Supplementary Material for "Finite-Memory Near-Optimal Learning for Markov Decision Processes with Long-Run Average Reward"

A EFFECTIVE HOEFFDING BOUNDS FOR UNKNOWN MDPS

We presently develop the main ideas required to prove Theorem 1. To do so, we follow the work of Tracol (2009).

Before we begin, we need to introduce some additional notation and work our way from bounds for unknown chains up to unknown MDPs.

A.1 REGULAR MARKOV CHAINS

Let $\mathcal{C} = (Q, P, R)$ be a Markov chain. We say $\mathcal{C}$ is *regular* if $|Q|$ is finite and it is both aperiodic and irreducible. It will be useful to recall the definitions of the latter two terms. A state q has *period* $p(q) = \gcd\{n > 0 : \Pr_{\mathcal{C}}^{q}[q_n = q]\}$. The MC $\mathcal{C}$ is said to be *aperiodic* if $p(q) = 1$ for all $q \in Q$; it is *irreducible* if $\mathcal{G}_{\mathcal{C}} = (Q, E)$, where $E \stackrel{\text{def}}{=} \{(q, q') \in Q \times Q : P(q, q') > 0\}$, is a strongly connected directed graph. In words, the chain is irreducible if the probability of reaching q' from q is nonzero for all $q, q' \in Q$.

It is well known that, finite irreducible, and thus regular, Markov chains have a unique *stationary distribution* $\pi \in \mathrm{Dist}(Q)$ (Norris, 1998).

Doeblin's condition. In the sequel we will state Hoeffding-like inequalities in terms of an *ergodicity coefficient* and its corresponding *mixing time*. To be precise, we recall a classical sufficient condition for an MC to be *uniformly ergodic* (Meyn and Tweedie, 1993).

Definition 8. *A Markov chain $\mathcal{C}$ satisfies Doeblin's condition if there exist $\lambda \in \mathbb{R}, \lambda > 0$, a probability measure φ over (subsets of) Q, and an integer $t \in \mathbb{N}$, such that*

$$\Pr_{\mathcal{C}}^{q_0}[q_t \in T] \geq \lambda\varphi(T)$$

for all $T \subseteq Q$ and all $q_0 \in Q$.

Henceforth, we refer to λ as the ergodicity coefficient and to t as the mixing time.

It is classical to show that, for regular unknown MCs, Doeblin's condition always holds (Meyn and Tweedie, 1993). We give a proof of this claim in order to convince the reader that we can compute an ergodicity coefficient and a mixing time even if the MC is unknown.

Lemma 2. *Let $\mathcal{C} = (Q, P, R)$ be an unknown regular MC and $\mathbf{p}_{\min}$ be a lower bound for all nonzero transition probabilities. One can compute values $\lambda \in \mathbb{Q}, \lambda > 0$ and $t \in \mathbb{N}$ (dependent only on $\mathbf{p}_{\min}$ and $|Q|$) such that $\Pr_{\mathcal{C}}^{q_0}[q_t \in T] \geq \lambda\varphi(T)$ for all $T \subseteq Q$ and all $q_0 \in Q$.*

Proof. First, since $\mathcal{C}$ is regular, we know it has a unique stationary distribution π. We now observe that, because of aperiodicity and irreducibility, we know there exists some t such that

$$\Pr_{\mathcal{C}}^{q_0}[q_t = s] > 0$$

for any $q_0, s \in Q$. To give one such t we need to recall some definitions.

Given a finite set $N = \{a_1, \ldots, a_\ell\} \subseteq \mathbb{N}_{>0}$ of positive integers such that $\gcd(N) = 1$ we write $g(N)$ to denote the Frobenius number: The maximal integer that cannot be obtained as a *conical combination* of the a_i, that is as a sum

$$k_1 a_1 + \cdots + k_\ell a_\ell$$

where $k_1, \ldots, k_\ell \in \mathbb{N}$. (Since the set of numbers that is not obtainable as such a conical combination is bounded by Schur's theorem, $g(N)$ indeed exists.)

Let us set

$$t = \max_{N \subseteq \{1,2,\ldots,|Q|\}} g(N) + 1.$$

It should be clear that, since $\mathcal{C}$ is regular, we have that

$$\forall q_0, s \in Q : \Pr_{\mathcal{C}}^{q_0}[q_t = s] \geq \mathbf{p}_{\min}^t > 0. \tag{1}$$

Indeed, because of aperiodicity and the definition of the Frobenius number, a run prefix of length t' and nonzero probability exists from any state to any other state in the chain for any $t' \geq t$. To conclude, we set $\lambda = \mathbf{p}_{\min}^t$ and $\varphi = \pi$ and observe that Equation (1) implies that $\Pr_{\mathcal{C}}^{q_0}[q_t = s] \cdot \lambda^{-1} \geq 1 \geq \varphi(s)$. The desired result follows. □

A.2 UNICHAIN MARKOV CHAINS

Let $\mathcal{C} = (Q, P, R)$ be a Markov chain. We say $\mathcal{C}$ is *unichain* if it contains a unique maximal (w.r.t. state-set inclusion) irreducible sub-MC. In other words, the MC consists of a single recurrent class of states plus a possibly empty set of transient states. Note that unichain Markov chains also have a unique stationary distribution since almost all runs reach the unique maximal irreducible sub-MC (Baier and Katoen, 2008, Theorem 10.120).

From Lemma 2 and Proposition 2 in the work of Tracol (2009) we get the following.

Proposition 2. *Let $\mathcal{C} = (Q, P, R)$ be an unknown finite unichain MC and $\mathbf{p}_{\min}$ be a lower bound for all nonzero transition probabilities. For all $\varepsilon \in (0, 1)$ one can compute $K_0 \in \mathbb{N}$ and $\alpha, \beta \in \mathbb{Q}, \alpha, \beta > 0$ (dependent only on $\mathbf{p}_{\min}$, $|Q|$, and ε) such that*

$$\Pr_{\mathcal{C}}^{q_0}[\rho : \mathbf{Avg}_k(\rho) \geq \mathrm{E}_{\mathcal{C}}^{q_0}[\mathbf{Avg}_k] - \varepsilon] \geq 1 - \alpha \cdot \exp(-k \cdot \beta \cdot \varepsilon^2)$$

for all $k \geq K_0$ and all $q_0 \in Q$.

Proof. Tracol originally achieves this by decomposing $\mathcal{C}$ into its transient set of states and the regular sub-MC $\mathcal{C}'$ it contains. He then proceeds to decompose $\mathcal{C}'$ into aperiodic components based on all residue classes modulo the period of states of $\mathcal{C}'$. Finally, he (essentially) uses Lemma 2 to obtain bounds for each such regular chain. Critically, to compute K_0, α, β, he uses only $\mathbf{p}_{\min}$, $|Q|$, the ergodicity coefficient of each regular sub-chain, as well as their mixing times.

Since $\mathcal{C}$ is finite and the period of $\mathcal{C}'$ is bounded by $|Q|$ we can compute K_0, α, β taking all possible such decompositions into account. This allows us the flexibility of not having to rely on more information about P than just $\mathbf{p}_{\min}$ and $|Q|$. □

A.3 FINITE-MEMORY UNICHAIN STRATEGIES FOR MDPS

A strategy σ for an MDP $\mathcal{M} = (Q, q_0, A, P, R)$ is said to be *unichain* if the induced MC $\mathcal{M}^\sigma$ is a unichain MC. In the sequel we will be interested in such strategies which are additionally finite-state encodable. To obtain bounds for MDPs under finite-state unichain strategies we start by specializing our definition of induced MC.

Consider an MDP $\mathcal{M} = (Q, q_0, A, P, R)$ and a finite-state strategy σ implemented by the stochastic Mealy machine $\mathcal{T} = (M, m_0, f_u, f_o)$. The induced MC $\mathcal{M}^\sigma$ can be constructed as the product of $\mathcal{M}$ and $\mathcal{T}$ to obtain a **finite** MC (Q', P', R') as follows.

- $Q' = (Q \times M \times A) \cup (Q \times M)$

- $P'(\langle q', m', a' \rangle | s) = f_o(a'|m, q) \cdot P(q'|q, a')$ for any $s \in \{\langle q, m, a \rangle, \langle q, m \rangle\}$ and $a' \in \alpha(q)$ with $(q, a', q') \in \mathrm{supp}(P)$ and $m' = f_u(m, q, r(q, a', q'))$
- and $R'(s, \langle q', m', a' \rangle) = R(q, a, q')$ for any $s \in \{\langle q, m, a \rangle, \langle q, m \rangle\}$

For convenience, we write $\Pr^{q_0}_{\mathcal{M}^\sigma}[\cdot]$ instead of $\Pr^{\langle q_0, m_0 \rangle}_{\mathcal{M}^\sigma}[\cdot]$.

We now argue that, for finite-state unichain strategies, the the limit of the expected averages and the expectation of the limit of the averages coincide. That is to say, the limit can be "pushed into the expectation operator".

Lemma 3. *Let $\mathcal{M} = (Q, q_0, A, P, R)$ be an unknown MDP and σ be a finite-state unichain strategy for $\mathcal{M}$.*

$$\mathrm{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}] = \liminf_{k \to \infty} \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{Avg}_k] = \limsup_{k \to \infty} \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{Avg}_k].$$

Proof. Our approach consists in applying Lebesgue's dominated convergence theorem, which gives sufficient conditions for the equivalence between the limit of the expectation of functions and the expectation of their limit. Simply stated, we need the (pointwise) limit of the $\mathbf{Avg}_k(\cdot)$ functions to almost-surely exist and a (finite-expectation) bound on $\mathbf{Avg}_k(\rho)$ for all ρ and all k. For the second point: recall that the reward function is bounded, i.e. all rewards are in $[0, 1]$, thus $0 \leq \mathbf{Avg}_k(\rho) \leq 1$ for all ρ and all k. It remains to prove the limit of the average functions almost always exists.

For finite irreducible Markov chains $\mathcal{C}$ the ergodic theorem (Norris, 1998, Theorem 1.10.2) tells us that

$$\Pr^{q_0}_{\mathcal{C}}\left[\rho : \liminf_{k \to \infty} \mathbf{Avg}_k(\rho) = \limsup_{k \to \infty} \mathbf{Avg}_k(\rho)\right] = 1$$

thus the limit almost-surely exists. To conclude, we observe that the above extends to product MCs obtained from finite-state unichain strategies and MDPs with a mean-payoff function (such as $\mathcal{M}^\sigma$) since the function is prefix-independent and almost all runs reach the unique maximal irreducible sub-MC (Baier and Katoen, 2008, Theorem 10.120). □

The final bound. The main tool used in the next section is the following result. It gives us Hoeffding-like bounds on the convergence of the finite averages observed while following finite-state unichain strategies. Additionally, it tells us that the expected mean-payoff value is obtained almost surely.

Lemma 4. *Let $\mathcal{M} = (Q, q_0, A, P, R)$ be an unknown MDP, $\mathbf{p}_{\min}$ be a lower bound for all nonzero transition probabilities, and σ be a finite-state unichain strategy for $\mathcal{M}$.*

1. *For all $\varepsilon \in (0, 1)$ one can compute $M(\varepsilon) \in \mathbb{N}$ (dependent only on $\mathbf{p}_{\min}$, $|Q|$, $|A|$, and the amount of memory used by σ) such that*

$$\Pr^{q_0}_{\mathcal{M}^\sigma}[\rho : \forall k \geq M(\varepsilon),\ \mathbf{Avg}_k(\rho) \geq \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{Avg}_k] - \varepsilon] \geq 1 - \varepsilon.$$

2. *It holds that $\Pr^{q_0}_{\mathcal{M}^\sigma}[\rho : \mathbf{MP}(\rho) \geq \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}]] = 1$.*

Proof.

Item 1 Recall the bound from Proposition 2 and let K_0 be the corresponding integer computed for the finite unichain MC $\mathcal{M}^\sigma$. We observe that for all ε we can compute a $K_1 \geq K_0$ such that

$$1 - \alpha \cdot \exp(-k \cdot \beta \cdot \varepsilon^2) \leq 1 - 2^{-k}$$

for all $k \geq K_1$.

The following inequality from (Berthon et al., 2017, Proof of Lemma 12), will be useful later.

$$\prod_{j=k}^{\infty}(1 - 2^{-j}) \geq \exp\left(-2^{-(k-2)}\right)$$

Towards a formula to compute $M(\varepsilon)$, we derive the following bounds from the above inequality.

$$\begin{aligned}
\exp\left(-2^{-(k-2)}\right) \geq 1-\varepsilon &\iff -2^{-(k-2)} \geq \ln(1-\varepsilon) \\
&\iff 2^{-(k-2)} \leq \ln(\varepsilon) \\
&\iff (k-2) \geq -\log_2(\ln(\varepsilon)) \\
&\iff k \geq 2 - \log_2(\ln(\varepsilon)) \qquad \Longrightarrow \prod_{j=k}^{\infty}(1-2^{-j}) \geq 1-\varepsilon
\end{aligned}$$

Let us now set $M(\varepsilon) \stackrel{\text{def}}{=} \max(K_1, 2-\log_2(\ln(\varepsilon)))$ and denote by E_ℓ the event

$$\bigcap_{k=M(\varepsilon)}^{\ell} \left\{\rho \mid \mathbf{Avg}_k(\rho) \geq \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}\left[\mathbf{Avg}_k\right] - \varepsilon\right\}.$$

It follows from the above arguments that the probability measure of E_ℓ is at least $\prod_{k=M(\varepsilon)}^{\ell}(1-2^{-k})$. Furthermore, we have that $E_{\ell'} \subseteq E_\ell$ for all $\ell \leq \ell'$. Hence, we get (Baier and Katoen, 2008, Page 756) that

$$\Pr^{q_0}_{\mathcal{M}^\sigma}\left[\bigcap_{\ell \geq M(\varepsilon)} E_\ell\right] \geq \prod_{k=M(\varepsilon)}^{\infty}(1-2^{-k}) \geq 1-\varepsilon$$

which concludes the proof.

Item 2 We will now make use of item 1 to prove item 2. Consider a sequence $(\varepsilon_i)_{i\in\mathbb{N}}$ such that $\varepsilon_i = 2^{-i}$. It should be clear that, if we write E_i for the event

$$\left\{\rho \mid \exists M \in \mathbb{N}, \forall k \geq M,\ \mathbf{Avg}_k(\rho) \geq \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}\left[\mathbf{Avg}_k\right] - \varepsilon_i\right\},$$

we have that $E_k \subseteq E_j$ for all $j \leq k$. Furthermore, it follows from item 1 that $\Pr^{q_0}_{\mathcal{M}^\sigma}[E_i] \geq 1-2^{-i}$ for all $i \geq 0$. Hence, we can once more use the limit of the probabilities of the E_i and conclude that

$$\Pr^{q_0}_{\mathcal{M}^\sigma}\left[\bigcap_{i\in\mathbb{N}} E_i\right] = \lim_{i\to\infty} 1-\varepsilon_i = 1.$$

The claim thus follows since

$$\begin{aligned}
\bigcap_{i\in\mathbb{N}} E_i &= \left\{\rho \mid \forall i \in \mathbb{N}, \exists M \in \mathbb{N}, \forall k \geq M,\ \mathbf{Avg}_k(\rho) \geq \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}\left[\mathbf{Avg}_k\right] - \varepsilon_i\right\} \\
&= \left\{\rho \;\middle|\; \liminf_{i\in\mathbb{N}_{>0}} \left(\mathbf{Avg}_k(\rho) - \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}\left[\mathbf{Avg}_k\right]\right) \geq 0\right\} \\
&= \left\{\rho \;\middle|\; \liminf_{i\in\mathbb{N}_{>0}} \mathbf{Avg}_k(\rho) \geq \limsup_{i\in\mathbb{N}_{>0}} \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}\left[\mathbf{Avg}_k\right]\right\} \\
&= \left\{\rho \;\middle|\; \mathbf{MP}(\rho) \geq \limsup_{i\in\mathbb{N}_{>0}} \mathrm{E}^{q_0}_{\mathcal{M}^\sigma}\left[\mathbf{Avg}_k\right]\right\} \qquad \text{by definition}
\end{aligned}$$

and the probability measure of the last event above is, with probability 1, equivalent to the set of runs whose mean payoff is at least the expected mean payoff by Lemma 3. □

B PROOF OF THEOREM 1

We begin by arguing that for all ε one can compute a value for L such that the sequence $(\widehat{P}_i)_{i\in\mathbb{N}}$ of approximate probabilistic transition functions computed by σ_ε are $\frac{\varepsilon}{4}$-close to the unknown function P with probability $1-\varepsilon$.

B.1 EXPLORATION

Our goal in this section is to prove the following result. It is stated with respect to *empirical approximations* $\widehat{P}_i$ of P. Technically, such $\widehat{P}_i$ can be obtained by dividing the number of times a transition (q, a, q') has been observed compared to the number of times action $a \in A(q)$ has been executed from q.

Proposition 3. *For all $\varepsilon \in (0,1)$ one can compute $L \in \mathbb{N}$ such that the sequence $(\widehat{P}_i)_{i\in\mathbb{N}}$ of approximate functions computed by σ_ε satisfies the following*

$$\forall i \in \mathbb{N},\ \Pr^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}}\left[\rho : \left\|P - \widehat{P}_i\right\|_\infty \leq \varepsilon\right] \geq 1 - \varepsilon.$$

The result is a corollary of Lemma 5. The Lemma gives us a bound on the number of $|Q|$-step episodes for which we need to exercise a uniformly-random exploration strategy to obtain the desired approximations of P and R with at least some given probability. It can be proved via a simple application of Hoeffding's inequality.

Lemma 5. *For all $\varepsilon, \delta \in (0,1)$ one can compute $n \in \mathbb{N}$ (exponential in $|Q|$ and polynomial in $|A|$, $\mathbf{p}_{\min}^{-1}$, $\ln(\delta^{-1})$, and ε^{-1}) such that following uniformly-random exploration strategy during n (potentially non-consecutive) episodes of $|Q|$-steps suffices to collect enough information so that the empirical approximation $\widehat{P}$ is such that*

$$\Pr\left(\left\|P - \widehat{P}\right\|_\infty \leq \varepsilon\right) \geq 1 - \delta.$$

B.2 EXPLOITATION

We will now build upon Proposition 3 and argue that we can also compute O large enough so as to ensure that the expected average reward of every episode is at least $\mathbf{Val}(\mathcal{M})$ with probability $1 - \varepsilon$. For a run $\rho = q_0 \dots q_i \dots q_{i+L+O} \dots$, let $\mathbf{Ep}_i$ denote $\mathbf{Avg}(q_i \dots q_{i+L+O})$.

Proposition 4. *For all $\varepsilon \in (0,1)$ and all $L \in \mathbb{N}$, one can compute $O \in \mathbb{N}$ such that*

$$\forall i \in \mathbb{N} : \mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}}\left[\mathbf{Ep}_i\right] \geq \mathbf{Val}(\mathcal{M}) - \varepsilon.$$

The proof of the above claim will have to make use of a "simulation lemma" since we are exercising an optimal strategy for a learnt model, not the actual unknown MDP.

Robustness (a.k.a. simulation) lemma. The following result captures the intuition that some expectation-optimal strategies for MDPs whose transition function have the same support are "robust". That is, when used to play in another MDP with the same support and close transition functions, they achieve near-optimal expectation. The specific lemma we use follows from the work of Solan (2003, Theorem 6) and Chatterjee (2012, Theorem 5).

Lemma 6. *Let $\varepsilon \in (0,1)$, $\widehat{P}$ be a probabilistic transition function and $\widehat{R}$ a reward function. For all memoryless deterministic expectation-optimal strategies σ for the MDP $(Q, q_0, A, \widehat{P}, \widehat{R})$ we have that*

$$\left|\mathrm{E}^{q_0}_{\mathcal{M}^\sigma}\left[\mathbf{MP}\right] - \mathbf{Val}(\mathcal{M})\right| \leq \varepsilon$$

if $\mathrm{supp}(P) = \mathrm{supp}(\widehat{P})$, $\left\|R - \widehat{R}\right\|_\infty \leq \frac{\varepsilon}{4}$, *and*

$$\left\|P - \widehat{P}\right\|_\infty \leq \frac{\varepsilon \cdot \mathbf{p}_{\min}}{24|Q|}.$$

It is important to note that there always exist memoryless deterministic expectation-optimal strategies (Gimbert, 2007) which are, furthermore, also unichain Bruyère et al. (2014).

Using the robustness lemma. We now turn to the proof of the proposition. In general terms, we will give a bound on the time we need to optimize using a strategy computed from the approximated MDP. The strategy gets us close to the desired value according to the robustness lemma, however, we have to be able to stop the exploitation. To do the latter, we make use of classical algorithms which give us **exact mixing** in (expected) bounded time even in unknown Markov chains.

Proof of Proposition 4. Consider the i-th episode of exploration and exploitation dictated by σ_ε. Let L be such that with probability at least $1 - \varepsilon/4$ we have that

- $\left\| P - \widehat{P}_i \right\|_\infty \leq \eta$ for $\eta < \mathbf{p}_{\min}$ (see Lemma 5) so that $\mathrm{supp}(P) = \mathrm{supp}(\widehat{P}_i)$ and $\widehat{R}_i = R$, and
- such that

$$\eta \leq \frac{\varepsilon \cdot \mathbf{p}_{\min}}{4 \cdot 24|Q|}$$

so that any unichain memoryless deterministic expectation-optimal strategy σ_i for $(Q, q_0, A, \widehat{P}_i, \widehat{R}_i)$ is $(\varepsilon/4)$-optimal for $\mathcal{M}$.

This means that the expectation of the error is bounded by $\varepsilon/2$. In symbols, we have the following.

$$\mathbf{Val}(\mathcal{M}) - \mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}}\left[\mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_i}}\left[\mathbf{MP}\right]\right] \leq \frac{\varepsilon}{2} \tag{2}$$

We will now make use of the fact that there is a randomized algorithm which stops a finite irreducible unknown Markov chain **precisely** when the stationary distribution is reached (See, e.g., Lovász and Winkler, 1995; Propp and Wilson, 1998). Crucially, the one from Lovász and Winkler (1995) gives a stopping time with expectation upper-bounded by a polynomial of the maximal hitting time. Using our notation, this bound is $M' := 6|Q|^2\mathbf{p}_{\min}^{2|Q|}$. (Note that the cited results have been developed for finite irreducible chains. Nevertheless, they clearly extend to finite unichain MCs and thus to $\mathcal{M}^{\sigma_i}$.)

From (Levin and Peres, 2017, Theorem 6.15) we know that for any timestep t after

$$M := 4\varepsilon^{-1}M' = \frac{24|Q|^2\mathbf{p}_{\min}^{2|Q|}}{\varepsilon}$$

we have that, for P the probabilistic transition relation of $\mathcal{M}^{\sigma_i}$ and π its stationary distribution, the following holds.

$$\frac{1}{t}\sum_{i=0}^{t-1} P^i(q, q') - \pi(q') \leq \frac{\varepsilon}{4}$$

for all $q, q' \in Q$.

We can now choose $O' \in \mathbb{N}$ to be any integer large enough so that

$$\frac{O'}{L + M + O'} \geq O' - \frac{\varepsilon}{4}.$$

Intuitively, this means the proportion of time we spend after having reached the stationary distribution ($\pm\varepsilon/4$) accounts for $1 - \varepsilon/4$ of the time the episode takes. Since the rewards are bounded in $[0, 1]$, this means (See, e.g., Norris, 1998; Puterman, 2005) that

$$\mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}}\left[\mathbf{Ep}_i\right] \geq \lim_{k\to\infty} \mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_i}}\left[\mathbf{Avg}_k\right] - \frac{\varepsilon}{2}.$$

From Lemma 3 we then have that the limit on the right-hand side can be replaced by $\mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_i}}\left[\mathbf{MP}\right]$. Together with Equation (2) we thus get

$$\mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}}\left[\mathbf{Ep}_i\right] \geq \mathbf{Val}(\mathcal{M}) - \varepsilon$$

and since i was arbitrary, the result holds with $O = M + O'$. □

Putting everything together. We are now ready to give a proof of the theorem.

Proof of Theorem 1. Using Proposition 4 we obtain that

$$\lim_{i\to\infty} \mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}}\left[\mathbf{Ep}_i\right] \geq \mathbf{Val}(\mathcal{M}) - \varepsilon.$$

Observe now that σ_ε is a finite-state unichain strategy since it randomly explores the whole end component and forgets all it has learned infinitely often. Hence, Lemma 3 holds and we have that

$$\lim_{i\to\infty} \mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}} [\mathbf{Ep}_i] = \lim_{i\to\infty} \mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}} [\mathbf{Avg}_i] = \mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}} [\mathbf{MP}] .$$

In turn, the latter implies that

$$\mathrm{E}^{q_0}_{\mathcal{M}^{\sigma_\varepsilon}} [\mathbf{MP}] \geq \mathbf{Val}(\mathcal{M}) - \varepsilon.$$

Furthermore, by Lemma 4 item 2 the expectation is achieved with probability 1. □

C PROOF OF PROPOSITION 1 (OPTIMALITY OF σ_ε)

Proof of Proposition 1. Observe that taking always action a yields a mean payoff of 0.5, and b yields p. Therefore, depending on whether $p < 0.5$, the former or the latter is optimal.

Let σ be a finite-memory strategy given by a Mealy machine with n states and let $p = 0.25$. Assume there exists a run in $\mathcal{M}^\sigma_p$ such that during some $n+1$ consecutive visits of q_0, action a is always chosen with probability 1. Then, during these visits, σ has revisited q_0 with the same memory state twice. Between these two visits, action a and the successor state are always chosen deterministically with probability 1. Since the transition function on memory states of the Mealy machine is also deterministic and because the choice of actions depends only on the memory states, σ continues looping through these memory states while always choosing action a deterministically. Thus, this infinite suffix of the run always switches between q_0 and q_1 with probability 1 and therefore this infinite suffix has probability 1. Moreover, by definition of a run, the remaining finite prefix of the run must have positive probability. These two properties still hold if we change p to 0.75 because the probabilities of the transitions between q_0 and q_1 remain unchanged and because all transitions for the finite prefix still have positive probability. Hence, we get that the complete run has positive probability in $\mathcal{M}^\sigma_{0.75}$. However, the run has a suboptimal mean payoff of 0.5 while the optimum would be 0.75.

Otherwise, we get that during all $n+1$ consecutive visits of q_0 on all runs in $\mathcal{M}^\sigma_{0.25}$, there is some positive probability of choosing action b, bounded from below by the smallest of these positive probabilities p' among all memory states. Consequently, by the law of large numbers, that action b will be chosen almost surely at least a $p'/(n+1)$-fraction of the time. Since $p = 0.25$, the law of large numbers also implies that action b almost surely gives an average of 0.25 reward. Thus, the mean payoff is almost surely at most $0.25 \cdot p'/(n+1) + 0.5 \cdot (1 - p'/(n+1)) < 0.5$, i.e. the mean payoff is almost surely suboptimal. □

D PROOF OF THEOREM 2 (CORRECTNESS OF ALGORITHM 1)

Proof of Theorem 2. First note that each PEC is contained in the single component (Q, A) of M at the beginning of the algorithm. Moreover, if a PEC (S', B') is contained in a component $(S, B) \in M$ during the algorithm, then the graph $\mathcal{G}_{S',B',\overline{T}}$ for the PEC is a subset of the graph $\mathcal{G}_{S,B,\overline{T}}$ for the component (S, B). Since $\mathcal{G}_{S',B',\overline{T}}$ is strongly connected, the states S' of the PEC are also strongly connected in $\mathcal{G}_{S,B,\overline{T}}$. Therefore, they all belong to a common SCC and hence the algorithm adds a component containing the PEC into M'. Inductively, it follows that each PEC is contained in one component of M after the algorithm finishes.

Since M stayed unchanged after the last iteration, for each component $(S, B) \in M$ the graph $\mathcal{G}_{S,B,\overline{T}}$ is strongly connected, and because $B = \underline{B}_S$, it follows that (S, B) is a PEC. Moreover, since every MPEC has an upper bound in M which is itself a PEC, by maximality this upper bound has to be the MPEC and thus every MPEC is contained in M. Furthermore, consider a PEC $(S, B) \in M$. Because M partitions the states of the MDP, (S, B) has only itself as upper bound in M. But every PEC is contained in some MPEC which itself, as argued before, has to be included in M. Thus, this MPEC in M is an upper bound to (S, B), hence the MPEC has to be (S, B) and therefore every component $(S, B) \in M$ is a MPEC.

Finally, because the algorithm always partitions the states of the components of M, it has to finish after at most $|Q|$ iterations and can therefore be implemented with a runtime of $\mathcal{O}(|Q||\overline{T}|)$ since SCCs can be calculated in linear time. □

E PROOF OF THEOREM 3 (CORRECTNESS OF ALGORITHM 2)

Proof of Theorem 3. First note that each SEC is contained in S at the beginning of the algorithm. Moreover, if a SEC is contained in S during the algorithm, then all its states will get added to S'. Indeed, consider the end of the while-loop. Then we have for each state $s \in S \setminus S'$ that all actions $a \in \overline{B}_S(s)$, which would include all actions of the SEC, are also in $\underline{B}_{Q \setminus S'}(s)$, i.e. all must and at least one must or may transition for this action stay outside of S'. But this means that it is possible to instantiate exactly those transitions, i.e. we can choose a transition relation T with $\operatorname{supp}(T, s, a) = \operatorname{supp}(\overline{T}, s, a) \cap (Q \setminus S')$ for all these s and a, giving rise to a graph $\mathcal{G}_{S,\overline{B}_S,T}$ where no state $s \in S \setminus S'$ has an edge to a node in S' by construction of T. So all states in $S \setminus S'$ have no path to $s_0 \in S'$ in that graph. Because the graph for the SEC is a subgraph of this graph, it follows that all states of the SEC have to be contained in S' by the second property of SECs. Inductively, it follows that each SEC is contained in S after the repeat-until-loop finishes.

Since each SEC is in S after the repeat-until-loop, thus also has only actions of $\overline{B}_S$, and since each SEC is a PEC and therefore strongly connected, it follows that all states of the SEC are reachable from s_0 in $\mathcal{G}_{S,\overline{B}_S,\overline{T}}$. Hence, all SECs are a subset of (S, B) after the end of the algorithm.

Next we show that the second constraint of SECs holds for (S, B) after the end of the algorithm. Consider an arbitrary transition relation T with $\operatorname{supp}(T, s) = A(s) = \operatorname{supp}(\overline{T}, s)$ for all $s \in S$. Since S stayed unchanged in the last iteration of the repeat-until-loop, we can show that all states in S have a path to s_0 in $\mathcal{G}_{S,\overline{B}_S,T}$. This can be done inductively by the number of the iteration when a state s has been added to S'. The action $a \notin \underline{B}_{Q \setminus S'}(s)$, for which it has been added, tells us that either a must transition or all may transitions for action a go to S'. Because $\underline{T} \subseteq T$ and $\operatorname{supp}(T, s) = \operatorname{supp}(\overline{T}, s)$, this means that either such a must transition or one of those may transitions has to exist in T. In all cases, there is a transition in T connecting s to some state in S'. Since S stayed unchanged, this action is in $\overline{B}_S$ after the repeat-until-loop. Moreover, since s is in S after the end of the algorithm, it is reachable from s_0 in $\mathcal{G}_{S,\overline{B}_S,\overline{T}}$. As $a \in \overline{B}_S(s)$, all successors $\operatorname{supp}(\overline{T}, s, a)$ of s after action a are therefore also reachable from s_0 in that graph, giving that they have to be in S after the end of the algorithm, in particular action a is also still in $\overline{B}_S(s)$. Using the transition of T from s to some state in S', by induction we get the desired path in $\mathcal{G}_{S,\overline{B}_S,T}$ to s_0.

Finally, strong connectivity of $\mathcal{G}_{S,\overline{B}_S,\overline{T}}$ follows from the fact that with $T = \overline{T}$, each state has path to s_0 in $\mathcal{G}_{S,\overline{B}_S,\overline{T}}$ by the previous argument, and every state is reachable from s_0 by construction. Since $B = \overline{B}_S$, this gives that (S, B) is a SEC, and because it contains all SECs, it has to be an MSEC.

The algorithm finishes after at most $|Q|$ iterations since it always removes at least one state from S in each iteration. Moreover, the while-loop can be implemented in linear time using reference counting. Thus, the algorithm can be implemented with a runtime of $\mathcal{O}(|Q||\overline{T}|)$. □

F PROOF OF THEOREM 4

Proof of Theorem 4. First we note that the L computed for Theorem 1 in Lemma 5 depends only on $\mathbf{p}_{\min}$, $|Q|$, $|A|$ and ε. It does not depend on the structure of $\mathcal{M}$, so we do not need the transition relation in order to compute L. Then we can also compute the corresponding O as in Proposition 4 which also depends only on the mentioned terms.

Let's fix a MEC (S, B) of $\mathcal{M}$ with $\Pr^{q_0}_{\mathcal{M}^{\sigma_p}}[\operatorname{Inf} \subseteq S] > 0$ and denote by $\mathcal{N} = (S, s_0, B, P|_{S \times A}, R|_{S \times A \times S})$ the restriction of $\mathcal{M}$ to that MEC. In particular, $\mathbf{Val}(\mathcal{N}) = \mathbf{Val}(\mathcal{M} \mid S, B)$. Then execution of σ_ε with our choice of L and O in $\mathcal{N}$ will give an ε-optimal mean-payoff almost surely according to Theorem 1. We might not have exactly the same L and O as those from Theorem 1 because it might hold $|S| < |Q|$, however, the L chosen here is definitely larger than the one from Theorem 1, and therefore the choice of O ensures that the result still holds.

For $n \in \mathbb{N}$, $q_n \in Q$ and (S', B') with $S' \subseteq Q$ and $B'(s) \subseteq A(s)$ for all $s \in S'$, define

$$\operatorname{Runs}^{q_0,q_n}_{S',B'}(\mathcal{M}^{\sigma_p}) \stackrel{\text{def}}{=} \{h_0 h_1 \ldots \in \operatorname{Runs}^{q_0}(\mathcal{M}^{\sigma_p}) : \operatorname{last}(h_n) = q_n \wedge (S', B') \text{ is the MPEC containing } q_n \text{ at time } n\}$$

as the set of all runs being in q_n at time step n and with (S', B') being the MPEC computed by σ_p at that time step. (S', B') could be different from the MPEC which one would compute using the original transition relation bounds since σ_p might update those bounds. Moreover, we write

$$\operatorname{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p}) \stackrel{\text{def}}{=} \left\{h_0 h_1 \cdots \in \operatorname{Runs}^{q_0,q_n}_{S',B'}(\mathcal{M}^{\sigma_p}) : \operatorname{last}(h_{n-1}) \notin S \wedge \forall k \geq n : \operatorname{last}(h_k) \in S\right\}$$

as the set of all runs from $\mathrm{Runs}^{q_0,q_n}_{S',B'}(\mathcal{M}^{\sigma_p})$ such that n is the first time step after which the runs stay in S forever. Then we get

$$\{\mathrm{Inf} \subseteq S\} = \bigcup_{n,q_n,S',B'} \mathrm{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p})$$

as a countable disjoint union. Set $\mathrm{Opt} = \{\rho \in \mathrm{Runs}^{q_0}(\mathcal{M}^{\sigma_p}) : \mathbf{MP}(\rho) \geq \mathbf{Val}(\mathcal{N}) - \varepsilon\}$ as the set of ε-optimal runs w.r.t. the mean-payoff achievable in $\mathcal{N}$, i.e. in the MEC (S, B). Then, the law of total probability gives us that

$$\Pr^{q_0}_{\mathcal{M}^{\sigma_p}}[\mathrm{Opt} \mid \mathrm{Inf} \subseteq S] = \sum_{n,q_n,S',B'} \Pr^{q_0}_{\mathcal{M}^{\sigma_p}}\left[\mathrm{Opt} \;\middle|\; \mathrm{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p})\right] \Pr^{q_0}_{\mathcal{M}^{\sigma_p}}\left[\mathrm{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p}) \;\middle|\; \mathrm{Inf} \subseteq S\right]$$

And therefore we get with $\mathcal{I} = \left\{(n, q_n, S', B') : \Pr^{q_0}_{\mathcal{M}^{\sigma_p}}\left[\mathrm{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p})\right] > 0\right\}$ that

$$\Pr^{q_0}_{\mathcal{M}^{\sigma_p}}[\mathrm{Opt} \mid \mathrm{Inf} \subseteq S] \geq \inf_{(n,q_n,S',B') \in \mathcal{I}} \Pr^{q_0}_{\mathcal{M}^{\sigma_p}}\left[\mathrm{Opt} \;\middle|\; \mathrm{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p})\right]$$

If we can show

$$\Pr^{q_0}_{\mathcal{M}^{\sigma_p}}\left[\mathrm{Opt} \;\middle|\; \mathrm{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p})\right] = 1$$

for all $(n, q_n, S', B') \in \mathcal{I}$, then this proves the claim.

Therefore, fix some $(n, q_n, S', B') \in \mathcal{I}$. Since (S, B) is a MEC, (S, B) is also a PEC and therefore contained in the MPEC (S', B'), i.e. $S \subseteq S'$ and $B(s) \subseteq B'(s)$ for all $s \in S$. We claim that $B'|_S = B$. Assume the opposite is true, i.e. there exists an $s \in S$ and $a \in B'(s) \setminus B(s)$. Since (S, B) is a MEC and $a \notin B(s)$, it holds $\mathrm{supp}(P(s, a)) \not\subseteq S$, so there exists a transition leading out of S after action a in state s. As S is strongly connected using the actions in B which are included in B', from every state in S there exists a path of length at most $|Q| - 1$ leading to state s. Since σ_p employs an exploration strategy λ in (S', B'), during all consecutive $|Q|$ steps of the execution of λ there is a (constant) positive probability that the strategy will choose the path to s and then leave the MEC using the transition after action a, so the probability of staying in S after these $|Q|$ steps is less than one. Since σ_p will execute infinitely many $|Q|$ steps with strategy λ, it follows that the probability of staying in S is zero, contradicting

$$\Pr^{q_0}_{\mathcal{M}^{\sigma_p}}\left[\mathrm{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p})\right] > 0$$

Finally, this means that λ is an exploration strategy in (S, B), thus from moment n onwards σ_p executes σ_ε in $\mathcal{N}$. Hence,

$$\Pr^{q_0}_{\mathcal{M}^{\sigma_p}}\left[\mathrm{Opt} \;\middle|\; \mathrm{Runs}^{q_0,q_n}_{S',B'}(\mathcal{M}^{\sigma_p})\right] = 1$$

and therefore also

$$\Pr^{q_0}_{\mathcal{M}^{\sigma_p}}\left[\mathrm{Opt} \;\middle|\; \mathrm{Runs}^{q_0,q_n}_{S',B',S}(\mathcal{M}^{\sigma_p})\right] = 1 \qquad \square$$

G PROOF OF LEMMA 1

Proof of Lemma 1. We start by showing that all s_0-EC-safe strategies σ only choose actions within the MSEC (S, B). Therefore, let σ be an s_0-EC-safe strategy. Assume there exists a history $h \in \mathrm{Hist}^{s_0}(\overline{\mathcal{M}})$ with $\mathrm{last}(h) \in S$ (and positive probability under σ) such that $\mathrm{supp}(\sigma(h)) \not\subseteq B(\mathrm{last}(h))$. Let h be the shortest history with this property.

We show in the following paragraphs that we can construct a transition function P' such that $P'|_{S\times A} = \overline{P}|_{S\times A}$, i.e. P' coincides with $\overline{P}$ on (S, B), and such that no state outside of S is in the same MEC as s_0. Then, since $\mathrm{supp}(\sigma(h)) \not\subseteq B(\mathrm{last}(h))$, we have that there exists an action $a \in \mathrm{supp}(\sigma(h)) \setminus B(\mathrm{last}(h))$, and because this action is not in the MSEC (S, B), it has to hold $\mathrm{supp}(\overline{T}, \mathrm{last}(h), a) \not\subseteq S$. But then this action cannot belong to the MEC containing s_0 in $\mathcal{M}' = (Q, s_0, A, P', R)$ since the MEC only has states from S (by the assumption on how P' is constructed), but $\mathrm{supp}(P'(\mathrm{last}(h), a)) = \mathrm{supp}(\overline{P}(\mathrm{last}(h), a)) = \mathrm{supp}(\overline{T}, \mathrm{last}(h), a) \not\subseteq S$, contradicting the second property of ECs. Moreover, because h is the shortest history with the given properties, h has only been in the MSEC (S, B) so far and therefore, since P' coincides with $\overline{P}$ on (S, B), we get that h also has positive probability under σ in $\mathcal{M}'$. Thus, this gives a contradiction to the fact that σ is s_0-EC-safe in $\mathcal{M}'$.

Instead of constructing P', we will actually construct a transition relation T which coincides with $\mathrm{supp}(\overline{P})$ on $S \times A$ and which has the property that if $\mathrm{supp}(P') = T$, then no state outside of S is in the same MEC as s_0. Then we can choose an arbitrary transition function P' with $\mathrm{supp}(P') = T$ and $P'|_{S\times A} = \overline{P}\big|_{S\times A}$ (this is possible since T and $\mathrm{supp}(\overline{P})$ coincide on $S \times A$) which then gives the desired properties of P'.

Since T should coincide with $\mathrm{supp}(\overline{P})$ on $S \times A$, we only have to define T outside of S. For this we use the same construction as the one in the proof of Theorem 3. Recall this construction. For all states $s \in S$ not added to S' in the while-loop, we instantiate T such that $\mathrm{supp}(T, s, a) = \mathrm{supp}(\overline{T}, s, a) \cap (Q \setminus S')$ for all $a \in \overline{B}_S(s)$ and $\mathrm{supp}(T, s, a) = \mathrm{supp}(\overline{T}, s, a)$ for all other actions. Here, S denotes the set S from the algorithm.

Then, the states removed from S in the first iteration of the algorithm have no transition in T connecting them to states in S', because at that point it did hold $S = Q$, hence $\overline{B}_S(s) = A(s)$ and therefore $\mathrm{supp}(T, s, a) \cap S' = \mathrm{supp}(\overline{T}, s, a) \cap (Q \setminus S') \cap S' = \varnothing$ for all $a \in A(s)$. In particular, they have no path to $s_0 \in S'$ and can therefore not be in the same MEC as s_0.

For all states s removed in later iterations from S, we have $\overline{B}_S(s) = \underline{B}_{Q\setminus S'}(s)$. Thus, there are two cases to consider for an action $a \in A(s)$. Either $a \notin \overline{B}_S(s)$. Then we have that $\mathrm{supp}(\overline{T}, s, a) \nsubseteq S$. But since states removed earlier on in the algorithm from S can't be in the same MEC as s_0, action a as well cannot belong to that MEC by the second property of ECs. Thus, the MEC could only contain actions $a \in \overline{B}_S(s) = \underline{B}_{Q\setminus S'}(s)$. But for those actions it holds $\mathrm{supp}(T, s, a) = \mathrm{supp}(\overline{T}, s, a) \cap (Q \setminus S')$, i.e. for these actions, T has no transition to any state of S'. Hence, no state in $S \setminus S'$ could have a path to any state of S' in the MEC. Thus, since $s_0 \in S'$, the second property of ECs tells us that the states in $S \setminus S'$ cannot be in the same MEC as s_0.

Overall we get that all states removed at some point from S cannot be in the same MEC as s_0 for the transition relation T. Hence, this construction provides the required transition relation T such that no state outside the MSEC (S, B) can be in the same MEC as s_0. This concludes the proof of the first part of the Lemma. Hence, all s_0-EC-safe strategies are in the MSEC w.r.t. s_0.

To show the reverse direction, we actually show the additional claim that every strategy σ which is in the MSEC (S, B) w.r.t. s_0 is also in the maximal EC (S', B') within the MSEC (S, B) containing s_0, i.e. (S', B') is an EC with $s_0 \in S' \subseteq S$, $B'(s) \subseteq B(s)$ for all $s \in S$ and it is the maximal EC with these properties. Since the strategy is in an EC containing s_0, this directly implies that the strategy is also in the MEC containing s_0, hence s_0-EC-safe. Showing this therefore concludes the complete proof.

We claim that S' consists of all states reachable from s_0 in $\mathcal{G}_{S,B,\mathrm{supp}(P)}$ and $B' = B|_{S'}$. On the one hand, since $\mathcal{G}_{S',B',\mathrm{supp}(P)}$ must be strongly connected by the second property of ECs and because (S', B') is a subset of (S, B), we get that $\mathcal{G}_{S',B',\mathrm{supp}(P)}$ is a subgraph of $\mathcal{G}_{S,B,\mathrm{supp}(P)}$ and therefore all states of S' have to be reachable from s_0 in that graph. On the other hand, by the second property of SECs, we have that all states in S' have a path to s_0 in $\mathcal{G}_{S,B,\mathrm{supp}(P)}$. But since all successors of S' in that graph are by construction also included in our set S', we get that this path also exists in $\mathcal{G}_{S',B',\mathrm{supp}(P)}$. This gives the strong connectivity, and the first property of ECs is satisfied by construction. So our choice of (S', B') is indeed an EC, and by the first argument, it is the maximal EC with the required properties.

In particular, this characterisation gives that all actions of $B(s)$ for a state $s \in S'$ in that EC are part of the EC, and therefore σ is in this maximal EC. □

H PROOF OF THEOREM 5

Proof of Theorem 5. We choose the L and O as in the proof of Theorem 4. Define (S, B) to be the maximal EC containing s_0 within the MSEC (S', B') w.r.t. s_0 and denote by $\mathcal{N} = (S, s_0, B, P|_{S\times A}, R|_{S\times A\times S})$ the restriction of $\mathcal{M}$ to that EC. As seen before, employing σ_ε with our choice of L and O in $\mathcal{N}$ would give us a mean-payoff of at least $\mathbf{Val}(\mathcal{N}) - \varepsilon$ almost surely.

Note that Lemma 1 gives us that s_0-EC-safe strategies are exactly the strategies in (S, B), hence $\mathbf{sVal}(\mathcal{M}, s_0) = \mathbf{Val}(\mathcal{N})$. Furthermore, this Lemma states that σ_s is in (S, B) as well, i.e. it chooses only actions from $B(s)$ for all $s \in S$. Because σ_s employs an exploration strategy λ in (S', B') and $B(s) \subseteq B'(s)$ for all $s \in S$, this implies that actually $B(s) = B'(s)$ for all $s \in S$, i.e. $B'|_S = B$. Thus, λ is an exploration strategy in (S, B), therefore σ_s is the

strategy σ_ε in $\mathcal{N}$ and finally this gives

$$\Pr^{s_0}_{\mathcal{M}^{\sigma_s}}[\rho : \mathbf{MP}(\rho) \geq \mathbf{sVal}(\mathcal{M}, s_0) - \varepsilon] = \Pr^{s_0}_{\mathcal{M}^{\sigma_s}}[\rho : \mathbf{MP}(\rho) \geq \mathbf{Val}(\mathcal{N}) - \varepsilon] = 1 \qquad \square$$

References

Baier, C. and Katoen, J.-P. (2008). *Principles of model checking*. MIT Press.

Berthon, R., Randour, M., and Raskin, J.-F. (2017). Threshold constraints with guarantees for parity objectives in Markov decision processes. In *ICALP*, pages 121:1–121:15. Dagstuhl.

Bruyère, V., Filiot, E., Randour, M., and Raskin, J.-F. (2014). Meet your expectations with guarantees: Beyond worst-case synthesis in quantitative games. In *STACS*, pages 199–213. Dagstuhl.

Chatterjee, K. (2012). Robustness of structurally equivalent concurrent parity games. In *FoSSaCS*, pages 270–285. Springer.

Gimbert, H. (2007). Pure stationary optimal strategies in Markov decision processes. In *STACS*, pages 200–211. Springer.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*. American Mathematical Soc.

Lovász, L. and Winkler, P. (1995). Exact mixing in an unknown Markov chain. *Electr. J. Comb.*

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer.

Norris, J. R. (1998). *Markov chains*. Cambridge University Press.

Propp, J. G. and Wilson, D. B. (1998). How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *J. Algorithms*, (2):170–217.

Puterman, M. L. (2005). *Markov Decision Processes*. Wiley-Interscience.

Solan, E. (2003). Continuity of the value of competitive Markov decision processes. *Journal of Theoretical Probability*, (4):831–845.

Tracol, M. (2009). Fast convergence to state-action frequency polytopes for MDPs. *Opereration Research Letters*, (2):123–126.