms

By Theorem 8 there is always a stationary deterministic policy giving optimal average reward. Since there are only finitely many such policies, in principle one could obtain an optimal policy by evaluating the induced MRP for each single policy. As there are however $|\mathcal{A}|^{|\mathcal{S}|}$ policies this approach is prohibitive in practice.

A better way to find an optimal policy is via so-called *optimality equations*. For communicating MDPs this system of equations is given by

$$\rho + \lambda(s) = \max_a \Big\{ r(s,a) + \sum_{s'} p(s'|s,a)\,\lambda(s') \Big\} \tag{11}$$

for each state s. Writing $\mathbf{r}_a$ and $\mathbf{P}_a$ for the mean reward vector and the transition matrix under action a, respectively, in vector notation one has

$$\boldsymbol{\rho} + \boldsymbol{\lambda} \;=\; \max_a \Big\{ \mathbf{r}_a + \mathbf{P}_a\,\boldsymbol{\lambda} \Big\}, \tag{12}$$

which is known as the *Bellman (optimality) equation*. Under our assumption that the MDP is communicating the optimal average reward is independent of the initial state so that the vector $\boldsymbol{\rho}$ will consist of identical entries ρ, i.e., $\boldsymbol{\rho} = \rho\mathbf{1}$, where $\mathbf{1}$ is the vector with all entries 1. Note that the Poisson equation (8) for MRPs is a special case of (12). One can show that in finite communicating MDPs there is always a solution (ρ, λ) to (12), cf. Theorem 9.1.4 of Puterman (1994). Further, in this case ρ is the optimal average reward, cf. Theorem 9.1.3 of Puterman (1994). Analogously to the Poisson equation λ is not unique and if (ρ, λ) is a solution of (12) so is $(\rho, \lambda + c)$. This does not mean however that

λ is always unique up to translation, cf. Exercise 38.11 on p. 478 of Lattimore and Szepesvári (2020). Importantly, from a solution (ρ, λ) of (11) one can construct an optimal stationary deterministic policy π^* by choosing in all states s

$$\pi^*(s) := \arg\max_a \Big\{ r(s,a) + \sum_{s'} p(s'|s,a)\, \lambda(s') \Big\},$$

cf. Theorem 9.1.7 of Puterman (1994). While the Bellman equation allows to find a policy that gives optimal average reward, there may be however other policies π that give optimal average reward ρ, which with the respective bias λ_π does not satisfy (12), cf. Example 8.4.3 of Puterman (1994).

Using the Bellman equation an optimal policy can be computed by linear programming, for details see Sections 8.8, 9.3, and 9.5.2 of (Puterman, 1994). However, there are more practical alternatives like *policy iteration* or *value iteration*. For the former we refer to Sections 8.6, 9.2, and 9.5.1, for the latter to Sections 8.5, 9.4, and 9.5.3 of (Puterman, 1994).

We conclude with the following result, which shows that for a solution of the Bellman equation one obtains a generalization of Proposition 4 to MDPs.

Proposition 10 (Lemma 38.3 of Lattimore and Szepesvári, 2020)**.** *Let* (ρ, λ) *be a solution of the Bellman equation* (11) *in a communicating MDP with rewards in* $[0, 1]$*. Then the bias span of* λ *is bounded by the diameter.*[7]

4. Markov Chain and MDP Approximation

In the context of reinforcement learning, the learner is usually considered to have no or only very limited knowledge of the underlying MDP. In particular, the rewards and transition probabilities are unknown and have to be estimated from samples taken when interacting with the MDP.

When the learner has observations of rewards and transition probabilities for each state-action pair, she can compute corresponding estimates and thus obtains an empirical estimate of the underlying MDP. Obviously, it is an interesting and practically relevant question how well this empirical MDP approximates the true one. Naturally, similar questions have also been considered in the Markov chain literature often under the notion of *perturbation bounds.*

[7] More generally, for arbitrary rewards not necessarily bounded in $[0, 1]$, it holds that $\mathrm{span}(\lambda) \leq D\, \mathrm{span}(r)$, cf. Bartlett and Tewari (2009).

4.1. Perturbation Bounds for Markov Chains

The common setting for perturbation bounds is the following. Given an irreducible Markov chain with transition matrix $\mathbf{P}$ and a perturbed and also irreducible Markov chain with transition matrix $\mathbf{P}'$, the question is how well does the stationary distribution $\boldsymbol{\mu}'$ computed from $\mathbf{P}'$ approximate the true stationary distribution $\boldsymbol{\mu}$. Usually, one obtains results that bound the distance $\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_p$ in terms of $\|\mathbf{P} - \mathbf{P}'\|_q$.

The assumption of both considered Markov chains being irreducible seems quite strong when thinking of applications to MDPs and reinforcement learning. Recall in particular that even in communicating MDPs not all policies induce an irreducible Markov chain. Actually, the perturbation results hold more generally in arbitrary Markov chains having the same recurrent classes and starting in the same initial state. While this is still difficult to guarantee in the reinforcement learning setting, one is usually interested in the optimal policy which in communicating MDPs always induces a unichain Markov chain with a single recurrent class. Moreover, we shall see that perturbation results hold under weaker conditions than usually assumed in the literature.

Before taking a closer look at Markov chain perturbation we briefly mention a different, yet related line of research that considers *sensitivity* as a derivative. Caswell (2013) investigates not only the sensitivity of the stationary distribution but also other parameters of the Markov chain such as mean first passage times or Kemeny's constant (cf. below). Interesting applications in ecology and demography are also provided and investigated in more depth in (Caswell, 2019).

4.1.1. Condition Numbers and Perturbation

Perturbation bounds not only depend on the distance $\|\mathbf{P} - \mathbf{P}'\|_q$ but also on an additional parameter of the true Markov chain[8] called *condition number* by Cho and Meyer (2001). Cho and Meyer (2001) compare various perturbation bounds in the literature and note that most of the considered condition numbers either depend on the group inverse[9] of $\mathbf{I} - \mathbf{P}$ or on the fundamental matrix $(\mathbf{I} - \mathbf{P} + \mathbf{1}\boldsymbol{\mu}^\top)^{-1}$. A more intuitive condition number

[8] Actually, the situation is symmetric, so that one can instead also consider a parameter of $\mathbf{P}'$ instead.

[9] The group inverse of a matrix $\mathbf{A}$ is the unique matrix $\mathbf{A}^\#$ for which $\mathbf{A}\mathbf{A}^\#\mathbf{A} = \mathbf{A}$, $\mathbf{A}^\#\mathbf{A}\mathbf{A}^\# = \mathbf{A}^\#$, and $\mathbf{A}\mathbf{A}^\# = \mathbf{A}^\#\mathbf{A}$.

defined in terms of mean passage times has been introduced by Cho and Meyer (2000).

Theorem 11 (Cho and Meyer, 2000)**.** *Let* $\mathbf{P}$, $\mathbf{P}'$ *be the transition matrices of two irreducible Markov chains with stationary distributions* μ, μ', *respectively. Then*

$$|\mu(s) - \mu'(s)| \leq \frac{\max_{s' \neq s} \tau(s', s)}{2\tau(s, s)} \cdot \|\mathbf{P} - \mathbf{P}'\|_\infty.$$

It is well-known that $\tau(s,s) = \frac{1}{\mu(s)}$ so that one immediately obtains the following corollary.

Corollary 12. *Let* $\mathbf{P}$, $\mathbf{P}'$ *be the transition matrices of two irreducible Markov chains with stationary distributions* μ, μ', *respectively. Then*

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_\infty \leq \tfrac{1}{2} \max_s \left\{ \mu(s) \max_{s' \neq s} \tau(s', s) \right\} \|\mathbf{P} - \mathbf{P}'\|_\infty.$$

The double max quantity on the right hand side is reminiscent of the definition of the diameter (cf. Definition 7), here weighted by the stationary distribution. Actually as noted by Hunter (2005), since $\mu(s) \leq 1$ one also arrives at a bound in terms of the diameter $D := \max_{s' \neq s} \tau(s', s)$ of the Markov chain, that is,

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_\infty \leq \tfrac{1}{2} D \, \|\mathbf{P} - \mathbf{P}'\|_\infty. \tag{13}$$

Example 20 below presents a simple Markov chain for which the error is indeed of order $D \, \|\mathbf{P} - \mathbf{P}'\|_\infty$.

Hunter (2005) presents a more general perturbation bound in terms of generalized matrix inverses that subsumes several of the bounds in the literature, including Theorem 11. Further he also investigates a few special cases (like perturbations in a single row of the transition matrix) where improved bounds are possible.

Beside the diameter the condition number of Corollary 12 also resembles the parameter

$$\bar{\eta}(s) := \sum_{s'} \mu(s') \, \tau(s, s'). \tag{14}$$

In irreducible Markov chains this quantity is actually independent of s. That is, for all states s it holds that

$$\bar{\eta}(s) = \bar{\eta},$$

and $\bar{\eta}$ is known as *Kemeny's constant* (Kemeny and Snell, 1960). The terms $\bar{\eta}(s)$ can be interpreted as the expected time to reach the stationary distribution in the following sense: One first picks a state s' randomly according to the stationary distribution and then considers how long it takes in expectation to get there, starting in state s. While this is a neat interpretation of $\bar{\eta}(s)$ it does not give an intuitive explanation why these terms are constant, a question which has not been answered in a satisfactory way yet. More on different interpretations and a history of Kemeny's constant as well as an overview of related results can be found in (Hunter, 2014).

In the following, instead of $\bar{\eta}$ we will use the paramater $\eta := \bar{\eta} - 1$, which sometimes is referred to as Kemeny's constant as well. It naturally arises when setting $\tau(s,s) = 0$ instead of $\frac{1}{\mu(s)}$ in (14), cf. Hunter (2014). Unlike the previous perturbation results the following theorem bounds the error in 1-norm, which will be most useful for our purposes.

Theorem 13 (Hunter, 2006)**.** *Let* $\mathbf{P}$*,* $\mathbf{P}'$ *be the transition matrices of two irreducible Markov chains with stationary distributions* $\boldsymbol{\mu}$*,* $\boldsymbol{\mu}'$*, respectively. Then*

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_1 \leq \tfrac{\eta}{2}\, \|\mathbf{P} - \mathbf{P}'\|_\infty.$$

While the mentioned perturbation results in the literature assume that both Markov chains are irreducible, the results sometimes hold more generally as the following simple example demonstrates.

Example 14. *Consider a general two state Markov chain with* $S = \{1,2\}$ *and transition matrix*

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

where we assume that $0 < p, q < 1$*. The stationary distribution is given by* $\boldsymbol{\mu} = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$*, the mean first passage times are* $\tau(1,2) = \frac{1}{p}$*,* $\tau(2,1) = \frac{1}{q}$*, and Kemeny's constant* $\eta = \frac{2}{p+q}$*. Comparing the stationary distribution* $\boldsymbol{\mu}$ *of* $\mathbf{P}$ *to the stationary distribution* $\boldsymbol{\mu}' = (1,0)$ *of the Markov chain with transition matrix* $\mathbf{P}' = \mathbf{I}$ *and initial state 1, we see that the respective error is*

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_\infty = \max\left\{1 - \frac{q}{p+q}, \frac{p}{p+q}\right\} = \frac{p}{p+q}. \tag{15}$$

Although the assumptions of Corollary 12 are not satisfied since $\mathbf{P}'$ *is not irreducible, it still provides a valid upper bound on the error in* (15)*. That is,* *we have*

$$\begin{aligned}
&\tfrac{1}{2}\max_{s}\big\{\mu(s)\max_{s'\neq s}\tau(s',s)\big\}\,\|\mathbf{P}-\mathbf{P}'\|_\infty \\
&= \tfrac{1}{2}\max\left\{\frac{q}{p+q}\cdot\frac{1}{q},\frac{p}{p+q}\cdot\frac{1}{p}\right\}\max\{2p,2q\} \\
&= \frac{\max\{p,q\}}{p+q}.
\end{aligned}$$

The case when the initial state is 2 is of course symmetric. Similarly, we have $\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_1 = \frac{2p}{p+q}$ *and the bound of Theorem 13 gives* $\frac{2\max\{p,q\}}{p+q}$*.*

This example is quite simple and it is an interesting question which perturbation bounds also hold when the perturbed Markov chain is not irreducible.[10] In Corollary 18 below we provide a respective generalization of the bound in (13). Note that the perturbation bounds presented in this section become vacuous when both Markov chains are not irreducible and condition numbers based on mean first passage times are not finite anymore.

4.2. Markov Chain and MDP Approximations

Equipped with the perturbation results of the previous section we are now going to derive bounds for more general MDP approximations. As Theorem 13 fits our purposes best, in the following we will use this result, although in principle other bounds could be used as well.

MDP Notation. We introduce some notation first. For an MDP M we denote by $\rho(M,\pi,s)$ the average reward of policy π in M when starting in s. If the reward is independent of s we simply use $\rho(M,\pi)$. Further, we write $\rho^*(M)$ for the optimal average reward of M. In case the optimal average reward depends on the initial state s, we use $\rho^*(M,s)$. Finally, we introduce notation for Kemeny's constant, the condition number used in Theorem 13. Consider for a policy π the Markov chain induced by π on M. This Markov chain may be multi-chain so that we write $\eta(M,\pi,s)$ for Kemeny's constant

[10] Recently, we were able to answer this question by providing an explicit expression of the bias in an MRP in terms of mean first passage times. This not only leads to an improvement of Theorem 13 but also shows that the perturbation bounds of Theorems 11 and 13 hold more generally. The results also give a new interpretation of Kemeny's constant in terms of the bias in MRPs with constant reward. Due to time and space constraints, these results could not be included in this chapter. For details we refer to (Ortner, 2024).

of the recurrent class containing state s. As before, if there is only a single irreducible class, we skip the notation for the initial state and write $\eta(M,\pi)$.

We now introduce the notion of $(\varepsilon_r,\varepsilon_p)$-*approximation* of an MDP where the parameters $\varepsilon_r,\varepsilon_p$ intuitively provide an upper bound on the amount of perturbation applied to rewards and transition probabilities, respectively.

Definition 15. *Given two MDPs $M=(\mathcal{S},\mathcal{A},r,p)$, $\bar{M}=(\mathcal{S},\mathcal{A},\bar{r},\bar{p})$ over the same state-action space we call $\bar{M}$ an $(\varepsilon_r,\varepsilon_p)$-approximation of M if for all s in $\mathcal{S}$ and all a in $\mathcal{A}$,*

$$\begin{aligned}\big|\bar{r}(s,a)-r(s,a)\big| &\leq \varepsilon_r, &(16)\\ \sum_{s'\in\mathcal{S}}\big|\bar{p}(s'|s,a)-p(s'|s,a)\big| &\leq \varepsilon_p. &(17)\end{aligned}$$

Note that the roles of M and $\bar{M}$ in this definition are symmetric. That is, if $\bar{M}$ is an $(\varepsilon_r,\varepsilon_p)$-approximation of M then M is also an $(\varepsilon_r,\varepsilon_p)$-approximation of $\bar{M}$.

Condition (17) fits the perturbation bounds of the previous section. That is, let $\mathbf{P}_a,\bar{\mathbf{P}}_a$ be the transition matrices of MDPs M and $\bar{M}$ under action a. Then (17) implies that $\|\mathbf{P}_a-\bar{\mathbf{P}}_a\|_\infty\leq\varepsilon_p$ for all actions a and more generally $\|\mathbf{P}_\pi-\bar{\mathbf{P}}_\pi\|_\infty\leq\varepsilon_p$ for the transition matrices $\mathbf{P}_\pi,\bar{\mathbf{P}}_\pi$ of any stationary deterministic policy π. Accordingly, the presented results about perturbation can be directly applied to the Markov chains induced by any policy π.

Still, in principle one could also consider alternative formulations of condition (17) using e.g. different norms or just bounding the difference of single transition probabilities instead of the sum. However, this would change only a few technical details in what follows.

In Section 6 on Markov chain and MDP aggregation below we will introduce a more general definition of approximation where the underlying state spaces of the two MDPs need not be the same.

Given Definition 15 one can obtain error bounds for MDP approximation as corollary to Theorem 13.

Corollary 16. *Let $\bar{M}$ be an $(\varepsilon_r,\varepsilon_p)$-approximation of an MDP M.*

(i) For any policy π that induces the same transition structure on M and $\bar{M}$ it holds that

$$|\rho(M,\pi,s)-\rho(\bar{M},\pi,s)|\leq\varepsilon_r+\tfrac{1}{2}\,\eta(M,\pi,s)\,\varepsilon_p. \tag{18}$$

(ii) Assume that M and $\bar{M}$ both have state independent optimal average reward ρ^* and $\bar{\rho}^*$, respectively. Let π^*, $\bar{\pi}^*$ be corresponding optimal policies on M and $\bar{M}$, and assume that they induce Markov chains with the same transition structure on M and $\bar{M}$. Then

$$\begin{aligned} |\rho^* - \bar{\rho}^*| &\leq \varepsilon_r + \tfrac{1}{2}\max\{\eta(M,\pi^*),\eta(M,\bar{\pi}^*)\}\,\varepsilon_p, \ \textit{and} \qquad (19) \\ \rho(M,\bar{\pi}^*) &\geq \rho^* - 2\varepsilon_r - \tfrac{1}{2}\big(\eta(M,\pi^*)+\eta(M,\bar{\pi}^*)\big)\,\varepsilon_p. \qquad (20) \end{aligned}$$

Proof. For (i) we consider the Markov chains induced by π on M and $\bar{M}$ with stationary distributions μ and $\bar{\mu}$, respectively. Writing r and $\bar{r}$ for the mean rewards under π in M and $\bar{M}$, we have

$$\begin{aligned} \big|\rho(M,\pi,s) - \rho(\bar{M},\pi,s)\big| &= \Big|\sum_{s'}\mu(s')\,r(s') - \sum_{s'}\bar{\mu}(s')\,\bar{r}(s')\Big| \\ &= \Big|\sum_{s'}\big(\mu(s')-\bar{\mu}(s')\big)\,r(s') + \sum_{s'}\bar{\mu}(s')\big(r(s')-\bar{r}(s')\big)\Big| \\ &\leq \sum_{s'}\big|\mu(s')-\bar{\mu}(s')\big|\,r(s') + \sum_{s'}\bar{\mu}(s')\big|r(s')-\bar{r}(s')\big| \\ &\leq \|\boldsymbol{\mu}-\bar{\boldsymbol{\mu}}\|_1 + \sum_{s'}\bar{\mu}(s')\,\varepsilon_r \\ &\leq \tfrac{1}{2}\,\eta(M,\pi,s)\,\varepsilon_p + \varepsilon_r, \end{aligned}$$

using Theorem 13 and (16) in the final step.

Concerning (ii), first assume that $\rho^* \geq \bar{\rho}^*$. Then

$$\rho^* - \bar{\rho}^* \ \leq\ \rho(M,\pi^*) - \rho(\bar{M},\pi^*) \ \leq\ \varepsilon_r + \tfrac{1}{2}\eta(M,\pi,s)\,\varepsilon_p$$

according to (18). The case when $\rho^* \leq \bar{\rho}^*$ is symmetric and (19) follows.

Finally,

$$\begin{aligned} \rho^* - \rho(M,\bar{\pi}^*) &\leq \rho^* - \rho(M,\bar{\pi}^*) + \bar{\rho}^* - \rho(\bar{M},\pi^*) \\ &\leq \big|\rho(M,\pi^*) - \rho(\bar{M},\pi^*)\big| + \big|\rho(\bar{M},\bar{\pi}^*) - \rho(M,\bar{\pi}^*)\big| \end{aligned}$$

and bounding both of these terms by (18) yields (20). □

The condition of having the same transition structure is obviously quite restrictive when considering reinforcement learning applications. In an unknown MDP environment it may take long until the learner has correctly identified all triples (s, a, s') for which $p(s'|s, a) > 0$. In particular, when working with the empirical MDP it may take many samples until a transition with very small probability has been observed. Moreover, *optimistic* reinforcement learning algorithms (Filippi et al., 2010) sometimes assign positive probabilities even to transitions that have never been observed (Auer and Ortner, 2007; Jaksch et al., 2010). However, as we will see in the next section at least for a suitable optimal policy one can obtain bounds under much weaker conditions.

4.3. An Improved Perturbation Bound

The following bound on the error made by an $(\varepsilon_r, \varepsilon_p)$-approximation generalizes a bound of Ortner (2007) from ergodic to communicating MDPs and slightly improves a respective bound of Ortner et al. (2014a).

Theorem 17. *Let M be a communicating MDP with diameter D and an optimal policy π^* that satisfies the Bellman equation* (12). *Then for all $(\varepsilon_r, \varepsilon_p)$-approximations $\bar{M}$ of M and any initial state s,*

$$\left|\rho^*(M) - \rho(\bar{M}, \pi^*, s)\right| \;\leq\; \varepsilon_r + \tfrac{D}{2}\,\varepsilon_p.$$

Comparing Theorem 17 to the results of the previous section, we see that there is no assumption about the transition structure of the approximation $\bar{M}$. Rather, $\bar{M}$ not even needs to be communicating. It is also easily possible to derive from Theorem 17 a corresponding result for Markov chains that states the same perturbation bound as (13) but without any irreducibility condition on the perturbed Markov chain.

Corollary 18. *Consider an irreducible Markov chain with transition matrix $\mathbf{P}$, stationary distribution $\boldsymbol{\mu}$, and diameter D. Let a perturbed Markov chain over the same state space have transition matrix $\mathbf{P}'$ and stationary distribution $\boldsymbol{\mu}'_s$ when starting in initial state s. Then for all states s,*

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'_s\|_\infty \leq \tfrac{D}{2}\,\|\mathbf{P} - \mathbf{P}'\|_\infty.$$

Proof. Choose an arbitrary state s' and add a deterministic reward function r to both Markov chains with $r(s') = 1$ and $r(s'') = 0$ for all states $s'' \neq s'$. As the reward function is the same for both arising MRPs the perturbed MRP is an $(\varepsilon_r, \varepsilon_p)$-approximation of the unperturbed MRP with $\varepsilon_r = 0$ and

$\varepsilon_p = \|\mathbf{P} - \mathbf{P}'\|_\infty$. Writing ρ and $\rho'(s)$ for the average reward in the two MRPs (when starting in state s) we get by Theorem 17

$$|\mu(s') - \mu'_s(s')| = |\rho - \rho'(s)| \;\leq\; \tfrac{D}{2}\,\varepsilon_p.$$

Since s' was chosen arbitrarily, the claim follows. □

When both the original and the approximated MDP are communicating, we can derive from Theorem 17 the following improvement over the result of Corollary 16 (ii). We note that under stronger assumptions even better bounds can be derived, cf. Section 5.2 of (Boone, 2024).

Corollary 19. *Let $\bar{M}$ be a communicating $(\varepsilon_r, \varepsilon_p)$-approximation of a communicating MDP M. Let π^* and $\bar{\pi}^*$ be optimal policies satisfying the Bellman equation* (12) *in M and $\bar{M}$, respectively. Further assume that $\rho(\bar{M}, \pi^*)$ and $\rho(M, \bar{\pi}^*)$ are independent of the initial state. Then writing $\rho^* := \rho(M, \pi^*)$, $\bar{\rho}^* := \rho(\bar{M}, \bar{\pi}^*)$ for the optimal average reward in M and $\bar{M}$, it holds that*

$$\begin{aligned} |\rho^* - \bar{\rho}^*| &\leq \varepsilon_r + \tfrac{1}{2} D\,\varepsilon_p, \text{ and} \\ \rho(M, \bar{\pi}^*) &\geq \rho^* - 2\varepsilon_r - D\,\varepsilon_p. \end{aligned}$$

The proof is the same as that for Corollary 16 (ii) now using Theorem 17 instead of (18). Note that the assumption of the approximation being communicating and having state independent average reward for policies π^* and $\bar{\pi}^*$ in both MDPs can also be established by adding small additional transition probabilities in a suitable way.

The following example provides a lower bound that shows that when the perturbation in the transition probabilities is of order ε then the error in the stationary distribution and in average reward can be of order εD.

Example 20. *Consider two Markov chains with transition matrices*

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}, \quad \mathbf{P}' = \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix},$$

where we assume $p, q < \frac{1}{2}$ and $2q > p > q$. Then the error in the transition probabilities is of order $\varepsilon := p - q$ and the stationary distributions are $\boldsymbol{\mu} = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$ and $\boldsymbol{\mu}' = \left(\frac{1}{2}, \frac{1}{2}\right)$, respectively. Hence, by definition of ε and since by assumption $q > \varepsilon$,

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_\infty \;=\; \frac{p}{p+q} - \frac{1}{2} \;=\; \frac{\varepsilon}{2(p+q)} \;=\; \frac{\varepsilon}{2(2q+\varepsilon)} \;>\; \frac{\varepsilon}{6q},$$

which is of order $D\varepsilon$, as $D = \max\left\{\frac{1}{p}, \frac{1}{q}\right\} = \frac{1}{q}$. Assigning reward 1 to one state and 0 to the other one shows that this lower bound also holds for the error in average reward.

4.4. Proof of Theorem 17

As the proof of Theorem 17 is quite instructive, we include it here in full detail.[11] Consider a communicating MDP M with optimal policy π^* and an MDP approximation $\bar{M}$ of M. The following result compares the quantity $\ell\rho^*(M)$ to the accumulated rewards in $\bar{M}$ when performing π^* for ℓ steps.

Lemma 21. *Consider a communicating MDP $M = (\mathcal{S}, \mathcal{A}, r, p)$ with diameter D and let $\bar{M} = (\mathcal{S}, \mathcal{A}, \bar{r}, \bar{p})$ be an $(\varepsilon_r, \varepsilon_p)$-approximation of M. Let π^* be an optimal policy of M that satisfies the Bellman equation* (12) *and assume that π^* is performed on $\bar{M}$ for ℓ steps. Then denoting by $\bar{v}^*(s)$ the number of times state s is visited among these ℓ steps, with probability at least $1-\delta$ it holds that*

$$\ell\rho^*(M) - \sum_{s\in\mathcal{S}} \bar{v}^*(s)\cdot\bar{r}(s,\pi^*(s)) \;\le\; \left(\varepsilon_r + \tfrac{D}{2}\cdot\varepsilon_p\right)\ell + D\sqrt{2\ell\log(1/\delta)} + D. \quad (21)$$

Proof. We abbreviate $r^*(s) := r(s,\pi^*(s))$ and $p^*(s'|s) := p(s'|s,\pi^*(s))$, and use $\bar{r}^*(s)$ and $\bar{p}^*(s'|s)$ accordingly. Then

$$\begin{aligned}\ell\rho^*(M) - \sum_{s\in\mathcal{S}} \bar{v}^*(s)\cdot\bar{r}^*(s) &= \sum_s \bar{v}^*(s)\big(\rho^*(M) - \bar{r}^*(s)\big)\\ =\;& \sum_s \bar{v}^*(s)\big(\rho^*(M) - r^*(s)\big) + \sum_s \bar{v}^*(s)\big(r^*(s) - \bar{r}^*(s)\big). \quad (22)\end{aligned}$$

For the first term in (22) we use the optimality equations (11) for π^* and replace

$$\rho^*(M) - r^*(s) = \sum_{s'} p^*(s'|s)\cdot\lambda^*(s') - \lambda^*(s),$$

writing $\lambda^* := \lambda_{\pi^*}$ for the bias of π^* on M. By Proposition 10, applying a suitable translation to $\boldsymbol{\lambda}^*$ we can assume without loss of generality that $\|\boldsymbol{\lambda}^*\|_\infty \le \frac{D}{2}$.

[11] A shorter proof can be found in (Boone, 2024), see Theorem II.1 on p. 50.

The second term in (22) can be bounded by $\ell\varepsilon_r$ according to (16) and we obtain

$$\begin{aligned}\ell\rho^*(M) - \sum_{s\in\mathcal{S}} \bar{v}^*(s)\cdot\bar{r}^*(s) \;&\leq\; \sum_s \bar{v}^*(s)\Big(\sum_{s'} p^*(s'|s)\,\lambda^*(s') - \lambda^*(s)\Big) + \ell\varepsilon_r \\ &=\; \ell\varepsilon_r + \sum_s \bar{v}^*(s)\Big(\sum_{s'} \bar{p}^*(s'|s)\,\lambda^*(s') - \lambda^*(s)\Big) \\ &\qquad + \sum_s \bar{v}^*(s)\Big(\sum_{s'} p^*(s'|s)\,\lambda^*(s') - \sum_{s'} \bar{p}^*(s'|s)\,\lambda^*(s')\Big). \qquad (23)\end{aligned}$$

The last term in (23) can be written and bounded as

$$\begin{aligned}&\sum_s \bar{v}^*(s)\sum_{s'}\Big(p^*(s'|s) - \bar{p}^*(s'|s)\Big)\,\lambda^*(s') \\ &\quad\leq\; \sum_s \bar{v}^*(s)\sum_{s'}\Big|p^*(s'|s) - \bar{p}^*(s'|s)\Big|\cdot\|\lambda^*\|_\infty \;\leq\; \ell\varepsilon_p\cdot\tfrac{D}{2}. \qquad (24)\end{aligned}$$

For the second term in (23), we write s_τ for the state visited at time step τ to obtain

$$\begin{aligned}\sum_s \bar{v}^*(s)\Big(\sum_{s'} \bar{p}^*(s'|s)\,\lambda^*(s') - \lambda^*(s)\Big) &= \sum_{\tau=1}^{\ell}\Big(\sum_{s'} \bar{p}^*(s'|s_\tau)\,\lambda^*(s') - \lambda^*(s_\tau)\Big) \\ &= \sum_{\tau=1}^{\ell}\Big(\sum_{s'} \bar{p}^*(s'|s_\tau)\,\lambda^*(s') - \lambda^*(s_{\tau+1})\Big) + \lambda^*(s_{\ell+1}) - \lambda^*(s_1). \quad (25)\end{aligned}$$

By Proposition 10, $\lambda^*(s_{\ell+1}) - \lambda^*(s_1) \leq \mathrm{span}(\lambda^*) \leq \mathrm{D}$, and the sequence

$$X_\tau := \sum_{s'} \bar{p}^*(s'|s_\tau)\,\lambda^*(s') - \lambda^*(s_{\tau+1})$$

is a martingale difference sequence with $|X_\tau| \leq D$. Thus, an application of Azuma-Hoeffding's inequality (e.g., Lemma 10 of Jaksch et al., 2010) to (25) yields that with probability $1-\delta$,

$$\sum_s \bar{v}^*(s)\Big(\sum_{s'} \bar{p}^*(s'|s)\,\lambda^*(s') - \lambda^*(s)\Big) \;\;\leq\;\; D\sqrt{2\ell\log(1/\delta)} + D. \qquad (26)$$

Combining (23), (24), and (26) we get

$$\ell\rho^*(M) - \sum_{s\in\mathcal{S}} \bar{v}^*(s)\cdot\bar{r}^*(s) \;\;\leq\;\; \Big(\varepsilon_r + \tfrac{D}{2}\cdot\varepsilon_p\Big)\,\ell + D\sqrt{2\ell\log(1/\delta)} + D,$$

which completes the proof. □

With Lemma 21 established, Theorem 17 easily follows. Divide (21) by ℓ, choose $\delta = 1/\ell$, and let $\ell \to \infty$. Since the average reward of a policy is not random, the result holds surely and not just with probability 1. $\square$

4.5. Applications

As already mentioned, in the context of reinforcement learning the results of the previous sections typically will be applied to an approximation of the underlying but unknown MDP using the empirical estimates for rewards and transition probabilities.

4.5.1. Estimation and Sufficient Precision

In order to guarantee that the empirical MDP $\bar{M}$ is a good approximation of the true MDP M, one has to make sure that the empirical estimates of the rewards and transition probabilities are close to their true values. In view of the perturbation results, which depend on the largest error, this has to hold for *all* state-action pairs. That is, assume a given target precision θ shall hold e.g. for the maximal allowed difference of the average reward of any policy in M and $\bar{M}$. Then we need to guarantee that (16) and (17) hold for all state-action pairs for sufficiently small $\varepsilon_r, \varepsilon_p$ so that in the bound of Corollary 16 (i) we have $\varepsilon_r + \frac{\eta}{2}\,\varepsilon_p \leq \theta$. Thus e.g. choosing $\varepsilon_r := \varepsilon_p := \frac{2\theta}{\eta+2}$ would guarantee the target precision θ. Note that while condition numbers of the true MDP M will usually not be known, due to the symmetric nature of the perturbation bounds the respective values for the estimated MDP $\bar{M}$ will suffice in principle. However, in the theoretical analysis the random nature of this quantity complicates matters, so that one usually tries to design algorithms that do not use condition numbers explicitly, cf. e.g. (Jaksch et al., 2010).

4.5.2. Confidence Intervals and Bounded Parameter MDPs

Confidence intervals are a widely used tool in reinforcement learning in order to quantify the reliability of the empirical estimates. This holds in particular when one aims at finite time guarantees that go beyond simple convergence to an optimal policy. The size of the confidence intervals basically corresponds to the precision achieved for the respective estimated quantities. Accordingly, the analysis of reinforcement learning algorithms often combines confidence intervals with perturbation results, see e.g. (Auer and Ortner, 2007).

More generally, MDPs with (confidence) intervals instead of single fixed values for rewards and transition probabilities have been investigated under the notion of *bounded parameter MDP (BP-MDP)* (Givan et al., 2000; Tewari and Bartlett, 2007). As in our case, the intervals usually indicate some kind of uncertainty with respect to the true values. Accordingly, beside reinforcement learning also questions of robustness have often been dealt with in the framework of BP-MDPs. On the other hand, a BP-MDP can also be considered to be a continuous set of MDPs, or alternatively, an MDP with continuous action sets. Respective optimization in an BP-MDP has also been used in the context of *optimism* in reinforcement learning (Filippi et al., 2010), and computing an MDP maximizing the optimal average reward in a BP-MDP is a common sub-routine of reinforcement learning algorithms (Auer and Ortner, 2007; Jaksch et al., 2010).

5. Estimating the Stationary Distribution from Trajectories

So far, we have computed an estimate for the stationary distribution by using estimates for the single transition probabilities, thus obtaining an empirical transition matrix. While this is quite intuitive, it is also possible to estimate the stationary distribution by just observing a trajectory of the underlying Markov chain. That way one can obtain confidence intervals for the stationary distribution of a Markov chain and the average reward of an MRP directly. The subsequent exposition presents respective results of Paulin (2015) and follows Ortner (2020). Another approach using sample paths to obtain confidence intervals for optimal average reward and other values of the MDP such as mean first passage times can be found in (Burnetas and Katehakis, 1997).

5.1. Concentration Inequalities for Uniformly Ergodic Markov Chains

In the following we assume that the considered Markov chains are *uniformly ergodic*, i.e., there are $\theta < 1$ and L such that

$$\max_{s \in \mathcal{S}} \|\boldsymbol{\mu}_s^t - \boldsymbol{\mu}\|_1 \;\leq\; L\theta^t .$$

We note that irreducible, aperiodic Markov chains are uniformly ergodic, cf. Remark 1.2 of Paulin (2015).

The first of the presented results is a version of McDiarmid's inequality for Markov chains.

Lemma 22 (Corollary 2.10 and Remark 2.11 of Paulin, 2015)**.** *Let $S_1, \ldots, S_t$ be a random sequence of states generated by a uniformly ergodic Markov chain with state space $\mathcal{S}$ and mixing time T_{mix}. Given a function $f : \mathcal{S}^t \to \mathbb{R}$ such that*

$$f(s_1, \ldots, s_t) - f(s'_1, \ldots, s'_t) \;\leq\; \sum_i c_i \, \mathbb{I}[s_i \neq s'_i] \tag{27}$$

for constants $c_1, \ldots, c_t$, it holds that

$$\mathbb{P}\Big\{ \big| f(S_1, \ldots, S_t) - \mathbb{E}[f(S_1, \ldots, S_t)] \big| \geq \varepsilon \Big\} \leq 2 \exp\left(- \frac{2\varepsilon^2}{9 \, \|\boldsymbol{c}\|_2^2 \, T_{\mathrm{mix}}} \right).$$

This result can be applied to any MRP with underlying uniformly ergodic Markov chain to obtain a concentration result for the empirical average reward after any t steps.

Corollary 23. *Consider an MRP $(\mathcal{S}, r, \mathbf{P}, S_1)$ where the Markov chain $(\mathcal{S}, \mathbf{P}, \nu)$ is uniformly ergodic with mixing time T_{mix}. Writing $\hat{\boldsymbol{\mu}}^t_{S_1}$ for the empirical distribution after t steps defined as $\hat{\mu}^t_{S_1}(s) := \frac{1}{t} \sum_{i=1}^t \mathbb{I}\{S_i = s \,|\, S_1\}$, it holds that*

$$\mathbb{P}\Big\{ \big| \boldsymbol{\mu}^{t\top}_{S_1} \mathbf{r} - \hat{\boldsymbol{\mu}}^{t\top}_{S_1} \mathbf{r} \big| \geq \varepsilon \Big\} \leq 2 \exp\left(- \frac{2\varepsilon^2 t}{9 T_{\mathrm{mix}}} \right).$$

Proof. We define $f(S_1, \ldots, S_t) := \frac{1}{t}\big(r(S_1) + \ldots + r(S_t)\big)$. Then condition (27) holds for $c_i = \frac{1}{t}$, $i = 1, \ldots, t$, so that we obtain the claim by application of Lemma 22. □

Note that the bound is trivial for $t < T_{\mathrm{mix}}(\varepsilon)$. On the other hand, when $t \geq T_{\mathrm{mix}}(\varepsilon)$, one additionally has $\|\boldsymbol{\mu}^t_{S_1} - \boldsymbol{\mu}\|_1 \leq \varepsilon$, so that also $|(\boldsymbol{\mu}^t_{S_1} - \boldsymbol{\mu})\mathbf{r}| \leq \varepsilon$. Hence, in this case one obtains

$$\mathbb{P}\Big\{ \big| \boldsymbol{\mu}^{\top} \mathbf{r} - \hat{\boldsymbol{\mu}}^{t\top}_{S_1} \mathbf{r} \big| \geq 2\varepsilon \Big\} \leq 2 \exp\left(- \frac{2\varepsilon^2 t}{9 T_{\mathrm{mix}}} \right).$$

Then fixing an error probability δ provides a confidence interval for the average reward $\rho = \boldsymbol{\mu}^\top \mathbf{r}$. That is, with probability at least $1 - \delta$, we have

$$\big| \boldsymbol{\mu}^{t\top}_{S_1} \mathbf{r} - \hat{\boldsymbol{\mu}}^{t\top}_{S_1} \mathbf{r} \big| \;\leq\; \sqrt{\frac{9 T_{\mathrm{mix}} \log \frac{2}{\delta}}{2t}} \text{ and } \big| \rho - \hat{\boldsymbol{\mu}}^{t\top}_{S_1} \mathbf{r} \big| \;\leq\; \sqrt{\frac{18 T_{\mathrm{mix}} \log \frac{2}{\delta}}{t}}.$$

Paulin (2015) also gives results for the difference between the empirical and the stationary distribution of uniformly ergodic Markov chains, now in terms of the total variation distance.

Lemma 24 (Proposition 2.18 of Paulin, 2015)**.**

$$\mathbb{P}\Big\{\big|d_{TV}(\mu,\hat{\mu}^t_{S_1}) - \mathbb{E}[d_{TV}(\mu,\hat{\mu}^t_{S_1})]\big| \geq \varepsilon\Big\} \leq 2\exp\left(-\frac{2\varepsilon^2 t}{9T_{\text{mix}}}\right).$$

Lemma 25 (Proposition 3.16 and the following remark of Paulin, 2015)**.**

$$\mathbb{E}[d_{TV}(\mu,\hat{\mu}^t_{S_1})] \leq \sum_{s\in\mathcal{S}} \min\left(\sqrt{\frac{8\mu(s)}{t\Gamma}}, \mu(s)\right),$$

where $\Gamma := \max_k \left\{\frac{\gamma(\mathbf{P}^{*k}\mathbf{P}^k)}{k}\right\}$ *is the pseudo-spectral gap of the chain. Here* $\mathbf{P}$ *is the transition kernel interpreted as linear operator,* $\mathbf{P}^*$ *is the adjoint of* $\mathbf{P}$*, and* $\gamma(\mathbf{P}^{*k}\mathbf{P}^k)$ *is the spectral gap of the self-adjoint operator* $\mathbf{P}^{*k}\mathbf{P}^k$*.*

Concerning the pseudo-spectral gap Γ more details can be found in (Paulin, 2015). For many purposes it will be sufficient to know that in uniformly ergodic Markov chains Γ can be bounded in terms of the mixing time T_{mix} as (cf. Proposition 3.4 of Paulin, 2015)

$$\tfrac{1}{\Gamma} \leq 2T_{\text{mix}}. \tag{28}$$

Combining this bound with Lemmas 24 and 25 gives the following confidence interval for the estimate of the stationary distribution.

Corollary 26. *In any uniformly ergodic Markov chain with probability at least* $1-\delta$*,*

$$d_{TV}(\mu,\hat{\mu}^t_{S_1}) \leq \sqrt{\frac{38ST_{\text{mix}}\log\frac{2}{\delta}}{t}}.$$

Proof. Set the error probability in Lemma 24 to δ. Then by (28), Lemma 25, and Jensen's inequality it follows that

$$d_{TV}(\mu,\hat{\mu}^t_{S_1}) \leq \sqrt{\frac{16ST_{\text{mix}}}{t}} + \sqrt{\frac{9T_{\text{mix}}\log\frac{2}{\delta}}{2t}} \leq \sqrt{\frac{38ST_{\text{mix}}\log\frac{2}{\delta}}{t}}. \qquad \square$$

5.2. Application to Reinforcement Learning

The results of the previous section provide an alternative route for reinforcement learning algorithms. Instead of computing estimates of transition probabilities, one evaluates trajectories of policies. Ortner (2020) suggests an algorithm that re-uses samples when evaluating policies. However with respect to computation the necessary evaluation of all policies is prohibitive in general and can only be recommended when dealing with a small set of policies to choose from as e.g. considered by Azar et al. (2013).

6. Markov Chain and MDP Aggregation

MDP aggregations lead to simplifications that speed up the computation of an optimal policy. Accordingly, MDP aggregations have a long history, see (Givan et al., 2003) for references. While a simplified policy on an MDP aggregation that performs well on the original MDP justifies the concept of aggregation for MDPs (Van Roy, 2006), for Markov chains an aggregation usually rather means a loss of information. For example, from the stationary distribution of an aggregated Markov chain it is in general not possible to obtain the stationary distribution of the original Markov chain. Thus, aggregation of Markov chains is usually considered of limited interest in itself, yet makes sense in MDP and reinforcement learning contexts. Here one is often interested only in an aggregated value such as the average reward that sometimes can be obtained from an aggregated stationary distribution. We note that there are still a few references that investigate lossless aggregation in Markov chains such as (Geiger et al., 2015).

In the following we concentrate on aggregation of MDPs and just note that any aggregation of a Markov chain can be considered as a special case thereof, e.g. naturally arising when considering a stationary deterministic policy on an aggregated MDP.

Formally, there are various ways how to express MDP aggregation. First, an aggregation of an MDP $M = (\mathcal{S}, \mathcal{A}, r, p, S_1)$ can be defined by a partition $\widehat{\mathcal{S}} = \{\hat{s}_1, \dots, \hat{s}_k\}$ of the state space $\mathcal{S}$, where $\widehat{\mathcal{S}}$ constitutes the state space of the aggregated MDP. Defining rewards and transition probabilities of the aggregated MDP in a suitable way (cf. details below) one obtains an aggregated MDP $\widehat{M} = (\widehat{\mathcal{S}}, \mathcal{A}, \hat{p}, \hat{r}, \hat{S}_1)$ where $\hat{S}_1$ contains S_1. Generally, one can define a surjective function $\varphi : \mathcal{S} \to \widehat{\mathcal{S}}$ with $\varphi(s) = \hat{s}$ iff $s \in \hat{s}$. This also provides an alternative view on state aggregation: Given two MDPs M, $\widehat{M}$ the latter can be defined to be an aggregation of the former if there is a surjective function $\varphi : \mathcal{S} \to \widehat{\mathcal{S}}$ such that rewards and transition probabilities in M, $\widehat{M}$ are compatible (cf. Definitions 27 and 29 below). In the following, we usually consider an aggregation to be such a surjective function φ mapping states to aggregated states.

6.1. Exact Aggregations

The simplest case when aggregation can be applied is when there are states having the same transition probability distributions and the same rewards under all actions. In this case it is obviously not necessary to distinguish between these states and they can be aggregated. That is, the aggregation function φ will map them to the same aggregated state. The corresponding aggregated MDP will have rewards and transition probabilities satisfying for all states s and all actions a,

$$\begin{aligned}\hat{r}\big(\varphi(s),a\big) &= r(s,a),\\ \hat{p}\big(\varphi(s')\,|\,\varphi(s),a\big) &= \sum_{s'':\varphi(s'')=\varphi(s')} p(s''\,|\,s,a).\end{aligned}$$

These two conditions can be taken as definition of an aggregation, which can be applied to more general situations. That is, in general it is not necessary that states which are aggregated have the same transition probability distributions.

Definition 27. *Given two MDPs $M=(\mathcal{S},\mathcal{A},r,p,S_1)$, $\widehat{M}=(\widehat{\mathcal{S}},\mathcal{A},\hat{r},\hat{p},\hat{S}_1)$, we call $\widehat{M}$ an aggregation of M if there is a surjective function $\varphi:\mathcal{S}\to\widehat{\mathcal{S}}$ such that $\hat{S}_1=\varphi(S_1)$ and for all s in $\mathcal{S}$, all $\hat{s}$ in $\widehat{\mathcal{S}}$, and all actions a,*

$$\begin{aligned}\hat{r}\big(\varphi(s),a\big) &= r(s,a),\\ \hat{p}\big(\hat{s}\,|\,\varphi(s),a\big) &= \sum_{s':\varphi(s')=\hat{s}} p(s'\,|\,s,a).\end{aligned}$$

Example 28. *Consider a Markov chain with state space $\mathcal{S}=\{1,2,3,4,5\}$ and transition matrix*

$$\mathbf{P}=\begin{pmatrix} 0 & 1/2 & 1/4 & 1/4 & 0\\ 1/2 & 0 & 0 & 1/2 & 0\\ 1 & 0 & 0 & 0 & 0\\ 1/2 & 1/2 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\end{pmatrix}.$$

The state space can be aggregated to $\hat{1}=\{1,2\}$, $\hat{2}=\{3,4,5\}$, and the aggregated Markov chain has transition matrix

$$\widehat{\mathbf{P}}=\begin{pmatrix} 1/2 & 1/2\\ 1 & 0\end{pmatrix}.$$

While this aggregation obviously comes with a loss of information, for an MRP in which the mean rewards coincide in states 1,2 as well as in states 3,4,5, the average reward could still be computed from the aggregated MRP.

6.2. Policies and their Extensions

Since the state space in the aggregated MDP $\widehat{M}$ is smaller than in the original MDP M there are fewer policies and not each policy in M has an equivalent in $\widehat{M}$. However, any policy $\hat{\pi} : \widehat{\mathcal{S}} \to \mathcal{A}$ defined on the aggregated MDP $\widehat{M}$ can be extended to a policy $\hat{\pi} : \mathcal{S} \to \mathcal{A}$ on the original MDP M by assigning to each state s in $\mathcal{S}$ the action $\hat{\pi}(\varphi(s))$. Moreover, it can be shown that an optimal policy $\hat{\pi}^*$ on $\widehat{M}$ is also optimal in M, cf. Theorem 4 of Ortner (2007).

6.3. Approximate Aggregations

If one allows some error, MDPs sometimes can be approximated by smaller aggregated MDPs. In the simplest case, one would try to aggregate states with similar rewards and transition probability distributions. Generally, the error bounds that have been considered in Section 4 can also be employed here so that there is a close link between MDP approximation and aggregation.

The following definition of approximate aggregation originally introduced by Ortner et al. (2014a) is a generalization of the notion of $(\varepsilon_r, \varepsilon_p)$-approximation in Definition 15 to the case where the two considered MDPs need not share a common state space.

Definition 29. *An MDP $\widehat{M} = (\widehat{\mathcal{S}}, \mathcal{A}, \hat{r}, \hat{p}, \hat{S}_1)$ is an $(\varepsilon_r, \varepsilon_p)$-aggregation of another MDP $M = (\mathcal{S}, \mathcal{A}, r, p, S_1)$ if there is a surjective function $\varphi : \mathcal{S} \to \widehat{\mathcal{S}}$ with $\varphi(S_1) = \hat{S}_1$ such that for all s in $\mathcal{S}$ and all actions a,*

$$\begin{aligned} \left|\hat{r}\big(\varphi(s), a\big) - r(s,a)\right| &\leq \varepsilon_r, &(29)\\ \sum_{\hat{s}' \in \widehat{\mathcal{S}}} \left| \hat{p}\big(\hat{s}' \mid \varphi(s), a\big) - \sum_{s' : \varphi(s') = \hat{s}'} p(s'|s,a) \right| &\leq \varepsilon_p. &(30)\end{aligned}$$

Although the aggregated MDP in general has different state space than the original MDP M, it is still possible to bound the error when using an $(\varepsilon_r, \varepsilon_p)$-aggregation $\widehat{M}$ instead of M by perturbation bounds for the Markov chains induced by a policy. This is possible by defining for a given $(\varepsilon_r, \varepsilon_p)$-aggregation $\widehat{M}$ of M a new MDP $\bar{M}$, which is an $(\varepsilon_r, \varepsilon_p)$-approximation of M and which has for each policy $\hat{\pi} : \widehat{\mathcal{S}} \to \mathcal{A}$ the same average reward as $\widehat{M}$.

Lemma 30. *Let $\widehat{M} = (\widehat{\mathcal{S}}, \mathcal{A}, \hat{r}, \hat{p})$ be an $(\varepsilon_r, \varepsilon_p)$-aggregation of an MDP $M = (\mathcal{S}, \mathcal{A}, r, p)$. Then there is an MDP $\bar{M} = (\mathcal{S}, \mathcal{A}, \bar{r}, \bar{p})$ that is an $(\varepsilon_r, \varepsilon_p)$-approximation of M such that for all policies $\hat{\pi} : \widehat{\mathcal{S}} \to \mathcal{A}$ it holds that*[12]

$$\rho(\widehat{M}, \hat{\pi}, \hat{s}) = \rho(\bar{M}, \hat{\pi}, s) \tag{31}$$

for all states s in $\mathcal{S}$ and all $\hat{s}$ in $\widehat{\mathcal{S}}$ with $\varphi(s) = \hat{s}$.

Proof. For each state s in $\mathcal{S}$ and each action a we set $\bar{r}(s,a) := \hat{r}(\varphi(s), a)$ and

$$\bar{p}(s'|s,a) := \frac{p(s'|s,a) \cdot \hat{p}\big(\varphi(s') \mid \varphi(s), a\big)}{\sum_{s'':\varphi(s'')=\varphi(s')} p(s''|s,a)} .$$

Note that $\bar{p}(\cdot|s,a)$ is indeed a probability distribution over $\mathcal{S}$. In particular,

$$\begin{aligned}
\sum_{s' \in \mathcal{S}} \bar{p}(s'|s,a) &= \sum_{s' \in \mathcal{S}} \frac{p(s'|s,a)}{\sum_{s'':\varphi(s'')=\varphi(s')} p(s''|s,a)} \cdot \hat{p}\big(\varphi(s') \mid \varphi(s), a\big) \\
&= \sum_{\hat{s}' \in \widehat{\mathcal{S}}} \sum_{s':\varphi(s')=\hat{s}'} \frac{p(s'|s,a)}{\sum_{s'':\varphi(s'')=\hat{s}'} p(s''|s,a)} \cdot \hat{p}\big(\hat{s}' \mid \varphi(s), a\big) \\
&= \sum_{\hat{s}' \in \widehat{\mathcal{S}}} \hat{p}\big(\hat{s}' \mid \varphi(s), a\big) \; = \; 1.
\end{aligned}$$

Now by definition, the rewards $\bar{r}(s,a) = \hat{r}\big(\varphi(s), a\big)$ and aggregated transition probabilities

$$\begin{aligned}
\sum_{s':\varphi(s')=\hat{s}'} \bar{p}(s'|s,a) &= \sum_{s':\varphi(s')=\hat{s}'} \frac{p(s'|s,a)}{\sum_{s'':\varphi(s'')=\hat{s}'} p(s''|s,a)} \cdot \hat{p}\big(\hat{s}' \mid \varphi(s), a\big) \\
&= \hat{p}\big(\hat{s}' \mid \varphi(s), a\big)
\end{aligned}$$

in $\bar{M}$ have the same values for all states s that are mapped to the same aggregated state by φ. Therefore, $\widehat{M}$ is an aggregation of $\bar{M}$ according to Definition 27, and it follows that (31) holds for $s, \hat{s}$ with $\varphi(s) = \hat{s}$.

Further by (29) and (30) we have

$$|r(s,a) - \bar{r}(s,a)| \; = \; \big|r(s,a) - \hat{r}\big(\varphi(s), a\big)\big| \; \leq \; \varepsilon_r$$

[12] In $\bar{M}$, actually the policy $\hat{\pi}$ extended to $\mathcal{S}$ as defined in Section 6.2 is considered.

as well as

$$\begin{aligned}\sum_{s'\in\mathcal{S}} \big|p(s'|s,a)-\bar{p}(s'|s,a)\big| &= \sum_{s'\in\mathcal{S}} p(s'|s,a)\cdot\left|1-\frac{\hat{p}(\varphi(s')|\varphi(s),a)}{\sum_{s'':\varphi(s'')=\varphi(s')} p(s''|s,a)}\right| \\ &= \sum_{\hat{s}'\in\widehat{\mathcal{S}}}\ \sum_{s':\varphi(s')=\hat{s}'} p(s'|s,a)\cdot\left|\frac{\sum_{s'':\varphi(s'')=\hat{s}'} p(s''|s,a)-\hat{p}(\hat{s}'|\varphi(s),a)}{\sum_{s'':\varphi(s'')=\hat{s}'} p(s''|s,a)}\right| \\ &= \sum_{\hat{s}'\in\widehat{\mathcal{S}}}\Big|\sum_{s'':\varphi(s'')=\hat{s}'} p(s''|s,a)-\hat{p}(\hat{s}'|\varphi(s),a)\Big| \ \le\ \varepsilon_p,\end{aligned}$$

which shows that $\bar{M}$ is an $(\varepsilon_r,\varepsilon_p)$-approximation of M. □

As noted in the proof, the MDP $\bar{M}$ in Lemma 30 is defined such that $\widehat{M}$ is an (exact) aggregation of $\bar{M}$. Accordingly, an optimal policy $\hat{\pi}^*$ of $\widehat{M}$ will also be optimal in $\bar{M}$. Since $\bar{M}$ is an $(\varepsilon_r,\varepsilon_p)$-approximation of M we can apply the results of Section 4 to obtain respective guarantees for the error caused by approximate aggregation. In particular, the following result is a generalization of Corollary 19.

Corollary 31. *Let $\widehat{M}$ be a communicating $(\varepsilon_r,\varepsilon_p)$-approximation of a communicating MDP M. Let π^* and $\hat{\pi}^*$ be optimal policies satisfying the Bellman equation* (12) *in M and $\widehat{M}$, respectively. Further assume that $\rho(\bar{M},\pi^*)$ and $\rho(M,\hat{\pi}^*)$ are independent of the initial state, where $\bar{M}$ is the MDP of Lemma 30. Then writing $\rho^*:=\rho(M,\pi^*)$, $\hat{\rho}^*:=\rho(\widehat{M},\hat{\pi}^*)$ for the optimal average reward in M and $\widehat{M}$, it holds that*

$$\begin{aligned}|\rho^*-\hat{\rho}^*| &\le \varepsilon_r+\tfrac{1}{2}D\,\varepsilon_p, \text{ and} \\ \rho(M,\hat{\pi}^*) &\ge \rho^*-2\varepsilon_r-D\,\varepsilon_p.\end{aligned}$$

Example 32. *The most extreme example of an $(\varepsilon_r,\varepsilon_p)$-aggregation is when all states of a Markov chain or an MDP are aggregated to a single state with transition matrix $\mathbf{P}=(1)$ under each action. While this can be done for all irreducible Markov chains and communicating MDPs for a sufficiently large ε_p it will in general make not much sense. An exception would be an MRP where all states have the same mean reward, so that differentiating between single states is not necessary if one is only interested in the average reward.*

6.4. Applications

Aggregation usually combines the topic of approximation with a simplification of the underlying MDP structure. When the MDP is explicitly given, this way one trades precision with computation time. In the context of reinforcement learning aggregation techniques in order to speed up computation make even more sense, as the MDP parameters are only estimated up to a certain precision, so that aggregation does not necessarily mean an additional loss of accuracy. An example for an online reinforcement learning algorithm which employs aggregation of states that have intersecting confidence intervals has been provided by Ortner (2013), which also notes the downside that finding suitable aggregations is hard in itself (cf. also Even-Dar and Mansour, 2003 for details).

7. Beyond Aggregation: $(\varepsilon_r, \varepsilon_p)$-structured MDPs and Restless Bandits

State aggregation is a natural concept that allows to simplify MDPs. With respect to reinforcement learning, the knowledge of the respective structure allows faster learning as an MDP with smaller state space can be learned using fewer samples. While the concept of state aggregation is useful, an MDP may exhibit more complex symmetries which allow to use a sample more than once although aggregation is not possible. An example for such a setting is the *restless bandit* problem (Whittle, 1988), which will be presented in this section. Beside introducing the more refined concept of $(\varepsilon_r, \varepsilon_p)$-structured MDPs, restless bandits also allow to apply other concepts presented in the previous sections.

7.1. Motivation

As a motivating example consider that you would like to follow two or even more soccer games running in parallel and broadcasted on different TV channels. If you only have a single TV set, you have to switch channels in order to get an update on each different game. Usually, one has certain preferences concerning different scenes in a soccer game, so that one is for example more interested in observing goals than a longer break caused by an injury of a player. Further, a soccer game follows some inherent logic so that the observer can try to predict whether something interesting is going to happen in the next few moments. Observing a goal is thus more likely after a team has been awarded a penalty than in the half-time break. In the model which we will define in more detail below we will assume that underlying each soccer game is a hidden Markov chain according to which the game evolves. The aim of the observer is to find a switching strategy that maximizes her rewards with respect to the observed scenes (for example the number of watched goals).

Obviously, the setting of this example appears in other natural applications where one has to switch between different options that evolve independently of the made choices. Another well-known example that fits this description is *cognitive radio* (Avner and Mannor, 2014) where one aims to keep track of the transmission channel offering the best quality (cf. below for details).

7.2. The Restless Bandit Setting

Choosing in discrete time steps $t = 1, 2, \dots$ from a fixed set of options is usually considered as an instance of a *multi-armed bandit* problem (Lattimore and Szepesvári, 2020). Accordingly, the options are called *arms*. In the standard setting the rewards of each arm are iid samples from a fixed distribution not known to the learner. In our case the rewards of each arm come from an underlying MRP with irreducible and aperiodic Markov chain that evolves independently of which arm is chosen.

More precisely, we consider a fixed set of K arms numbered from 1 to K. Underlying each arm j there is an MRP with state space $\mathcal{S}_j$, transition matrix $\mathbf{P}_j$ and mean rewards $r_j(s)$ in each state s in $\mathcal{S}_j$. For each MRP there is a designated initial state at step $t = 1$. When an arm j is selected at time t, one receives a random reward drawn from the reward distribution of the state s of the chosen arm at time t. Then all MRPs evolve in parallel according to their current states and their transition matrices.

What is important is that (as in our introductory soccer game example) the only observations one can rely on is the feedback from selecting arms. That is, in particular current states of arms not chosen are unknown so that one only obtains an update for an arm by picking it.

7.3. Examples

Before considering a suitable MDP representation of the setting we give two further examples of (Ortner et al., 2014b) for illustration.

7.3.1. Cognitive Radio

Returning to the application of cognitive radio, consider K radio channels each of which can be either busy or available. The goal is to always pick a channel which is available so that we assign deterministic rewards of either 0 when the chosen channel is busy, or 1 if it is available. The underlying MRPs accordingly have two states 0 and 1.

When the probability of staying in either state is high then it is optimal to keep playing an arm giving a reward of 1 also in the subsequent time step and to switch otherwise. On the other hand, if the probability of a transition to the other state is higher than staying in the current one, then it is better to switch after observing a reward of 1 and to stay otherwise.

7.3.2. Exploration

Consider another restless bandit setting with two arms. Arm 1 with small probability changes between two states giving reward 0 and 1, respectively, as in the previous example. Arm 2 gives a constant and deterministic reward of $\frac{1}{2}$. The optimal policy will keep playing arm 1 as long as it gives reward 1 and change to arm 2 when the observed reward of arm 1 is 0. Every now and then the optimal policy will have to check out arm 1 so that it does not miss when it changes back to the state giving reward 1. This shows that the *exploration-exploitation payoff* (Fruit, 2019) is important in the restless bandit setting even if the scenario is fully known. In particular, the optimal policy will accept a temporary loss in order to collect information that will lead to higher reward in later steps, which is the type of behavior a learning agent has to adopt in many reinforcement learning problems.

7.4. The Optimal Policy

In general, neither the form of the optimal policy nor its computation are easy (Papadimitriou and Tsitsiklis, 1999). Concerning the optimal policy one can show that in general there is no index-based policy that is optimal. An *index-based policy* computes a single quantity for each arm and decides based on this *index* which arm to choose. The computation of the index takes into account only observations concerning the arm itself and not any of the other arms.

Theorem 33 (Theorem 4 of Ortner et al., 2014b)**.** *For each index-based policy π there is a restless bandit problem in which π behaves suboptimally.*

Thus, while index policies are still attractive in order to at least approximate an optimal policy (Whittle, 1988; Guha et al., 2010; Grünewälder and Khaleghi, 2019), the form of the optimal policy is in general more complex. We will return to that point after introducing an MDP representation of the restless bandit setting.

7.5. MDP Representation

The restless bandit setting can be formalized as an MDP as follows (Tekin and Liu, 2011; Ortner et al., 2014b). The state space consists of vectors of the form

$$(s_j, n_j)_{j=1}^K := (s_1, n_1, s_2, n_2, \dots, s_K, n_K),$$

where each s_j is an element of the state space $\mathcal{S}_j$ of arm j, and the n_j's are pairwise distinct natural numbers. The interpretation of this vector is as follows. For each arm j the last observed state s_j as well as the number of time steps are noted that have passed since this observation took place. As there always has to be an arm that has been chosen at the previous step, one of the n_j's has to be 1, while we also introduce the option of n_j's being 0 in case an arm j has not been chosen at all. In the latter case, as respective state s_j the initial state of arm j will be used.

The actions are obviously the arms $1, 2, \dots, K$ and when choosing arm i in state $(s_j, n_j)_{j=1}^K$ the successor state in the MDP can only change the observed state of arm i. For this arm i the successor state has $n_i = 1$, while the states of the arms $j \neq i$ not chosen remain the same and the respective counts n_j increase by 1. The transition probabilities correspond to the n_i-step transition probabilities of arm i, that is, the respective entries in the power matrix $\mathbf{P}_i^{n_i}$. More precisely, the transition probability from state $(s_j, n_j)_{j=1}^K$ to state $(s'_j, n'_j)_{j=1}^K$ under action i is given by

$$p\big((s'_j, n'_j)_{j=1}^K \,|\, (s_j, n_j)_{j=1}^K, i\big) = \begin{cases} p_i^{n_i}(s'_i|s_i), & \text{if } n'_i = 1 \text{ and } \forall j \neq i: \\ & n'_j = n_j + 1, s_j = s'_j, \\ 0 & \text{else.} \end{cases} \tag{32}$$

Finally, the reward for choosing arm i in state $(s_j, n_j)_{j=1}^K$ is taken from the reward distribution of the observed state of the chosen arm i, so that the respective mean reward is given by $\sum_{s \in \mathcal{S}_i} p_i^{n_i}(s|s_i) \cdot r_i(s)$.

By construction the number of states of the introduced MDP representation is countably infinite. However, considering that after a finite number of T steps each n_j cannot be larger than T it is sufficient to consider a finite MDP where the state space is restricted to states $(s_j, n_j)_{j=1}^K$ with $n_j \leq T$.

As the MDP representation contains all information at the disposal of the learner, the optimal policy of a restless bandit problem corresponds to the optimal policy in the MDP representation. That way we have not only found a neat way to represent the setting but can also formally pin down the optimal policy. Note that an optimal policy will operate only on a finite part of the infinite MDP representation, except in case there is an arm that is never played. Such arms can however simply be deleted from the representation so that considering a finite representation is always sufficient. In principle, the computation of an optimal policy in this finite MDP representation can be done using standard algorithms as those briefly mentioned in Section 3.4.

7.6. $(\varepsilon_r, \varepsilon_p)$*-structured MDPs*

The MDP representation of the restless bandit setting has some interesting properties that facilitiate learning. First, we have already seen that the transition probability distributions are sparse, that is, most transition probabilities are known to be 0 so that there is no need to sample them. Furthermore, also the remaining positive transition probabilities have a special structure. Generally, the transition probability for choosing arm i in state $(s_j, n_j)_{j=1}^K$ only depends on s_i and n_i and not on the other entries of the state vector. Moreover, these transition probabilities can actually be observed in different states. That is, by (32) the transition probability distributions for choosing arm i in two states $(s_j, n_j)_{j=1}^K$, $(s'_j, n'_j)_{j=1}^K$ are basically the same when $s_i = s'_i$ and $n_i = n'_i$, only the respective successor states are different. Finally, for sufficiently large n_j the Markov chain underlying arm j can be considered to be close to its stationary distribution. Accordingly, states that differ only with respect to arms j with $n_j \geq T_{\text{mix}}^j(\varepsilon)$ can be aggregated, where $T_{\text{mix}}^j(\varepsilon)$ denotes the ε-mixing time of the Markov chain underlying arm j. The introduced error depends on ε and can be bounded by Corollary 31.

While the latter aggregation of states can be formulated within the aggregation formalization introduced in Section 6, the structural properties concerning the individual transition probabilities cannot. The following notion of $(\varepsilon_r, \varepsilon_p)$-*structured MDP* however can grasp also similarities on the level of state-action pairs instead of just states.

Definition 34 (Ortner et al., 2014b)**.** *Let* $M = (\mathcal{S}, \mathcal{A}, r, p, S_1)$ *be an MDP and* $c : \mathcal{S} \times \mathcal{A} \to \mathcal{C}$ *be a coloring function that assigns each state-action* *pair a color from a given set of colors $\mathcal{C}$. We call the pair (M,c) an $(\varepsilon_r,\varepsilon_p)$-structured MDP if for each two state-action pairs (s,a), (s',a') with $c(s,a) = c(s',a')$ there is a bijective translation function $\varphi_{s,a,s',a'} : \mathcal{S} \to \mathcal{S}$ such that*

$$\begin{aligned} |r(s,a) - r(s',a')| &\leq \varepsilon_r, \\ \sum_{s''} \big|p(s''|s,a) - p\big(\varphi_{s,a,s',a'}(s'') \,|\, s',a'\big)\big| &\leq \varepsilon_p. \end{aligned}$$

While $(\varepsilon_r,\varepsilon_p)$-structured MDPs can represent weaker notions of similarity than aggregation, in general the MDP cannot be reduced to a smaller one. Aggregation of two states s,s' in the sense of Definition 27 would only be possible when $c(s,a) = c(s',a)$ for all actions a and the associated translation function $\varphi_{s,a,s',a}$ is the identity. However, even if this is not the case, with respect to reinforcement learning the additional structural information allows the learner to speed up the learning process.

For the MDP representation of the restless bandit we can give the following coloring function. We choose $c((s_j,n_j)_{j=1}^K, i) = c((s'_j,n'_j)_{j=1}^K, i')$ whenever $i = i'$, $s_i = s'_i$, and either $n_i = n'_i$ or $n_i, n'_i \geq T^i_{\mathrm{mix}}(\varepsilon)$. The corresponding translation function maps the state

$$(s_1, n_1+1, \ldots, s_{i-1}, n_{i-1}+1, s, 1, s_{i+1}, n_{i+1}+1, \ldots, s_K, n_K+1)$$

to state

$$(s'_1, n'_1+1, \ldots, s'_{i-1}, n'_{i-1}+1, s, 1, s'_{i+1}, n'_{i+1}+1, \ldots, s'_K, n'_K+1).$$

Intuitively, the translation function maps the possible successor states of $(s_j,n_j)_{j=1}^K$ to the corresponding successor states of $(s'_j,n'_j)_{j=1}^K$ when picking arm i.

7.7. Notes and Applications

The restless bandit setting combines all the topics we have encountered before. It also introduces the notion of $(\varepsilon_r,\varepsilon_p)$-structured MDPs, which generalizes the concept of aggregation. Taking a look at the single Markov chains C_π induced by a policy $\pi : \mathcal{S} \to \mathcal{A}$, all states s,s' for which $c(s,\pi(s)) = c(s',\pi(s'))$ could be aggregated in C_π. Thus, the coloring function stores all respective structural information of all the $|\mathcal{A}|^{|\mathcal{S}|}$ Markov chains C_π.

A reinforcement learning algorithm for restless bandits and more generally for $(\varepsilon_r,\varepsilon_p)$-structured MDPs can be found in (Ortner et al., 2014b). The given algorithm is able to exploit the underlying structure by making use of the results presented in this section.

References

Auer, P., Ortner, R., 2007. Logarithmic online regret bounds for reinforcement learning, in: Adv. Neural Inf. Process. Syst. 19, pp. 49–56.

Avner, O., Mannor, S., 2014. Concurrent bandits and cognitive radio networks, in: Machine Learning and Knowledge Discovery in Databases – European Conference, ECML PKDD 2014. Proceedings, Part I, Springer. pp. 66–81.

Azar, M.G., Lazaric, A., Brunskill, E., 2013. Regret bounds for reinforcement learning with policy advice, in: Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2013, Proceedings, Part I, Springer. pp. 97–112.

Bartlett, P.L., Tewari, A., 2009. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs, in: Proc. 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009, pp. 25–42.

Boone, V., 2024. Optimal Regrets in Markov decision processes. PhD thesis. Université Grenoble Alpes. URL: `https://hal.science/tel-04874467v1`.

Burnetas, A.N., Katehakis, M.N., 1997. On confidence intervals from simulation of finite Markov chains. Math. Oper. Res. 46, 241–250.

Caswell, H., 2013. Sensitivity analysis of discrete Markov chains via matrix calculus. Linear Algebra Appl. 438, 1727–1745.

Caswell, H., 2019. Sensitivity Analysis: Matrix Methods in Demography and Ecology. Springer, Berlin.

Cho, G.E., Meyer, C.D., 2000. Markov chain sensitivity measured by mean first passage times. Linear Algebra Appl. 316, 21–28.

Cho, G.E., Meyer, C.D., 2001. Comparison of perturbation bounds for the stationary distribution of a Markov chain. Linear Algebra Appl. 335, 137–150.

Even-Dar, E., Mansour, Y., 2003. Approximate equivalence of Markov decision processes, in: Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, pp. 581–594.

Filippi, S., Cappé, O., Garivier, A., 2010. Optimism in reinforcement learning and Kullback-Leibler divergence, in: 48th Annual Allerton Conference on Communication, Control, and Computing, IEEE. pp. 115–122.

Fruit, R., 2019. Exploration-exploitation dilemma in Reinforcement Learning under various form of prior knowledge. PhD thesis. Université de Lille 1, Sciences et Technologies. URL: `https://theses.hal.science/tel-02388395`.

Geiger, B.C., Petrov, T., Kubin, G., Koeppl, H., 2015. Optimal Kullback-Leibler aggregation via information bottleneck. IEEE Trans. Autom. Control. 60, 1010–1022.

Givan, R., Dean, T., Greig, M., 2003. Equivalence notions and model minimization in Markov decision processes. Artif. Intell. 147, 163–223.

Givan, R., Leach, S.M., Dean, T., 2000. Bounded-parameter Markov decision processes. Artif. Intell. 122, 71–109.

Grünewälder, S., Khaleghi, A., 2019. Approximations of the restless bandit problem. J. Mach. Learn. Res. 20, 14:1–14:37.

Guha, S., Munagala, K., Shi, P., 2010. Approximation algorithms for restless bandit problems. J. ACM 58, 3:1–3:50.

Hordijk, A., Yushkevich, A.A., 2002. Blackwell optimality, in: Handbook of Markov decision processes. Kluwer Acad. Publ., Boston, MA. volume 40 of *Internat. Ser. Oper. Res. Management Sci.*, pp. 231–267.

Hunter, J.J., 2005. Stationary distributions and mean first passage times of perturbed Markov chains. Linear Algebra Appl. 410, 217–243.

Hunter, J.J., 2006. Mixing times with applications to perturbed Markov chains. Linear Algebra Appl. 417, 108–123.

Hunter, J.J., 2014. The role of Kemeny's constant in properties of Markov chains. Comm. Statist. Theory Methods 43, 1309–1321.

Jaksch, T., Ortner, R., Auer, P., 2010. Near-optimal regret bounds for reinforcement learning. J. Mach. Learn. Res. 11, 1563–1600.

Kallenberg, L., 2016. Markov decision processes. URL: `https://www.math.leidenuniv.nl/~kallenberg/Lecture-notes-MDP.pdf`. unpublished lecture notes.

Kallenberg, L.C.M., 2002. Classification problems in MDPs, in: Hou, Z., Filar, J.A., Chen, A. (Eds.), Markov Processes and Controlled Markov Chains. Springer US, Boston, MA, pp. 151–165.

Kemeny, J., Snell, J., 1960. Finite Markov Chains. Van Nostrand.

Lattimore, T., Szepesvári, C., 2020. Bandit Algorithms. Cambridge University Press.

Levin, D.A., Peres, Y., Wilmer, E.L., 2006. Markov chains and mixing times. American Mathematical Society.

Lewis, M.E., Puterman, M.L., 2002. Bias optimality, in: Handbook of Markov decision processes. Kluwer Acad. Publ., Boston, MA. volume 40 of *Internat. Ser. Oper. Res. Management Sci.*, pp. 89–111.

Norris, J.R., 1998. Markov chains. Cambridge series in statistical and probabilistic mathematics, Cambridge University Press.

Ortner, R., 2007. Pseudometrics for state aggregation in average reward Markov decision processes, in: Algorithmic Learning Theory, 18th International Conference, ALT 2007, Proceedings, pp. 373–387.

Ortner, R., 2013. Adaptive aggregation for reinforcement learning in average reward Markov decision processes. Ann. Oper. Res. 208, 321–336.

Ortner, R., 2020. Regret bounds for reinforcement learning via Markov chain concentration. J. Artif. Intell. Res. 67, 115–128.

Ortner, R., 2024. A note on the bias and Kemeny's constant in Markov reward processes with an application to Markov chain perturbation `arXiv:arXiv/2408.04454`.

Ortner, R., Maillard, O., Ryabko, D., 2014a. Selecting near-optimal approximate state representations in reinforcement learning, in: Algorithmic Learning Theory – 25th International Conference, ALT 2014, pp. 140–154.

Ortner, R., Ryabko, D., Auer, P., Munos, R., 2014b. Regret bounds for restless Markov bandits. Theor. Comput. Sci. 558, 62–76.

Papadimitriou, C.H., Tsitsiklis, J.N., 1999. The complexity of optimal queuing network control. Math. Oper. Res. 24, 293–305.

Paulin, D., 2015. Concentration inequalities for Markov chains by Marton couplings and spectral methods. Electron. J. Probab. 20, 1–32.

Puterman, M.L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, Inc., New York, NY, USA.

Tekin, C., Liu, M., 2011. Adaptive learning of uncontrolled restless bandits with logarithmic regret, in: 49th Annual Allerton Conference on Communication, Control, and Computing, IEEE. pp. 983–990.

Tewari, A., Bartlett, P.L., 2007. Bounded parameter Markov decision processes with average reward criterion, in: Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007, pp. 263–277.

Van Roy, B., 2006. Performance loss bounds for approximate value iteration with state aggregation. Math. Oper. Res. 31, 234–244.

Whittle, P., 1988. Restless bandits: Activity allocation in a changing world. J. Appl. Probab. 25, pp. 287–298.

Woess, W., 2009. Denumerable Markov chains. Generating functions, boundary theory, random walks on trees. EMS Textbooks in Mathematics, European Mathematical Society (EMS), Zürich.