Markov Reward Models and Markov Decision Processes in Discrete and Continuous Time: Performance Evaluation and Optimization

Alexander Gouberman and Markus Siegle

Department of Computer Science
Universität der Bundeswehr, München, Germany
{alexander.gouberman,markus.siegle}@unibw.de

Abstract. State-based systems with discrete or continuous time are often modelled with the help of Markov chains. In order to specify performance measures for such systems, one can define a reward structure over the Markov chain, leading to the Markov Reward Model (MRM) formalism. Typical examples of performance measures that can be defined in this way are time-based measures (e.g. mean time to failure), average energy consumption, monetary cost (e.g. for repair, maintenance) or even combinations of such measures. These measures can also be regarded as target objects for system optimization. For that reason, an MRM can be enhanced with an additional control structure, leading to the formalism of Markov Decision Processes (MDP).

In this tutorial, we first introduce the MRM formalism with different types of reward structures and explain how these can be combined to a performance measure for the system model. We provide running examples which show how some of the above mentioned performance measures can be employed. Building on this, we extend to the MDP formalism and introduce the concept of a policy. The global optimization task (over the huge policy space) can be reduced to a greedy local optimization by exploiting the non-linear Bellman equations. We review several dynamic programming algorithms which can be used in order to solve the Bellman equations exactly. Moreover, we consider Markovian models in discrete and continuous time and study value-preserving transformations between them. We accompany the technical sections by applying the presented optimization algorithms to the example performance models.

1 Introduction

State-based systems with stochastic behavior and discrete or continuous time are often modelled with the help of Markov chains. Their efficient evaluation and optimization is an important research topic. There is a wide range of application areas for such kind of models, coming especially from the field of Operations Research, e.g. economics [4,23,31] and health care [14,36], Artificial Intelligence, e.g. robotics, planning and automated control [11,40] and Computer Science [2,6, 12,13,21,25,34]. In order to specify performance and dependability measures for

A. Remke and M. Stoelinga (Eds.): ROCKS Autumn School 2012, LNCS 8453, pp. 156–241, 2014.

such systems, one can define a reward structure over the Markov chain, leading to the Markov Reward Model (MRM) formalism. Typical examples of performance measures that can be defined in this way are time-based measures (e.g. mean time to failure), average energy consumption, monetary cost (e.g. for repair, maintenance) or even combinations of such measures. These measures can also be regarded as target objects for system optimization. For that reason, an MRM can be enhanced with an additional control structure, leading to the formalism of Markov Decision Processes (MDP) [20]. There is a huge number of optimization algorithms for MDPs in the literature, based on dynamic programming and linear programming [7, 8, 17, 33] – all of them rely on the Bellman optimality principle [5].

In many applications, the optimization criteria are a trade-off between several competing goals, e.g. minimization of running cost and maximization of profit at the same time. For sure, in these kinds of trade-off models, it is important to establish an optimal policy which in most cases is not intuitive. However, there are also examples of target functions with no trade-off character (e.g. pure lifetime maximization [16]) which can also lead to counterintuitive optimal policies. Therefore, using MDPs for optimization of stochastic systems should not be neglected, even if a heuristically established policy seems to be optimal.

In order to build up the necessary theoretical background in this introductory tutorial, we first introduce in Sect. 2 the discrete-time MRM formalism with finite state space and define different types of reward measures typically used in performance evaluation, such as total reward, discounted reward and average reward. In contrast to the majority of literature, we follow a different approach to deduce the definition of the discounted reward through a special memoryless horizon-expected reward. We discuss properties of these measures and create a fundamental link between them, which is based on the Laurent series expansion of the discounted reward (and involves the deviation matrix for Markov chains). We derive systems of linear equations used for evaluation of the reward measures and provide a running example based on a simple queueing model, in order to show how these performance measures can be employed.

Building on this, in Sect. 3 we introduce the MDP formalism and the concept of a policy. The global optimization task (over the huge policy space) can be reduced to a greedy local optimization by exploiting the set of non-linear Bellman equations. We review some of the basic dynamic programming algorithms (policy iteration and value iteration) which can be used in order to solve the Bellman equations. As a running example, we extend the queueing model from the preceding section with a control structure and compute optimal policies with respect to several performance measures.

From Sect. 4 on we switch to the continuous-time setting and present the CT-MRM formalism with a reward structure consisting of the following two different types: impulse rewards which measure discrete events (i.e. transitions) and rate rewards which measure continuous time activities. In analogy to the discrete-time case, we discuss the performance measures given by the total, horizon-expected, discounted and average reward measures. In order to be able to evaluate these

measures, we present model transformations that can be used for discretizing the CTMRM to a DTMRM by embedding or uniformization [24]. As a third transformation, we define the continuization which integrates the discrete impulse rewards into a continuous-time rate, such that the whole CTMRM possesses only rate rewards. We further study the soundness of these transformations, i.e. the preservation of the aforementioned performance measures. Similar to DTMRMs, the discounted reward measure can be expanded into a Laurent series which once again shows the intrinsic structure between the measures. We accompany Sect. 4 with a small wireless sensor network model.

In Sect. 5 we finally are able to define the continuous-time MDP formalism which extends CTMRMs with a control structure, as for discrete-time MDPs. With all the knowledge collected in the preceding sections, the optimization algorithms for CTMDPs can be performed by MDP algorithms through time discretization. For evaluation of the average reward measure, we reveal a slightly different version of policy iteration [17], which can be used for the continuization transformation. As an example, we define a bridge circuit CTMDP model and optimize typical time-based dependability measures like mean time to failure and the availability of the system.

Figure 1.1 shows some dependencies between the sections in this tutorial in form of a roadmap. One can read the tutorial in a linear fashion from beginning to end, but if one wants to focus on specific topics it is also possible to skip certain sections. For instance, readers may wish to concentrate on Markov Reward Models in either discrete or continuous time (Sects. 2 and 4), while neglecting the optimization aspect. Alternatively, they may be interested in the discrete time setting only, ignoring the continuous time case, which would mean to read only Sects. 2 (on DTMRMs) and 3 (on MDPs). Furthermore, if the reader is interested in the discounted reward measure, then he may skip the average reward measure in every subsection.

Readers are assumed to be familiar with basic calculus, linear algebra and probability theory which should suffice to follow most explanations and derivations. For those wishing to gain insight into the deeper mathematical structure of MRMs and MDPs, additional theorems and proofs are provided, some of which require more involved concepts such as measure theory, Fubini's theorem or Laurent series. For improved readability, long proofs are moved to the Appendix in Sect. A.

The material presented in this tutorial paper has been covered previously by several authors, notably in the books of Puterman [33], Bertsekas [7, 8], Bertsekas/Tsitsiklis [10] and Guo/Hernandez-Lerma [17]. However, the present paper offers its own new points of view: Apart from dealing also with non-standard measures, such as horizon-expected reward measures and the unified treatment of rate rewards and impulse rewards through the concept of continuization, the paper puts an emphasis on transformations between the different model classes by embedding and uniformization. The paper ultimately anwers the interesting question of which measures are preserved by those transformations.

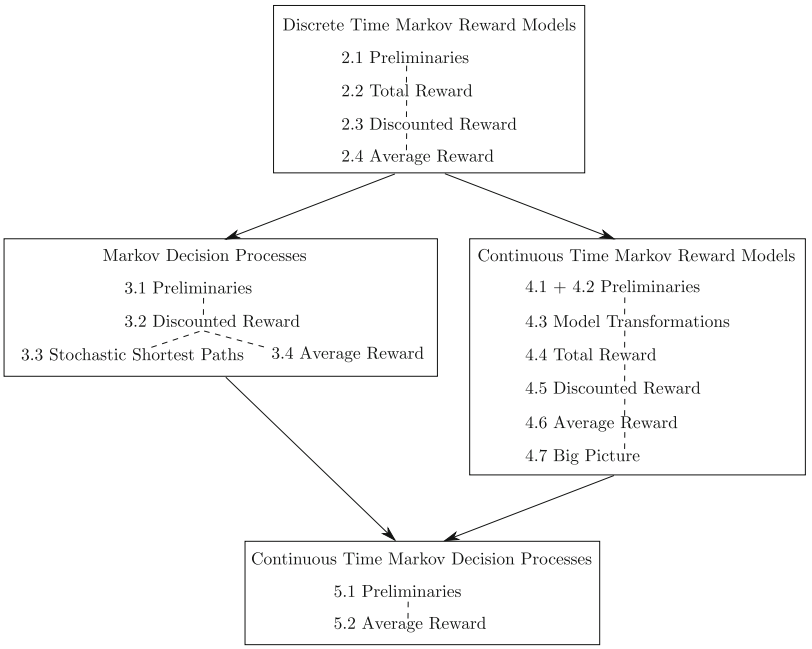

Fig. 1.1. Roadmap of the tutorial with dependencies between the sections and subsections. As an example, Sect. 3.3 on stochastic shortest paths, needs Sect. 3.2 and therefore also Sect. 2.3 but not Sect. 2.4. The Big Picture in Sect. 4.7 is one of the main goals of this tutorial and also necessary for the section on CTMDPs.

Symbols

B^A	set of functions $f : A \to B$
2^A	power set of A
$\mathcal{D}(S)$	set of probability distributions over S
P	probability measure, probability transition matrix
P_s	probability measure assuming initial state s
$\mathbb{E}_s$	expectation from initial state s
$\mathbb{1}_A(x)$	indicator function, $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 otherwise
$\delta_{s,s'}$	Kronecker-δ, $\delta_{s,s'} = 1$ if $s = s'$ and 0 otherwise
I	identity matrix
$\mathbf{1}$	column vector consisting of ones in each entry
$\overline{p}$	$\overline{p} = 1 - p$
$\ker(A)$	kernel of a matrix A
γ	discount factor, $0 < \gamma < 1$
α	discount rate, $\alpha > 0$
V	value function
g	average reward
h	bias

2 Discrete Time Markov Reward Models

2.1 Preliminaries

Definition 2.1. *For a finite set S let $\mathcal{D}(S) := \left\{\delta\colon S \to [0,1] \mid \sum_{s\in S} \delta(s) = 1\right\}$ be the space of discrete probability distributions over S.*

Definition 2.2. *A **discrete-time Markov chain (DTMC)** is a structure $\mathcal{M} = (S, P)$ consisting of a finite set of states S (also called the state space of $\mathcal{M}$) and a transition function $P\colon S \to \mathcal{D}(S)$ which assigns to each state s the probability $P(s,s') := (P(s))(s')$ to move to state s' within one transition. A **discrete-time Markov Reward Model (DTMRM)** enhances a DTMC (S,P) by a reward function $R\colon S \times S \to \mathbb{R}$ and is thus regarded as a structure $\mathcal{M} = (S,P,R)$.*

In the definition, the rewards are defined over transitions, i.e. whenever a transition from s to s' takes place, a reward $R(s,s')$ is gained. Alternatively, a DTMRM can also be defined with state-based rewards $R\colon S \to \mathbb{R}$. There is a correspondence between transition-based and state-based reward models: A state-based reward $R(s)$ can be converted into a transition-based reward by defining $R(s,s') := R(s)$ for all $s' \in S$. On the other hand, a transition-based reward $R(s,s')$ can be transformed by expectation into its state-based version by defining $R(s) := \sum_{s'\in S} R(s,s')P(s,s')$. Of course, this transformation can not be inverted, but as we will see, the reward measures that we consider do not differ. Note that for state-based rewards there are two canonical but totally different possibilities to define the point in time, when such a reward can be gained: either when a transition into the state or out of the state is performed. This corresponds to the difference in the point of view for "arrivals" and "departures" of jobs in queueing systems. When working with transition-based rewards $R(s,s')$ as in Definition 2.2, then such a confusion does not occur since $R(s,s')$ is gained in state s after transition to s' and thus its expected value $R(s)$ corresponds to the "departure" point of view. For our purposes we will mix both representions and if we write R for a reward then it is assumed to be interpreted in a context-dependent way as either the state-based or the transition-based version.

Each bijective representation

$$\varphi\colon S \to \{1,2,\ldots,n\}\,,\ n := |S| \tag{2.1}$$

of the state space as natural numbered indices allows to regard the rewards and the transition probabilities as real-valued vectors in $\mathbb{R}^n$ respectively matrices in $\mathbb{R}^{n\times n}$. We indirectly take such a representation φ, especially when we talk about P and R in vector notation. In this case the transition function $P : S \to \mathcal{D}(S)$ can be regarded as a ***stochastic*** matrix, i.e. $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1} = (1,\ldots,1)^T$ is the column vector consisting of all ones.

Example 2.1 (Queueing system). As a running example, consider a system consisting of a queue of capacity k and a service unit which can be idle, busy or

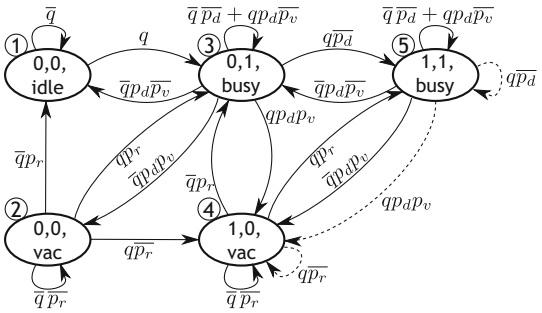

Fig. 2.1. Queueing model with queue capacity $k = 1$ and a service unit which can be idle, busy or on vacation. State $(0, 1, \text{busy})$ represents 0 jobs in the queue, 1 job in service and server is busy. The dashed transitions represent the event that an incoming job is discarded. The parameter values are $q = 0.25$, $p_d = 0.5$, $p_v = 0.1$ and $p_r = 0.25$. Overlined probabilities are defined as $\overline{p} := 1 - p$.

on vacation (Fig. 2.1) [38]. A job arrives with probability $q = 0.25$ and gets enqueued if the queue is not full. If the server is idle and there is some job waiting in the queue, the server immediately gets busy and processes the job. At the end of each unit of service time, the server is done with the job with probability $p_d = 0.5$ and can either go idle (and possibly getting the next job) or the server needs vacation with probability $p_v = 0.1$. From the vacation mode the server returns with probability $p_r = 0.25$. Figure 2.1 shows the model for the case of queue capacity $k = 1$, where transitions are split into regular transitions (solid lines) and transitions indicating that an incoming job gets discarded (dashed lines). The reward model is as follows: For each accomplished service a reward of $R_{\text{acc}} = \$100$ is gained, regardless of whether the server moves to idle or to vacation. However, the loss of a job during arrival causes a cost of $C_{\text{loss}} = -\$1000$. Therefore we consider the following state-based reward structures:

$$R_{\text{profit}} = \begin{pmatrix} 0 \\ 0 \\ R_{\text{acc}} p_d \\ 0 \\ R_{\text{acc}} p_d \end{pmatrix}, \; R_{\text{cost}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ C_{\text{loss}} q\overline{p_r} \\ C_{\text{loss}} \left(q\overline{p_d} + q p_d p_v\right) \end{pmatrix}, \; R_{\text{total}} = R_{\text{profit}} + R_{\text{cost}}, \tag{2.2}$$

where $\overline{p} := 1 - p$ for some probability p. □

There are several ways how the rewards gained for each taken transition (respectively for a visited state) can contribute to certain measures of interest. In typical applications of performance evaluation, rewards are accumulated over some time period or a kind of averaging over rewards is established. In order to be able to define these reward measures formally, we need to provide some basic knowledge on the stochastic state process induced by the DTMC part of a DTMRM.

2.1.1 Sample Space

For a DTMC $\mathcal{M} = (S, P)$ define Ω as the set of infinite paths, i.e.

$$\Omega := \left\{(s_0, s_1, s_2, \dots) \in S^{\mathbb{N}} \mid P(s_{i-1}, s_i) > 0 \text{ for all } i \geq 1\right\} \tag{2.3}$$

and let $\mathcal{B}(\Omega)$ be the Borel σ-algebra over Ω generated by the cylinder sets

$$C(s_0, s_1, \dots, s_N) := \{\omega \in \Omega \mid \omega_i = s_i \ \forall i \leq N\}.$$

Each $s \in S$ induces a probability space $(\Omega, \mathcal{B}(\Omega), P_s)$ with a probability distribution $P_s \colon \mathcal{B}(\Omega) \to \mathbb{R}$ over paths, such that for each cylinder set $C(s_0, \dots, s_N) \in \mathcal{B}(\Omega)$

$$P_s\left(C(s_0, \dots, s_N)\right) = \delta_{s_0,s} P(s_0, s_1) P(s_1, s_2) \dots P(s_{N-1}, s_N),$$

where δ is the Kronecker-δ, i.e. $\delta_{s,s'} = 1$ if $s = s'$ and 0 otherwise.

Definition 2.3. *The DTMC $\mathcal{M}$ induces the stochastic* ***state process*** $(X_n)_{n\in\mathbb{N}}$ *over Ω which is a sequence of S-valued random variables such that $X_n(\omega) := s_n$ for $\omega = (s_0, s_1, \dots) \in \Omega$.*

Note that

$$P_s(X_n = s') = \sum_{s_1,\dots,s_{n-1}} P(s, s_1) P(s_1, s_2) \dots P(s_{n-2}, s_{n-1}) P(s_{n-1}, s') = P^n(s, s'),$$

where $\sum_{s_1,\dots,s_{n-1}}$ denotes summation over all tuples $(s_1, \dots, s_{n-1}) \in S^{n-1}$. The process X_n also fulfills the ***Markov property*** (or ***memorylessness***): For all $s, s', s_0, s_1, \dots s_{n-1} \in S$ it holds that

$$P_{s_0}(X_{n+1} = s' \mid X_1 = s_1, \dots, X_{n-1} = s_{n-1}, X_n = s) = P(s, s'), \tag{2.4}$$

i.e. if the process is in state s at the current point in time n, then the probability to be in state s' after the next transition does not depend on the history of the process consisting of the initial state $X_0 = s_0$ and the traversed states $X_1 = s_1, \dots, X_{n-1} = s_{n-1}$ up to time $n-1$.
We denote the expectation operator over $(\Omega, \mathcal{B}(\Omega), P_s)$ as $\mathbb{E}_s$. For a function $f \colon S^{n+1} \to \mathbb{R}$ it holds

$$\mathbb{E}_s\left[f(X_0, X_1, \dots, X_n)\right] = \sum_{s_1,\dots,s_n} f(s, s_1 \dots, s_n) P(s, s_1) P(s_1, s_2) \dots P(s_{n-1}, s_n).$$

In vector representation we often write $\mathbb{E}[Y] := (\mathbb{E}_s[Y])_{s\in S}$ for the vector consisting of expectations of a real-valued random variable Y.

2.1.2 State Classification

In the following, we briefly outline the usual taxonomy regarding the classification of states for a discrete-time Markov chain $\mathcal{M} = (S, P)$. The state process X_n induced by $\mathcal{M}$ allows to classify the states S with respect to their recurrence

and reachability behavior. If $X_0 = s$ is the initial state then the random variable $M_s := \inf\{n \geq 1 \mid X_n = s\} \in \mathbb{N} \cup \{\infty\}$ is the first point in time when the process X_n returns to s. If along a path $\omega \in \Omega$ the process never returns to s then $M_s(\omega) = \inf \emptyset = \infty$. If there is a positive probability to never come back to s, i.e. $P_s(M_s = \infty) > 0$ then the state s is called ***transient***. Otherwise, if $P_s(M_s < \infty) = 1$ then s is ***recurrent***. We denote S^t as the set of transient states and S^r as the set of recurrent states. A state s' is ***reachable*** from s (denoted by $s \to s'$), if there exists $n \in \mathbb{N}$ with $P_s(X_n = s') > 0$. The notion of reachability induces the (communcation) equivalence relation

$$s \leftrightarrow s' \quad \Leftrightarrow \quad s \to s' \text{ and } s' \to s.$$

This relation further partitions the set of recurrent states S^r into the equivalence classes S_i^r, $i = 1, \ldots, k$ such that the whole state space S can be written as the disjoint union $S = \bigcup_{i=1}^k S_i^r \cup S^t$. Each of these equivalence classes S_i^r is called a ***closed recurrent class***, since (by the communication relation) for every $s \in S_i^r$ there are no transitions out of this class, i.e. $P(s, s') = 0$ for all $s' \in S \backslash S_i^r$. For this reason each S_i^r is a minimal closed subset of S, i.e. there is no proper nonempty subset of S_i^r which is closed. In case a closed recurrent class consists only of one state s, then s is called ***absorbing***. The DTMC $\mathcal{M}$ is ***unichain*** if there is only one recurrent class ($k = 1$) and if in addition $S^t = \emptyset$ then $\mathcal{M}$ is ***irreducible***. A DTMC that is not unichain will be called ***multichain***. The queueing system in Example 2.1 is irreducible, since every state is reachable from every other state. For discrete-time Markov chains there are some peculiarities regarding the long-run behavior of the Markov chain as $n \to \infty$. If $X_0 = s$ is the initial state and the limit $\rho_s(s') := \lim_{n\to\infty} P_s(X_n = s')$ exists for all s', then $\rho_s \in \mathcal{D}(S)$ is called the ***limiting distribution*** from s. In general this limit does not need to exist, since the sequence $P_s(X_n = s') = P^n(s, s')$ might have oscillations between distinct accumulation points. This fact is related to the periodicity of a state: From state s the points in time of possible returns to s are given by the set $R_s := \{n \geq 1 \mid P_s(X_n = s) > 0\}$. If all $n \in R_s$ are multiples of some natural number $d \geq 2$ (i.e. $R_s \subseteq \{kd \mid k \in \mathbb{N}\}$) then the state s is called ***periodic*** and the ***periodicity*** of s is the largest such integer d. Otherwise s is called ***aperiodic*** and the periodicity of s is set to 1. The periodicity is a class property and means that for every closed recurrent class S_i^r and for all $s, s' \in S_i^r$ the periodicity of s and s' is the same. A Markov chain which is irreducible and aperiodic is often called ***ergodic*** in the literature. As an example, a two-state Markov chain with transition probability matrix

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

is irreducible but not ergodic, since it is periodic with periodicity 2.

One can show that a recurrent state s in a DTMC with finite state space is aperiodic if and only if for all $s' \in S$ the sequence $P^n(s, s')$ converges. Therefore, the limiting distribution ρ_s exists (for s recurrent) if and only if s is aperiodic. In this case $\rho_s(s') = \sum_t \rho_s(t) P(t, s')$ for all s', which is written in vector notation by $\rho_s = \rho_s P$. This equation is often interpreted as the invariance

(or stationarity) condition: If $\rho_s(s')$ is the probability to find the system in state s' at some point in time, then this probability remains unchanged after the system performs a transition. In general, there can be several distributions $\rho \in \mathcal{D}(S)$ with the invariance property $\rho = \rho P$ and any such distribution ρ is called a ***stationary distribution***. It holds that the set of all stationary distributions forms a simplex in $\mathbb{R}^S$ and the number of vertices of this simplex is exactly the number of recurrent classes k. Therefore, in a unichain model $\mathcal{M}$ there is only one stationary distribution ρ and if $\mathcal{M}$ is irreducible then $\rho(s) > 0$ for all $s \in S$. Since a limiting distribution is stationary it further holds that if $\mathcal{M}$ is unichain and aperiodic (or even ergodic), then for all initial states s the limiting distribution ρ_s exists and $\rho_s = \rho$ is the unique stationary distribution and thus independent of s.

We will draw on the stationary distributions (and also the periodic behavior) of a Markov chain in Sect. 2.4, where we will outline the average reward analysis. In order to make the intuition on the average reward clear, we will also work with the splitting $S = \bigcup_{i=1}^{k} S_i^r \cup S^t$ into closed recurrent classes and transient states. The computation of this splitting can be performed by the Fox-Landi state classification algorithm [15]. It finds a representation φ of S (see (2.1)) such that P can be written as

$$P = \begin{pmatrix} P_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & P_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & 0 & \dots & P_k & 0 \\ \widetilde{P}_1 & \widetilde{P}_2 & \widetilde{P}_3 & \dots & \widetilde{P}_k & \widetilde{P}_{k+1} \end{pmatrix} \tag{2.5}$$

where $P_i \in \mathbb{R}^{r_i \times r_i}$, $\widetilde{P}_i \in \mathbb{R}^{t \times r_i}$, $\widetilde{P}_{k+1} \in \mathbb{R}^{t \times t}$ with $r_i := |S_i^r|$ for $i = 1, \dots, k$ and $t := |S^t|$. The matrix P_i represents the transition probabilities within the i-th recurrent class S_i^r and $\widetilde{P}_i$ the transition probabilities from transient states S^t into S_i^r if $i = 1, \dots, k$, respectively transitions within S^t for $i = k+1$. Every closed recurrent class S_i^r can be seen as a DTMC $\mathcal{M}^i = (S_i^r, P_i)$ by omitting incoming transitions from transient states. It holds that $\mathcal{M}^i$ is irreducible and thus has a unique stationary distribution $\rho_i \in \mathcal{D}(S^{r_i})$ with $\rho_i(s) > 0$ for all $s \in S^{r_i}$. This distribution ρ_i can be extended to a distribution $\rho_i \in \mathcal{D}(S)$ on S by setting $\rho_i(s) := 0$ for all $s \in S \setminus S_i^r$. Note that ρ_i is also stationary on $\mathcal{M}$. Since transient states S^t of $\mathcal{M}$ are left forever with probability 1 (into the recurrent classes), every stationary distribution $\rho \in \mathcal{D}(S)$ fulfills that $\rho(s) = 0$ for all $s \in S^t$. Thus, an arbitrary stationary distribution $\rho \in \mathcal{D}(S)$ is a convex combination of all the ρ_i, i.e. $\rho(s) = \sum_{i=1}^{k} a_i \rho_i(s)$ with $a_i \geq 0$ and $\sum_{i=1}^{k} a_i = 1$. (This forms the k-dimensional simplex with vertices ρ_i as mentioned above.)

2.1.3 Reward Measure

We now consider a DTMRM $\mathcal{M} = (S, P, R)$ and want to describe in the following sections several ways to accumulate the rewards $R(s, s')$ along paths $\omega = (s_0, s_1, \dots) \in \Omega$. As an example, for a fixed $N \in \mathbb{N} \cup \{\infty\}$ (also called the

horizon length) we can accumulate the rewards for the first N transitions by simple summation: the reward gained for the i-th transition is $R(s_{i-1}, s_i)$ and is summed up to $\sum_{i=1}^{N} R(s_{i-1}, s_i)$, which is regarded as the value of the path ω for the first N transitions. The following definition introduces the notion of a value for state-based models, with which we will be concerned in this tutorial.

Definition 2.4. *Consider a state-based model $\mathcal{M}$ with state space S and a real vector space $\mathcal{V}$. A **reward measure** $\mathcal{R}$ is an evaluation of the model $\mathcal{M}$ that maps $\mathcal{M}$ with an optional set of parameters to the **value** $V \in \mathcal{V}$ of the model. If $\mathcal{V}$ is a vector space of functions over S, i.e. $\mathcal{V} = \mathbb{R}^S = \{V : S \to \mathbb{R}\}$, then a value $V \in \mathcal{V}$ is also called a **value function** of the model.*

Note that we consider in this definition an arbitrary state-based model, which can have discrete time (e.g. DTMRM or MDP, cf. Sect. 3) or continuous time (CTMRM or CTMDP, cf. Sects. 4 and 5). We will mainly consider vector spaces $\mathcal{V}$ which consist of real-valued functions. Beside value functions $V \in \mathbb{R}^S$ which map every state $s \in S$ to a real value $V(s) \in \mathbb{R}$, we will also consider value functions that are time-dependent. For example, if T denotes a set of time values then $\mathcal{V} = \mathbb{R}^{S \times T}$ consists of value functions $V : S \times T \to \mathbb{R}$ such that $V(s,t) \in \mathbb{R}$ is the real value of state $s \in S$ at the point in time $t \in T$. In typical applications T is a discrete set for discrete-time models (e.g. $T = \mathbb{N}$ or $T = \{0, 1, \ldots, N\}$), or T is an interval for continuous-time models (e.g. $T = [0, \infty)$ or $T = [0, T_{\max}]$). The difference between the notion of a reward measure $\mathcal{R}$ and its value function V is that a reward measure can be seen as a measure type which needs additional parameters in order to be able to formally define its value function V. Examples for such parameters are the horizon length N, a discount factor γ (in Sect. 2.3) or a discount rate α (in Sect. 4.5). If clear from the context, we use the notions reward measure and value (function) interchangeably.

2.2 Total Reward Measure

We now define the finite-horizon and infinite-horizon total reward measures which formalize the accumulation procedure along paths by summation as mentioned in the motivation of Definition 2.4. The finite-horizon reward measure is used as a basis upon which all the following reward measures will be defined.

Definition 2.5. *Let $\mathcal{M}$ be a DTMRM with state process $(X_n)_{n \in \mathbb{N}}$ and $N < \infty$ a fixed finite horizon length. We define the **finite-horizon total value function** $V_N \colon S \to \mathbb{R}$ by*

$$V_N(s) := \mathbb{E}_s \left[\sum_{i=1}^{N} R(X_{i-1}, X_i) \right]. \tag{2.6}$$

*If for all states s the sequence $\mathbb{E}_s \left[\sum_{i=1}^{N} |R(X_{i-1}, X_i)| \right]$ converges with $N \to \infty$, we define the **(infinite-horizon) total value function** as*

$$V_\infty(s) := \lim_{N \to \infty} V_N(s).$$

In general $V_N(s)$ does not need to converge as $N \to \infty$. For example, if all rewards for every recurrent state are strictly positive, then accumulation of positive values diverges to ∞. Even worse, if the rewards have different signs then their accumulation can also oscillate. In order not to be concerned with such oscillations, we impose as a stronger condition the absolute convergence for the infinite-horizon case as in the definition.

As next we want to provide a method which helps to evaluate the total reward measure for the finite and infinite horizon cases. The proof of the following theorem can be found in the Appendix (page 234).

Theorem 2.1 (Evaluation of the Total Reward Measure).

(i) The finite-horizon total value $V_N(s)$ can be computed iteratively through

$$V_N(s) = R(s) + \sum_{s' \in S} P(s, s') V_{N-1}(s'),$$

where $V_0(s) := 0$ for all $s \in S$.

(ii) If the infinite-horizon total value $V_\infty(s)$ exists, then it solves the system of linear equations

$$V_\infty(s) = R(s) + \sum_{s' \in S} P(s, s') V_\infty(s'). \tag{2.7}$$

We formulate the evaluation of the total value function in vector notation:

$$V_N = R + PV_{N-1} = \sum_{i=1}^{N} P^{i-1} R. \tag{2.8}$$

For the infinite-horizon total value function it holds

$$V_\infty = R + PV_\infty \quad \text{respectively} \quad (I - P)V_\infty = R. \tag{2.9}$$

Note that Theorem 2.1 states that if V_∞ exists, then it solves (2.9). On the other hand the system of equations $(I-P)X = R$ with the variable X may have several solutions, since P is stochastic and thus the rank of $I-P$ is not full. The next proposition shows a necessary and sufficient condition for the existence of V_∞ in terms of the reward function R. Furthermore, it follows that if V_∞ exists then $V_\infty(s) = 0$ on all recurrent states s and V_∞ is also the unique solution to $(I - P)X = R$ with the property that $X(s) = 0$ for all recurrent states s. A proof (for aperiodic Markov chains) can be found in the Appendix on page 235.

Proposition 2.1. *For a DTMRM (S, P, R) let $S = \bigcup_{i=1}^{k} S_i^r \cup S^t$ be the partitioning of S into k closed recurrent classes S_i^r and transient states S^t. The infinite-horizon total value function V_∞ exists if and only if for all $i = 1, \ldots, k$ and for all $s, s' \in S_i^r$ it holds that*

$$R(s, s') = 0.$$

Example 2.2. Let us go back to the queueing model introduced in Example 2.1. The finite-horizon total value function for the first 30 transitions and reward function $R := R_{\text{total}}$ is shown in Fig. 2.2 for the initial state $s_{\text{init}} = (0, 0, \text{idle})$.

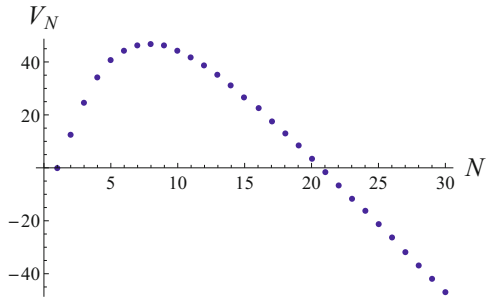

Fig. 2.2. Finite-horizon total value function with horizon length $N = 30$ for the queueing model in Example 2.1 and initial state $s_{\text{init}} = (0, 0, \text{idle})$

As one can see, at the beginning the jobs need some time to fill the system (i.e. both the queue and the server) and thus the expected accumulated reward increases. But after some time steps the high penalty of $C_{\text{loss}} = -\$1000$ for discarding a job outweighs the accumulation of the relatively small reward $R_{\text{acc}} = \$100$ for accomplishing a job and the total value decreases. The infinite-horizon total value does not exist in this model, since $V_N(s_{\text{init}})$ diverges to $-\infty$. However, in case the total reward up to the first loss of a job is of interest, one can introduce an auxiliary absorbing state loss with reward 0, which represents that an incoming job has been discarded (Fig. 2.3).

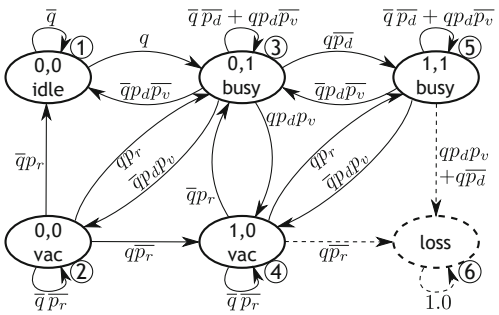

Fig. 2.3. Queueing model enhanced with an auxiliary absorbing state 'loss' representing the loss of an incoming job due to a full queue

Since the single recurrent state loss in this DTMRM has reward 0, the total value function exists and fulfills (2.7) (respectively (2.9)) with $R(s) := R_{\text{profit}}(s)$ for

$s \neq \text{loss}$ (see (2.2)) and $R(\text{loss}) := 0$. Note that $I - P$ has rank 5 since P defines an absorbing state. From $R(\text{loss}) = 0$ it follows that the total value function V_∞ is the unique solution for (2.7) with the constraint $V_\infty(\text{loss}) = 0$ and is given by

$$V_\infty \approx (1221.95, 980.892, 1221.95, 659.481, 950.514, 0)^T . \qquad \square$$

2.3 Horizon-Expected and Discounted Reward Measure

In order to be able to evaluate and compare the performance of systems in which the total value does not exist, we need other appropriate reward measures. In this and the following subsection, we will present two other typically used reward measures: the discounted and the average reward measure. Roughly speaking, the average reward measures the derivation of the total value with respect to the horizon length N, i.e. its average growth. The discounted measure can be used if the horizon length for the system is finite but a priori unknown and can be assumed as being random (and memoryless). In order to define the discounted reward measure, we first introduce the more general horizon-expected reward measure.

Definition 2.6. *Let $\mathcal{M} = (S, P, R)$ be a DTMRM and consider a random horizon length N for $\mathcal{M}$, i.e. N is a random variable over $\mathbb{N}$ that is independent of the state process X_n of $\mathcal{M}$. Let $V_{(N)}$ denote the random finite-horizon total value function that takes values in $\left\{V_n \in \mathbb{R}^S \mid n \in \mathbb{N}\right\}$. Define the* ***horizon-expected value function*** *by*

$$V(s) := \mathbb{E}\left[V_{(N)}(s)\right],$$

if the expectation exists for all $s \in S$, i.e. $|V_{(N)}(s)|$ has finite expectation.

In order to be formally correct, the random variable $V_{(N)}(s)$ is the conditional expectation $V_{(N)}(s) = \mathbb{E}_s\left[\sum_{i=1}^{N} R(X_{i-1}, X_i) \mid N\right]$ and thus if $N = n$ then $V_{(N)}(s)$ takes the value $\mathbb{E}_s\left[\sum_{i=1}^{N} R(X_{i-1}, X_i) \mid N = n\right] = \mathbb{E}_s\left[\sum_{i=1}^{n} R(X_{i-1}, X_i)\right] = V_n(s)$. By the law of total expectation it follows that

$$V(s) = \mathbb{E}\left[\mathbb{E}_s\left[\sum_{i=1}^{N} R(X_{i-1}, X_i) \mid N\right]\right] = \mathbb{E}_s\left[\sum_{i=1}^{N} R(X_{i-1}, X_i)\right],$$

i.e. $V(s)$ is a joint expectation with respect to the product of the probability measures of N and all the X_i.

The following lemma presents a natural sufficient condition that ensures the existence of the horizon-expected value function.

Lemma 2.1. *If the random horizon length N has finite expectation $\mathbb{E}[N] < \infty$ then $V(s)$ exists.*

Proof. Since the state space is finite there exists $C \in \mathbb{R}$ such that $|R(s,s')| \leq C$ $\forall s, s' \in S$. Therefore

$$|V_n(s)| \leq \mathbb{E}_s \left[\sum_{i=1}^{n} |R(X_{i-1}, X_i)| \right] \leq n \cdot C$$

and thus

$$\mathbb{E}\left[|V_{(N)}(s)|\right] = \sum_{n=0}^{\infty} |V_n(s)| \cdot P(N=n) \leq C \cdot \mathbb{E}[N] < \infty. \qquad \square$$

In many applications the horizon length is considered to be memoryless, i.e. $P(N > n+m \mid N > m) = P(N > n)$ and is therefore geometrically distributed. This fact motivates the following definition.

Definition 2.7. *For $\gamma \in (0,1)$ let N be geometrically distributed with parameter $1-\gamma$, i.e. $P(N=n) = \gamma^{n-1}(1-\gamma)$ for $n = 1, 2, \ldots$. In this case the horizon-expected value function is called* ***discounted value function*** *with* ***discount factor*** *γ (or just γ-discounted value function) and is denoted by V^γ.*

As for the total value function in Sect. 2.2 we can explicitly compute the discounted value function:

Theorem 2.2 (Evaluation of the Discounted Reward Measure). *For a discount factor $\gamma \in (0,1)$ it holds that*

$$V^\gamma(s) = \lim_{n \to \infty} \mathbb{E}_s \left[\sum_{i=1}^{n} \gamma^{i-1} R(X_{i-1}, X_i) \right]. \tag{2.10}$$

Furthermore, V^γ is the unique solution to the system of linear equations

$$V^\gamma(s) = R(s) + \gamma \sum_{s' \in S} P(s,s') V^\gamma(s') \tag{2.11}$$

which is written in vector notation as

$$(I - \gamma P) V^\gamma = R.$$

Proof. Let N be geometrically distributed with parameter γ. By conditional expectation we get

$$\begin{aligned} V^\gamma(s) &= \mathbb{E}\left[\mathbb{E}_s \left[\sum_{i=1}^{N} R(X_{i-1}, X_i) \mid N \right] \right] = \sum_{n=1}^{\infty} \mathbb{E}_s \left[\sum_{i=1}^{n} R(X_{i-1}, X_i) \right] P(N=n) \\ &= (1-\gamma) \sum_{i=1}^{\infty} \mathbb{E}_s \left[R(X_{i-1}, X_i) \right] \sum_{n=i}^{\infty} \gamma^{n-1} = \sum_{i=1}^{\infty} \mathbb{E}_s \left[R(X_{i-1}, X_i) \right] \gamma^{i-1}, \end{aligned}$$

which gives (2.10). The derivation of the linear equations (2.11) is completely analogous to the total case by comparing (2.10) to (2.6) (see proof of Theorem 2.1). Since $I - \gamma P$ has full rank for $\gamma \in (0,1)$ the solution is unique. $\square$

Equation (2.10) yields also another characterization of the discounted reward measure. Along a path $\omega = (s_0, s_1, s_2, \dots)$ the accumulated discounted reward is $R(s_0, s_1) + \gamma R(s_1, s_2) + \gamma^2 R(s_2, s_3) + \dots$. The future rewards $R(s_i, s_{i+1})$ are reduced by the factor $\gamma^i < 1$ in order to express some kind of uncertainty about the future reward value (e.g. induced by inflation). In many applications, the discounted reward measure is used with a high discount factor close to 1 which still avoids possible divergence of the infinite-horizon total value. Qualitatively speaking, for high $\gamma < 1$ the sequence γ^i decreases for the first few points in time i slowly, but exponentially to 0. If we assume that the rewards $R(s)$ are close to each other for each state s, then the rewards accumulated within the first few time steps approximately give the discounted value.

Remark 2.1. Note that the discounted value function can be equivalently characterized as an infinite-horizon total value function by adding an absorbing and reward-free final state *abs* to the state space S such that *abs* is reachable from any other state with probability $1 - \gamma$ and any other transition probability is multiplied with γ (see Fig. 2.4). Since *abs* is eventually reached on every path within a finite number of transitions with probability 1 and has reward 0, it characterizes the end of the accumulation procedure. The extended transition probability matrix $P' \in \mathbb{R}^{(|S|+1)\times(|S|+1)}$ and the reward vector $R' \in \mathbb{R}^{|S|+1}$ are given by

$$P' = \begin{pmatrix} \gamma P & (1-\gamma)\mathbf{1} \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad R' = \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where $\mathbf{1} = (1, \dots, 1)^T$. Since *abs* is the single recurrent state and $R(abs) = 0$ it follows by Proposition 2.1 that V_∞ exists. Furthermore, it satisfies $(I - P')V_\infty = R'$ and because $V_\infty(abs) = 0$ it is also the unique solution with this property. On the other hand $(V^\gamma, 0)^T$ is also a solution and thus $V_\infty = (V^\gamma, 0)^T$.

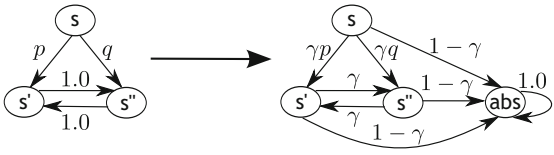

Fig. 2.4. Equivalent characterization of the γ-discounted reward measure as a total reward measure by extending the state space with an absorbing reward-free state

Example 2.3. As an example we analyze the queueing model from Example 2.1 with respect to the discounted reward measure. Figure 2.5 shows $V^\gamma(s)$ for the inital state $s_{\text{init}} = (0, 0, \text{idle})$ as a function of γ. As we see, for small values of γ the discounted values are positive, since the expected horizon length $\frac{1}{1-\gamma}$ is also small and thus incoming jobs have a chance to be processed and not discarded within that time. However for γ approaching 1, the expected horizon length gets larger and the accumulation of the negative reward $C_{\text{loss}} = -\$1000$ for discarding jobs in a full queue prevails. □

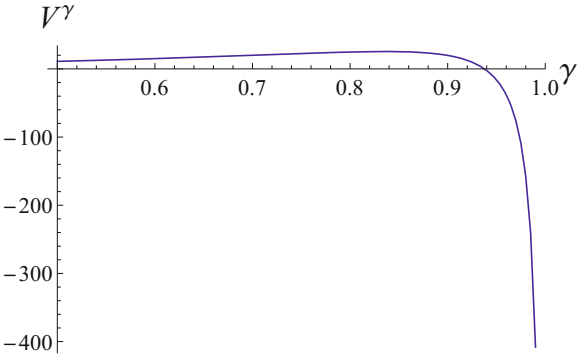

Fig. 2.5. Discounted reward $V^\gamma(s_{\text{init}})$ as a function of γ for the queueing model in Example 2.1 and initial state $s_{\text{init}} = (0, 0, \text{idle})$

2.4 Average Reward Measure

We now provide another important reward measure for the case that the horizon length is infinite (and not random as assumed in Sect. 2.3). We assume for this section that the reader is familiar with the concept of periodicity as presented in Sect. 2.1.2. If for a DTMRM $\mathcal{M} = (S, P, R)$ the infinite-horizon total value function V_∞ does not exist, then either $V_N(s)$ diverges to $\pm\infty$ for some state s or $V_N(s)$ is oscillating over time. The problem with this measure is that V_∞ infinitely often collects the rewards and sums them all up. Instead of building such a total accumulation, one can also measure the system by considering the gained rewards only per time step. As an example, consider an ergodic model $\mathcal{M}$ in the long run with limiting distribution $\rho := \rho_s$ given by $\rho_s(s') := \lim_{n\to\infty} P^n(s, s')$ for an arbitrary intial state s. Then in steady-state the system is rewarded at each time step with $\rho R = \sum_s \rho(s)R(s) \in \mathbb{R}$, i.e. an average of all the rewards $R(s)$ weighted by the probability $\rho(s)$ that the system occupies state s in steady-state. But averaging the rewards $R(s)$ shall not be restricted to only those models $\mathcal{M}$ for which a limiting distribution exists. First of all, the limiting distribution ρ_s can depend on the initial state s, if $\mathcal{M}$ has several closed recurrent classes. More important, the limiting distribution might even not exist, as one can see for the periodic model with

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

However, in this case one would also expect an average reward of 3 for each time step and for every state, since P is irreducible and has the unique stationary distribution $\rho = (0.5, 0.5)$. This means that the average reward measure shall also be applicable to models for which the limiting distribution does not exist. Instead of computing an average per time step in steady-state, one can also think of calculating an average in the long-run by accumulating for each horizon length N the total value V_N and dividing it by the time N. The limit of the

sequence of these finite horizon averages establishes the desired long-run average. As we will see in Proposition 2.2, this long-run average always converges, independent of the structure of the underyling DTMRM. Furthermore, in case the limiting distribution exists then the steady-state average and the long-run average coincide and thus the long-run average fulfills the desired requirements from the motivation.

Definition 2.8. *The **average reward measure** (or **gain**) of a DTMRM with value function $g(s)$ is defined by*

$$g(s) := \lim_{N\to\infty} \frac{1}{N} V_N(s)$$

if the limit exists.

In the following, we summarize well-known results from linear algebra which first of all directly imply that the average reward exists (at least for finite state spaces) and furthermore also allow us to provide methods for its evaluation.

Definition 2.9. *For a stochastic matrix P define the **limiting matrices** P^∞ and P^* as:*

$$P^\infty := \lim_{N\to\infty} P^N \quad \text{and} \quad P^* := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} P^{i-1},$$

if the limits exist. (P^* is the Cesàro limit of the sequence P^i.)

Suppose that P^* exists. Then the average reward can be computed by

$$g = P^* R, \tag{2.12}$$

since by (2.8) it holds that

$$g(s) = \lim_{N\to\infty} \frac{1}{N} V_N(s) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left(P^{i-1} R\right)(s) = \left(P^* R\right)(s).$$

Note also that if P^∞ exists, then the i-th row in P^∞ (with $i = \varphi(s)$, see (2.1)) represents the limiting distribution ρ_i of the model, given that the initial state of the system is s. By the motivation from above it should also hold that $g(s) = \rho_i R$. The following proposition relates these two quantities to each other. We refer for proof to [33].

Proposition 2.2. *Consider a DTMC $\mathcal{M} = (S, P)$ with finite state space.*

(i) The limiting matrix P^* exists.
(ii) If P^∞ exists, then $P^* = P^\infty$.
(iii) If P is aperiodic then P^∞ exists and if in addition P is unichain (or ergodic) with limiting distribution ρ then $\rho P = \rho$ and P^∞ has identical rows ρ, i.e.

$$P^* = P^\infty = \mathbf{1}\rho,$$

where $\mathbf{1} = (1, \ldots, 1)^T$ is the column vector consisting of all ones.

From Proposition 2.2(ii) it follows that the definition of the average reward corresponds to its motivation from above in the case that the limiting distribution ρ_i exists for all i. However, in the case of periodic DTMRMs (when the limiting distribution is not available), Proposition 2.2(i) ensures that at least P^* exists and due to (2.12) this is sufficient for the computation of the average reward.

Remark 2.2. The limiting matrix P^* satisfies the equalities

$$PP^* = P^*P = P^*P^* = P^*. \tag{2.13}$$

P^* can be computed by partitioning the state space $S = \cup_{i=1}^k S_i^r \cup S^t$ into closed recurrent classes S_i^r and transient states S^t which results in a representation of P as in (2.5). Let the row vector $\rho_i \in \mathbb{R}^{r_i}$ denote the unique stationary distribution of P_i, i.e. $\rho_i P_i = \rho_i$. Then

$$P^* = \begin{pmatrix} P_1^* & 0 & 0 & \dots & 0 & 0 \\ 0 & P_2^* & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & 0 & \dots & P_k^* & 0 \\ \widetilde{P}_1^* & \widetilde{P}_2^* & \widetilde{P}_3^* & \dots & \widetilde{P}_k^* & 0 \end{pmatrix} \tag{2.14}$$

where $P_i^* = \mathbf{1}\rho_i$ has identical rows $\rho_i \in \mathbb{R}^{r_i}$ and $\widetilde{P}_i^* = (I - \widetilde{P}_{k+1})^{-1}\widetilde{P}_i P_i^*$ consists of trapping probabilities from transient states into the i-th recurrent class. It follows that the average reward g is constant on each recurrent class, i.e. $g(s) = g_i := \rho_i R_i$ for all $s \in S_i^r$ where $R_i \in \mathbb{R}^{r_i}$ is the vector of rewards on S_i^r. On transient states the average reward g is a weighted sum of all the g_i with weights given by the trapping probabilities.

We want to provide another method to evaluate the average reward measure, because it will be useful for the section on MDPs. This method relies on the key aspect of a Laurent series decomposition which also links together the three proposed measures total reward, discounted reward and average reward. Consider the discounted value $V^\gamma = (I - \gamma P)^{-1}R$ as a function of γ (cf. Theorem 2.2). If the total value function $V_\infty = (I - P)^{-1}R$ exists then V^γ converges to V_∞ as $\gamma \nearrow 1$. But what happens if V_∞ diverges to ∞ or $-\infty$? In this case V^γ has a pole singularity at $\gamma = 1$ and can be expanded into a Laurent series. Roughly speaking, a Laurent series generalizes the concept of a power series for (differentiable) functions f with poles, i.e. points c at the boundary of the domain of f with $\lim_{x\to c} f(x) = \pm\infty$. In such a case, f can be expanded in some neighborhood of c into a function of the form $\sum_{n=-N}^{\infty} a_n(x-c)^n$ for some $N \in \mathbb{N}$ and $a_n \in \mathbb{R}$, which is a sum of a rational function and a power series. In our case, since $\gamma \mapsto V^\gamma$ might have a pole at $\gamma = 1$, the Laurent series is of the form $V^\gamma = \sum_{n=-N}^{\infty} a_n(\gamma - 1)^n$ for γ close to 1. The coefficients a_n in this expansion are given in Theorem 2.3 in the sequel which can be deduced from the following Lemma. A proof for the lemma can be found in [33].

Lemma 2.2 (Laurent Series Expansion). *For a stochastic matrix P the matrix $(I - P + P^*)$ is invertible. Let*

$$H := (I - P + P^*)^{-1} - P^*.$$

There exists $\delta > 0$ such that for all $0 < \rho < \delta$ the Laurent series of the matrix-valued function $\rho \mapsto (\rho I + (I - P))^{-1}$ is given by

$$(\rho I + (I - P))^{-1} = \rho^{-1} P^* + \sum_{n=0}^{\infty} (-\rho)^n H^{n+1}.$$

Theorem 2.3 (Laurent Series of the Discounted Value Function). *Let $\mathcal{M} = (S, P, R)$ be a DTMRM. For a discount factor $\gamma < 1$ close to 1 write $\gamma(\rho) := \frac{1}{1+\rho}$ where $\rho = \frac{1-\gamma}{\gamma} > 0$ and consider the discounted value $V^{\gamma(\rho)}$ as a function of ρ.*

(i) The Laurent series of $\rho \mapsto V^{\gamma(\rho)}$ at 0 is given (for small $\rho > 0$) by

$$V^{\gamma(\rho)} = (1 + \rho) \left(\rho^{-1} g + \sum_{n=0}^{\infty} (-\rho)^n H^{n+1} R \right), \tag{2.15}$$

where g is the average reward.

(ii) It holds that

$$V^{\gamma} = \frac{1}{1-\gamma} g + h + f(\gamma) \tag{2.16}$$

where $h := HR$ and f is some function with $\lim_{\gamma \nearrow 1} f(\gamma) = 0$. Furthermore,

$$g = \lim_{\gamma \nearrow 1} (1 - \gamma) V^{\gamma}.$$

Proof. (i) We apply the Laurent series from Lemma 2.2 as follows:

$$\begin{aligned} V^{\gamma} &= (I - \gamma P)^{-1} R = (1 + \rho)(\rho I + (I - P))^{-1} R \\ &= (1 + \rho) \left(\rho^{-1} P^* R + \sum_{n=0}^{\infty} (-\rho)^n H^{n+1} R \right) \end{aligned}$$

and the claim follows from (2.12). Furthermore, by substituting $\gamma = (1 + \rho)^{-1}$

$$V^{\gamma} = \frac{1}{\gamma} \left(\frac{\gamma}{1-\gamma} g + h + \sum_{n=1}^{\infty} \left(\frac{\gamma - 1}{\gamma} \right)^n H^{n+1} R \right) = \frac{1}{1-\gamma} g + h + f(\gamma)$$

where $f(\gamma) := \dfrac{1-\gamma}{\gamma} h + \displaystyle\sum_{n=1}^{\infty} \left(\frac{\gamma - 1}{\gamma} \right)^n H^{n+1} R$ and $f(\gamma) \to 0$ when $\gamma \nearrow 1$ such that (ii) follows. □

The vector h in (2.16) is called the ***bias*** for the DTMRM. We provide an equivalent characterization for h that allows a simpler interpretation of the term "bias" as some sort of deviation. If the reward function R is replaced by the average reward g, such that in every state s the average reward $g(s)$ is gained instead of $R(s)$, then the finite horizon total reward is given by $G_N(s) := \mathbb{E}_s \left[\sum_{i=1}^{N} g(X_i) \right]$,

where X_i is the state process. In this case $\Delta_N := V_N - G_N$ describes the deviation in accumulation between the specified reward R and its corresponding average reward $g = P^*R$ within N steps. By (2.8) it holds that

$$\Delta_N = \sum_{n=0}^{N-1} P^n(R-g) = \sum_{n=0}^{N-1}(P^n - P^*)R.$$

Note that $P^n - P^* = (P-P^*)^n$ for all $n \geq 1$ which follows from (2.13) and $\sum_{k=0}^{n}(-1)^k\binom{n}{k} = 0$ applied on $(P-P^*)^n = \sum_{k=0}^{n}\binom{n}{k}P^{n-k}(-P^*)^k$. If we assume that Δ_N converges for any reward function R then $\sum_{n=0}^{\infty}(P-P^*)^n$ converges and it holds that $\sum_{n=0}^{\infty}(P-P^*)^n = (I-(P-P^*))^{-1}$. It follows

$$\lim_{N\to\infty}\Delta_N = \sum_{n=0}^{\infty}(P^n - P^*)R = \left((I-P^*) + \sum_{n=1}^{\infty}(P-P^*)^n\right)R =$$
$$\left(\sum_{n=0}^{\infty}(P-P^*)^n - P^*\right)R = \left((I-(P-P^*))^{-1} - P^*\right)R = HR = h.$$

Therefore, the bias $h(s)$ is exactly the long-term deviation between $V_N(s)$ and $G_N(s)$ as $N \to \infty$. This means that the value $h(s)$ is the excess in the accumulation of rewards beginning in state s until the system reaches its steady-state. Remember that g is constant on recurrent classes. Thus, for a recurrent state s it holds that $G_N(s) = \mathbb{E}_s\left[\sum_{i=1}^{N} g(X_i)\right] = g(s)N$ is linear in N and $g(s)N + h(s)$ is a linear asymptote for $V_N(s)$ as $N \to \infty$, i.e. $V_N(s) - (g(s)N + h(s)) \to 0$ (see Fig. 2.6). The matrix H is often called the ***deviation matrix*** for the DTMRM since it maps any reward function R to the corresponding long-term deviation represented by the bias $h = HR$.

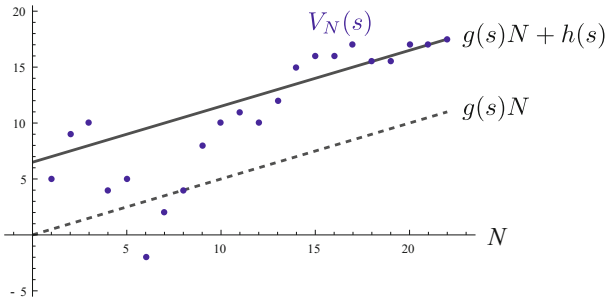

Fig. 2.6. Interpretation of the bias $h(s)$ as the limit of the deviation $V_N(s) - G_N(s)$ as $N \to \infty$. For a recurrent state s it holds that $G_N(s) = g(s)N$.

Another characterization for the bias can be given by considering Δ_N as a finite-horizon total value function for the average-corrected rewards $R - g$, i.e.

$\Delta_N(s) = \mathbb{E}_s\left[\sum_{i=1}^{N}(R(X_i) - g(X_i))\right]$. For this reason the bias h is a kind of infinite-horizon total value function for the model $(S, P, R - g)$ [1].

In the above considerations we have assumed that Δ_N converges (for any reward function R), which is guaranteed if the DTMRM is aperiodic [33]. On the other hand, one can also construct simple periodic DTMRMs for which Δ_N is oscillating. For this reason, there is a similar interpretation of the bias h if the periodicity of P is averaged out (by the Cesàro limit). This is related to the distinction between the two limiting matrices P^∞ and P^* from Definition 2.9 and goes beyond the scope of this tutorial. In Sect. 4 we will introduce the average reward for continuous-time models (where periodicity is not a problem) and define the deviation matrix H by a continuous-time analogon of the discrete representation $H = \sum_{n=0}^{\infty}(P^n - P^*)$.

Remark 2.3. In the Laurent series expansion of V^γ respectively $V^{\gamma(\rho)}$ as in (2.15) the vector value $H^{n+1}R$ for $n \geq 0$ is often called the n-bias of V^γ and therefore the bias $h = HR$ is also called the 0-bias. We will see in Sect. 3.4 (and especially Remark 3.6) that these values play an important role in the optimization of MDPs with respect to the average reward and the n-bias measures.

We now provide some methods for computing the average reward based on the bias h of a DTMRM $\mathcal{M} = (S, P, R)$. If $\mathcal{M}$ is ergodic then from Proposition 2.2 it holds that $P^* = \mathbf{1}\rho$ and thus the average reward $g = P^*R = \mathbf{1}(\rho R)$ is constantly ρR for each state. In the general case (e.g. $\mathcal{M}$ is multichain or periodic), the following theorem shows how the bias h can be involved into the computation of g. The proof is based on the following equations that reveal some further connections between P^* and H:

$$P^* = I - (I - P)H \quad \text{and} \quad HP^* = P^*H = 0. \tag{2.17}$$

These equations can be deduced from the defining equation for H in Lemma 2.2 together with (2.13).

Theorem 2.4 (Evaluation of the Average Reward Measure). *The average reward g and the bias h satisfy the following system of linear equations:*

$$\begin{pmatrix} I - P & 0 \\ I & I - P \end{pmatrix} \begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} 0 \\ R \end{pmatrix}. \tag{2.18}$$

Furthermore, a solution (u, v) to this equation implies that $u = P^*R = g$ is the average reward and v differs from the bias h up to some $w \in \ker(I - P)$, i.e. $v - w = h$.

[1] Note that in Definition 2.5 we restricted the existence of the infinite-horizon total value function to an absolute convergence of the finite-horizon total value function Δ_N. By Proposition 2.1 this is equivalent to the fact that the rewards are zero on recurrent states. For the average-corrected model this restriction is in general not satisfied. The reward function $R - g$ can take both positive and negative values on recurrent states which are balanced out by the average reward g such that Δ_N is converging (at least as a Cesàro limit).

Proof. We first show that g and h are solutions to (2.18). From $PP^* = P^*$ it follows that $Pg = PP^*R = P^*R = g$ and from (2.17) we have

$$(I - P)h = (I - P)HR = (I - P^*)R = R - g$$

and thus (g, h) is a solution to (2.18). Now for an arbitrary solution (u, v) to (2.18) it follows that

$$(I-P+P^*)u = (I-P)u+P^*u+P^*(I-P)v = (I-P)u+P^*(u+(I-P)v) = 0+P^*R.$$

Since $I - P + P^*$ is invertible by Lemma 2.2, we have

$$u = (I - P + P^*)^{-1}P^*R = \left((I - P + P^*)^{-1} - P^* + P^*\right)P^*R = HP^*R + P^*R.$$

From (2.17) it holds that $HP^* = 0$ and thus $u = P^*R = g$. Furthermore, since both h and v fulfill (2.18) it follows that

$$(I - P)v = R - g = (I - P)h$$

and thus $w := v - h \in \ker(I - P)$, such that $h = v - w$. □

From (2.18) it holds that $h = (R - g) + Ph$, which reflects the motivation of the bias h as a total value function for the average-corrected model $(S, P, R - g)$. Furthermore, the theorem shows that the equation $h = (R - u) + Ph$ is only solvable for $u = g$. This means that there is only one choice for u in order to balance out the rewards R such that the finite-horizon total value function Δ_N for the model $(S, P, R - u)$ converges (as a Cesàro limit).

Remark 2.4. Assume that $S = \bigcup_{i=1}^{k} S_i^r \cup S^t$ is the splitting of the state space of $\mathcal{M}$ into closed recurrent classes S_i^r and transient states S^t.

(i) As we saw from (2.12) and (2.14) one can directly compute g from the splitting of S. Equation (2.18) shows that such a splitting is not really necessary. However, performing such a splitting (e.g. by the Fox-Landi algorithm [15]) for the computation of g by P^*R can be more efficient than simply solving (2.18) [33].

(ii) In order to compute g from (2.18) it is enough to compute the bias h up to $\ker(I - P)$. The dimension of $\ker(I - P)$ is exactly the number k of closed recurrent classes S_i^r. Hence, if the splitting of S is known, then $v(s)$ can be set to 0 for some arbitrary chosen state s in each recurrent class S_i^r (leading to a reduction in the number of equations in Theorem 2.4).

(iii) In order to determine the bias h, it is possible to extend (2.18) to

$$\begin{pmatrix} I - P & 0 & 0 \\ I & I - P & 0 \\ 0 & I & I - P \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ R \\ 0 \end{pmatrix}. \tag{2.19}$$

It holds that if (u, v, w) is a solution to (2.19) then $u = g$ and $v = h$ [33]. In a similar manner, one can also establish a system of linear equations in order to compute the n-bias values (see Remark 2.3), i.e. the coefficients in the Laurent series of the discounted value function $V^{\gamma(\rho)}$.

The following corollary drastically simplifies the evaluation of the average reward for the case of unichain models $\mathcal{M} = (S, P, R)$. In this case, the state space can be split to $S = S^r \cup S^t$ and consists of only one closed recurrent class S^r.

Corollary 2.1. *For a unichain model (S, P, R) the average reward g is constant on S. More precisely, $g = g_0\mathbf{1}$ for some $g_0 \in \mathbb{R}$ and in order to compute g_0 one can either solve*

$$g_0\mathbf{1} + (I - P)h = R \tag{2.20}$$

or compute $g_0 = \rho R$, where ρ is the unique stationary distribution of P.

The proof is obvious and we leave it as an exercise to the reader.

Note that (2.20) is a reduced version of (2.18) since $(I-P)g = 0$ for all constant g. Many models in applications are unichain or even ergodic (i.e. irreducible and aperiodic), thus the effort for the evaluation of the average reward is reduced by (2.20). If it is a priori not known if a model is unichain or multichain, then either a model classification algorithm can be applied (e.g. Fox-Landi [15]) or one can directly solve (2.18). In the context of MDPs an analogous classification into unichain and multichain MDPs is applicable. We will see in Sect. 3.4 that Theorem 3.8 describes an optimization algorithm, which builds upon (2.18). In case the MDP is unichain, this optimization algorithm can also be built upon the simpler equation (2.20), thus gaining in efficiency. However, the complexity for the necessary unichain classification is shown to be NP-hard [39]. We refer for more information on classification of MDPs to [19].

Example 2.4. We want to finish this section by showing an example with multichain structure based on the queuing model from Example 2.1. Assume a queue with capacity size k in which jobs are enqueued with probability q and a server with the processing states "idle" and "busy" (with no vacation state). Once again, an accomplished job is rewarded $R_{\text{acc}} = \$100$ and a discarded job costs $C_{\text{loss}} = -\$1000$. Additional to the processing behavior a server can also occupy one of the following modes: "normal", "intense" or "degraded" (see Fig. 2.7).

In normal mode the server accomplishes a job with probability $p_{d,n} = 0.5$. From every normal state the server degrades with probability $r_d = 0.01$. In this degraded mode jobs are accomplished with a lower probability $p_{d,d} = 0.25$. If the system is in normal mode, the queue is full and a job enters the system, then the server moves from the normal mode to the intense mode. This move can only happen, if the system does not degrade (as in state $(1, 1, \text{normal}, \text{busy})$). In the intense mode the processing probability increases to $p_{d,i} = 1.0$ but with the drawback that a job can be served not correctly with probability 0.1. A non-correctly served job behaves as if it would be lost, i.e. the job involves a cost of $C_{\text{loss}} = -\$1000$. Being in intense mode or degraded mode, there is no way to change to any other mode. This means that both the intense mode and the degraded mode represent closed recurrent classes in the state space and thus the model is multichain.

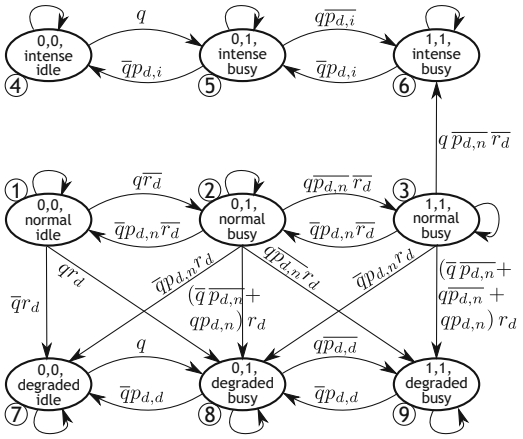

Fig. 2.7. Queueing model with queue capacity $k = 1$ and a service unit with processing behavior idle or busy and processing modes normal, intense or degraded. State $(0, 1, \text{normal}, \text{busy})$ represents 0 jobs in the queue, 1 job served, server is busy and in normal mode. The parameter values are $q = 0.25$, $p_{d,n} = 0.5$, $p_{d,i} = 1.0$ and $p_{d,d} = 0.25$ and $r_d = 0.01$. Probabilities for self-loops complete all outgoing probabilities to 1.0.

By solving (2.18) with P as in Fig. 2.7 and

$$R = \left(\begin{array}{c} 0 \\ R_{\text{acc}}p_{d,n} \\ R_{\text{acc}}p_{d,n} \\ \hline 0 \\ R_{\text{acc}}p_{d,i} \cdot 0.9 + C_{\text{loss}}p_{d,i} \cdot 0.1 \\ R_{\text{acc}}p_{d,i} \cdot 0.9 + C_{\text{loss}}p_{d,i} \cdot 0.1 \\ \hline 0 \\ R_{\text{acc}}p_{d,d} \\ R_{\text{acc}}p_{d,d} \end{array}\right) + \left(\begin{array}{c} 0 \\ 0 \\ C_{\text{loss}}q\overline{p_{d,n}} \\ \hline 0 \\ 0 \\ C_{\text{loss}}q\overline{p_{d,i}} \\ \hline 0 \\ 0 \\ C_{\text{loss}}q\overline{p_{d,d}} \end{array}\right) = \left(\begin{array}{c} 0 \\ 50 \\ -75 \\ \hline 0 \\ -10 \\ -10 \\ \hline 0 \\ 25 \\ -162.5 \end{array}\right)$$

the average reward can be computed to

$$g \approx (-25.8, -24.8, -19.8 \,|\, -2.5, -2.5, -2.5 \,|\, -50, -50, -50)^T .$$

As we see the average reward is constant on recurrent classes, i.e. -2.5 in the intense mode and -50 in the degraded mode. For the transient states (represented by the normal mode) the average reward is a state-dependent convex combination of the average rewards for the recurrent classes, since the probability to leave the normal mode to one of the recurrent classes depends on the particular transient state. The bias h can be determined from (2.19) to

$$h \approx (1853.3, 1811.3, 1213.9 \,|\, 2.5, -7.5, -17.5 \,|\, 363.6, 163.6, -436.364) .$$

This means that if $(0, 0, \text{normal}, \text{idle})$ is the initial state then the reward accumulation process of the finite-horizon total value function V_N follows the linear function $-25.8 \cdot N + 1853.3$ asymptotically as $N \to \infty$ (see Fig. 2.6). □

3 Markov Decision Processes

This section is devoted to Markov Decision Processes with discrete time. Section 3.1 provides the necessary definitions and terminology, and Sect. 3.2 introduces the discounted reward measure as the first optimization criterion. We present its most important properties and two standard methods (value iteration and policy iteration) for computing the associated optimal value function. Depending on the reader's interest, one or both of the following two subsections may be skipped: Stochastic Shortest Path Problems, the topic of Sect. 3.3, are special MDPs together with the infinite-horizon total reward measure as optimization criterion. The final subsection (Sect. 3.4) addresses the optimization of the average reward measure for MDPs, where we involve the bias into the computation of the average-optimal value function.

3.1 Preliminaries

We first state the formal definition of an MDP and then describe its execution semantics. Roughly speaking, an MDP extends the purely stochastic behavior of a DTMRM by introducing actions, which can be used in order to control state transitions.

Definition 3.1. *A **discrete-time Markov Decision Process (MDP)** is a structure $\mathcal{M} = (S, Act, e, P, R)$, where S is the finite state space, $Act \neq \emptyset$ a finite set of actions, $e\colon S \to 2^{Act} \setminus \emptyset$ the action-enabling function, $P\colon S \times Act \to \mathcal{D}(S)$ an action-dependent transition function and $R\colon S \times Act \times S \to \mathbb{R}$ the action-dependent reward function. We denote $P(s,a,s') := (P(s,a))(s')$.*

From a state $s \in S$ an enabled action $a \in e(s)$ must be chosen which induces a probability distribution $P(s,a)$ over S to target states. If a transition to s' takes place then a reward $R(s,a,s')$ is gained and the process continues in s'. In analogy to DTMRMs we denote $R(s,a) := \sum_{s' \in S} P(s,a,s')R(s,a,s')$ as the expected reward that is gained when action a has been chosen and a transition from state s is performed. The mechanism which chooses an action in every state is called a policy. In the theory of MDPs there are several possibilities to define policies. In this tutorial, we restrict to the simplest type of policy.

Definition 3.2. *A **policy** is a function $\pi\colon S \to Act$ with $\pi(s) \in e(s)$ for all $s \in S$. Define $\Pi \subseteq Act^S$ as the set of all policies.*

A policy π of an MDP $\mathcal{M}$ resolves the non-deterministic choice between actions and thus reduces $\mathcal{M}$ into a DTMRM $\mathcal{M}^\pi := (S, P^\pi, R^\pi)$, where

$$P^\pi(s,s') := P(s,\pi(s),s') \quad \text{and} \quad R^\pi(s,s') := R(s,\pi(s),s').$$

Remark 3.1. In the literature one often finds more general definitions of policies in which the choice of an action a in state s does not only depend on the current state s but also

– on the history of both the state process and the previously chosen actions and
– can be randomized, i.e. the policy prescribes for each state s a probability distribution $\pi(s) \in \mathcal{D}(e(s))$ over all enabled actions and action a is chosen with probability $(\pi(s))(a)$.

The policy type as in Definition 3.2 is often referred to "stationary Markovian deterministic". Here, deterministic is in contrast to randomized and means that the policy assigns a fixed action instead of some probability distribution over actions. A policy is Markovian, if the choice of the action does not depend on the complete history but only on the current state and point in time of the decision. A Markovian policy is stationary, if it takes the same action a everytime it visits the same state s and is thus also independent of time. For simplicity we stick to the type of stationary Markovian deterministic policies as in Definition 3.2 since this is sufficient for the MDP optimization problems we discuss in this tutorial. The more general types of policies are required if e.g. the target function to be optimized is of a finite-horizon type or if additional constraints for optimization are added to the MDP model (see also Remark 3.3).

Example 3.1 (Queueing model). We consider the queueing model introduced in Example 2.4. Assume that the server can be either idle or busy and operate in normal or intense mode. Figure 3.1 shows an MDP for queue capacity size $k = 2$ and the two actions "keep" and "move", which enable swichting between the processing modes.

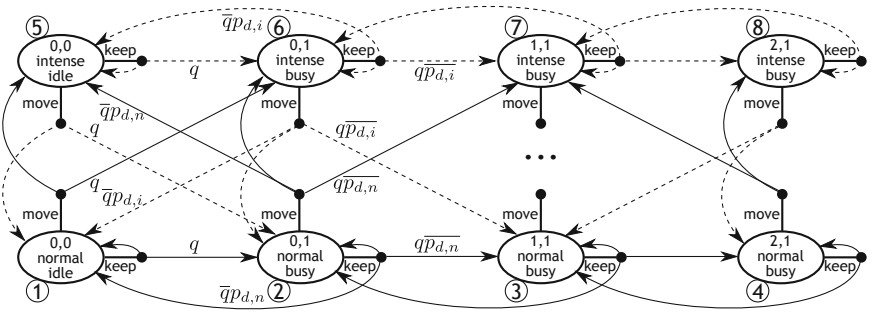

Fig. 3.1. An excerpt of the MDP queueing model with queue capacity $k = 2$, a service unit with processing behavior idle or busy and processing modes normal or intense. In each state the "keep" action keeps the current processing mode, while the "move" action changes the mode. State $(0, 1, \text{normal}, \text{busy})$ represents 0 jobs in the queue, 1 job served, server is busy and in normal mode. The parameter values are $q = 0.25$, $p_{d,n} = 0.5$ and $p_{d,i} = 1.0$. For better overview transitions from normal mode are bold, whereas transitions from intense mode are dashed.

A job enters the system with probability $q = 0.25$. If the queue is full then the system refuses the job, which causes a cost of $C_{\text{loss}} = -\$1000$. A normal operating server accomplishes the job with probability $p_{d,n} = 0.5$, whereas in intense

mode the server succeeds with probability $p_{d,i} = 1.0$. If a job is accomplished the system is rewarded with $R_{\text{acc}} = \$100$. In contrast to the normal mode, in intense mode the system raises a higher operating cost of $C_{\text{int}} = -\$10$ per time step. Furthermore a change from normal to intense mode causes additionally $C_{\text{move}} = -\$50$. All together the rewards can be represented by

$$R^{\text{keep}} = \begin{pmatrix} 0 \\ R_{\text{acc}} p_{d,n} \\ R_{\text{acc}} p_{d,n} \\ R_{\text{acc}} p_{d,n} + C_{\text{loss}} q\overline{p_{d,n}} \\ C_{\text{int}} \\ R_{\text{acc}} p_{d,i} + C_{\text{int}} \\ R_{\text{acc}} p_{d,i} + C_{\text{int}} \\ R_{\text{acc}} p_{d,i} + C_{\text{loss}} q\overline{p_{d,i}} + C_{\text{int}} \end{pmatrix}, \quad R^{\text{move}} = R^{\text{keep}} + \begin{pmatrix} C_{\text{move}} \\ C_{\text{move}} \\ C_{\text{move}} \\ C_{\text{move}} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \quad (3.1)$$

□

3.1.1 Classification of MDPs

The state classification introduced in Sect. 2.1.2 for DTMRMs also implies a classification of MDP models. An MDP $\mathcal{M}$ is called ***unichain***, if for all policies $\pi \in \Pi$ the induced DTMRM $\mathcal{M}^\pi$ is unichain. Otherwise, if there is a policy π, for which $\mathcal{M}^\pi$ has at least two closed recurrent classes, then $\mathcal{M}$ is called ***multichain***. As for DTMRMs, this classification will be mainly used for the analysis and optimization of the average reward in Sect. 3.4. There are also other classification criteria possible, e.g. regarding the reachability under policies [33]. However, in this tutorial, we only need the above described classification with respect to the chain structure of the MDP.

3.1.2 Reward Measure and Optimality

As in Sect. 2.1.3 we choose a reward measure $\mathcal{R}$ (see Definition 2.4) which will be applied to the MDP $\mathcal{M}$. For a policy $\pi \in \Pi$ let $V^\pi \in \mathcal{V}$ denote the value of $\mathcal{R}$ for the induced DTMRM $\mathcal{M}^\pi$ (if it is defined). In this section, we will only work with $\mathcal{V} = \mathbb{R}^S$. This allows us to define a value function $V^* \in \mathcal{V}$ of $\mathcal{R}$ for the complete MDP model $\mathcal{M}$.

Definition 3.3. *Let $\mathcal{M}$ be an MDP with reward measure $\mathcal{R}$. Define the **(optimal) value function** V^* of $\mathcal{M}$ by*

$$V^*(s) := \sup_{\pi \in \Pi} V^\pi(s), \quad (3.2)$$

*if $V^*(s)$ is finite for all s. A policy $\pi^* \in \Pi$ is called **optimal** if V^{π^*} is defined and*

$$\forall s \in S \ \forall \pi \in \Pi : \ V^{\pi^*}(s) \geq V^\pi(s). \quad (3.3)$$

Note that in the definition of V^* the supremum is taken in a state-dependent way. Furthermore, since the state and action spaces are finite, the policy space

Π is finite as well. Therefore, the supremum in the above definition is indeed a maximum. It follows that if for all $\pi \in \Pi$ the value $V^\pi(s)$ is defined and finite then $V^*(s)$ is also finite for all s.

In case $\mathcal{R}$ is the infinite-horizon total reward measure, we allow in contrast to Definition 2.5 the value $V^\pi(s)$ to converge improperly to $-\infty$. Taking the supremum in (3.2) doesn't care of this kind of convergence, if there is at least one policy providing a finite value $V^\pi(s)$. The same holds for (3.3) since here V^{π^*} has to exist in the sense of Definition 2.5 (and thus be finite).

Note further that through the definition of an optimal policy it is not clear if an optimal policy π^* exists, since π^* has to fulfill the inequality in (3.3) uniformly over all states. Definition 3.3 gives rise to the following natural question: Under which conditions does an optimal policy π^* exist and how is it related to the optimal value V^*? These questions will be answered in the following subsections.

3.2 Discounted Reward Measure

We first address the above mentioned questions in the case of the discounted reward measure with discount factor $\gamma \in (0, 1)$, since this measure is analytically simpler to manage than the infinite-horizon or the average reward measure. For motivation, let us first provide some intuition on the optimization problem. Assume we have some value function $V : S \to \mathbb{R}$ and we want to check whether V is optimal or alternatively in what way can we modify V in order to approach the optimal value V^*. When in some state s one has a choice between enabled actions $a \in e(s)$, then for each of these actions one can perform a look-ahead step and compute $Q(s, a) := R(s, a) + \gamma \sum_{s' \in S} P(s, a, s')V(s')$. The value $Q(s, a)$ combines the reward $R(s, a)$ gained for the performed action and the expectation over the values $V(s')$ for the transition to the target state s' induced by action a. If now $Q(s, a') > V(s)$ for some $a' \in e(s)$ then clearly one should improve V by the updated value $V(s) := Q(s, a')$ or even better choose the best improving action and set $V(s) := \max_{a \in e(s)} Q(s, a)$. This update procedure can be formalized by considering the ***Bellman operator*** $\mathcal{T} \colon \mathbb{R}^S \to \mathbb{R}^S$ which assigns to each value function $V \in \mathbb{R}^S$ its update $\mathcal{T}V := \mathcal{T}(V) \in \mathbb{R}^S$ defined by

$$(\mathcal{T}V)(s) := \max_{a \in e(s)} \left\{ R(s, a) + \gamma \sum_{s' \in S} P(s, a, s')V(s') \right\}.$$

Note that $\mathcal{T}$ is a non-linear operator on the vector space $\mathbb{R}^S$, since it involves maximization over actions. If we proceed iteratively, a sequence of improving value functions V is generated and the hope is that this sequence convergences to the optimal value function V^*. In case V is already optimal, there should be no strict improvement anymore possible. This means that for every state s the value $V(s)$ is maximal among all updates $Q(s, a),\ a \in e(s)$ on $V(s)$, i.e.

$$V(s) = \max_{a \in e(s)} \left\{ R(s, a) + \gamma \sum_{s' \in S} P(s, a, s')V(s') \right\}. \tag{3.4}$$

This non-linear fixed-point equation $V = \mathcal{T}V$ is also known as the ***Bellman optimality equation*** and we have to solve it, if we want to detemine V^*. The following Theorem 3.1 establishes results on existence and uniqueness of solutions to this equation. Furthermore, it also creates a connection between the optimal value function V^* and optimal policies π^*.

Theorem 3.1 (Existence Theorem). *Consider an MDP (S, Act, e, P, R) and the discounted reward measure with discount factor $\gamma \in (0, 1)$.*

(i) There exists an optimal value $(V^\gamma)^*$ which is the unique fixed point of $\mathcal{T}$, i.e. the Bellman optimality equation holds:

$$(V^\gamma)^* = \mathcal{T}\,(V^\gamma)^*\,. \tag{3.5}$$

(ii) There exists an optimal policy π^* and it holds that $(V^\gamma)^{\pi^*} = (V^\gamma)^*$.
(iii) Every optimal policy π^* can be derived from the optimal value $(V^\gamma)^*$ by

$$\pi^*(s) \in \operatorname*{argmax}_{a \in e(s)} \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s,a,s')\,(V^\gamma)^*\,(s') \right\}.$$

The complete proof can be found in [33]. The key ingredient for this proof relies on the following lemma, which provides an insight into the analytical properties of the Bellman operator $\mathcal{T}$.

Lemma 3.1. *(i) $\mathcal{T}$ is monotonic, i.e. if $U(s) \le V(s)$ for all $s \in S$ then $(\mathcal{T}U)(s) \le (\mathcal{T}V)(s)$ for all s.*
(ii) $\mathcal{T}$ is a contraction with respect to the maximum norm $||V|| := \max_{s \in S} |V(s)|$, i.e. there exists $q \in \mathbb{R}$ with $0 \le q < 1$ such that

$$||\mathcal{T}U - \mathcal{T}V|| \le q||U - V||.$$

The constant q is called Lipschitz constant and one can choose $q := \gamma$.

Remark 3.2. Lemma 3.1(ii) allows to apply the Banach fixed point theorem on the contraction $\mathcal{T}$ which ensures existence and uniqueness of a fixed point V_{fix}. Furthermore the sequence

$$V_{n+1} := \mathcal{T}V_n \tag{3.6}$$

converges to V_{fix} for an arbitrary initial value function $V_0 \in \mathbb{R}^S$. From the monotonicity property of $\mathcal{T}$ it can be shown that the sequence in (3.6) also converges to the optimal value $(V^\gamma)^*$ and thus $V_{\text{fix}} = (V^\gamma)^*$.

Writing (3.5) in component-wise notation yields exactly the Bellman optimality equation (3.4) as motivated, for which $(V^\gamma)^*$ is the unique solution. Note that $(V^\gamma)^*$ is defined in (3.2) by a maximization over the whole policy space Π, i.e. for each state s all policies $\pi \in \Pi$ have to be considered in order to establish the supremum. In contrast, the Bellman optimality equation reduces this global optimization task into a local state-wise optimization over the enabled actions $a \in e(s)$ for every $s \in S$. Note also that from Theorem 3.1 one can deduce that

$\mathcal{T}((V^\gamma)^{\pi^*}) = (V^\gamma)^{\pi^*}$, such that an optimal policy π^* can be also considered as a "fixed-point" of the Bellman operator $\mathcal{T}$. However, an optimal policy does not need to be unique.

From the Bellman equation (3.4) several algorithms based on fixed-point iteration can be derived which can be used in order to compute the optimal policy together with its value function. The following value iteration algorithm is based on (3.6). Its proof can be found in the Appendix on page 236.

Theorem 3.2 (Value Iteration). *For an arbitrary initial value function $V_0 \in \mathbb{R}^S$ define the sequence of value functions*

$$V_{n+1}(s) := (\mathcal{T}V_n)(s) = \max_{a \in e(s)} \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V_n(s') \right\}.$$

Then V_n converges to $(V^\gamma)^*$. As a termination criterion choose $\varepsilon > 0$ and continue iterating until $||V_{n+1} - V_n|| < \frac{1-\gamma}{2\gamma}\varepsilon$ and let

$$\pi_\varepsilon(s) \in \operatorname*{argmax}_{a \in e(s)} \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V_{n+1}(s') \right\}. \tag{3.7}$$

Then $||V^{\pi_\varepsilon} - (V^\gamma)^*|| < \varepsilon$.

The value iteration algorithm iterates on the vector space $\mathbb{R}^S$ of value functions. From an arbitrary value function an improving policy can be generated by (3.7). In contrast, the following policy iteration algorithm iterates on the policy space Π. From a policy π its value can be generated the other way round by solving a system of linear equations.

Theorem 3.3 (Policy Iteration). *Let $\pi_0 \in \Pi$ be an initial policy. Define the following iteration scheme.*

1. ***Policy evaluation**: Compute the value V^{π_n} of π_n by solving*

$$(I - \gamma P^{\pi_n}) V^{\pi_n} = R^{\pi_n}$$

and define the set of improving actions

$$A_{n+1}(s) := \operatorname*{argmax}_{a \in e(s)} \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V^{\pi_n}(s') \right\}.$$

Termination: If $\pi_n(s) \in A_{n+1}(s)$ for all s then π_n is an optimal policy.

2. ***Policy improvement**: Otherwise choose an improving policy π_{n+1} such that $\pi_{n+1}(s) \in A_{n+1}(s)$ for all $s \in S$.*

The sequence of values V^{π_n} is non-decreasing and policy iteration terminates within a finite number of iterations.

Proof. By definition of π_{n+1} it holds for all s that

$$V^{\pi_{n+1}}(s) = \max_{a \in e(s)} \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V^{\pi_n}(s') \right\}$$
$$\geq R(s, \pi_n(s)) + \gamma \sum_{s' \in S} P(s, \pi_n(s), s') V^{\pi_n}(s') = V^{\pi_n}(s).$$

Since there are only finitely many policies and the values V^{π_n} are non-decreasing, policy iteration terminates in a finite number of iterations. Clearly, if $\pi_n(s) \in A_{n+1}(s)$ for all s then π_n is optimal, since

$$V^{\pi_n}(s) = \max_{a \in e(s)} \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V^{\pi_n}(s') \right\} = \left(\mathcal{T} V^{\pi_n}\right)(s).$$

The conclusion follows by Theorem 3.1. □

Both presented algorithms value iteration and policy iteration create a converging sequence of value functions. For value iteration we mentioned in Remark 3.2 that the generated sequence $V_{n+1} = \mathcal{T} V_n$ converges to the fixed point of $\mathcal{T}$ which is also the global optimal value $(V^\gamma)^*$ of the MDP since $\mathcal{T}$ is monotonic. Same holds for the sequence V^{π_n} in policy iteration, since V^{π_n} is a fixed-point of $\mathcal{T}$ for some $n \in \mathbb{N}$ and thus $V^{\pi_n} = (V^\gamma)^*$. The convergence speed of these algorithms is in general very slow. Value iteration updates in every iteration step the value function V_n on every state. This means especially that states s that in the current iteration step do not contribute to a big improvement $|V_{n+1}(s) - V_n(s)|$ in their value will be completely updated like every other state. However, it can be shown that convergence in value iteration can also be guaranteed, if every state is updated infinitely often [37]. Thus, one could modify the order of updates to states regarding their importance or contribution in value improvement (asynchronous value iteration).

Policy iteration on the other hand computes at every iteration step the exact value V^{π_n} of the current considered policy π_n, by solving a system of linear equations. If π_n is not optimal, then after improvement to π_{n+1} the effort for the accurate computation of V^{π_n} is lost. Therefore, the algorithms value iteration and policy iteration just provide a foundation for potential algorithmic improvements. Examples for such improvements are relative value iteration, modified policy iteration or action elimination [33]. Of course, heuristics which use model-dependent meta-information can also be considered in order to provide a good initial value V_0 or initial policy π_0.

Note that MDP optimization underlies the curse of dimensionality: The explosion of the state space induces an even worse explosion of the policy space since $|\Pi| \in O\left(|Act|^{|S|}\right)$. There is a whole branch of Artificial and Computational Intelligence, which develops learning algorithms and approximation methods for large MDPs (e.g. reinforcement learning [37], evolutionary algorithms [29], heuristics [28] and approximate dynamic programming [8, 10, 26, 27, 32]).

Example 3.2. We now want to optimize the queueing model MDP from Example 3.1 by applying both algorithms value iteration and policy iteration. Table 3.1 shows a comparison between values for the initial state $s_{\text{init}} = (0, 0, \text{normal}, \text{idle})$ under the policies π^{normal} (respectively π^{intense}) which keeps normal (intense) mode or moves to normal (intense) mode and the discount-optimal policy π^* for $\gamma = 0.99$, i.e.

$$\pi^{\text{normal}} = \begin{pmatrix} \text{keep} \\ \text{keep} \\ \text{keep} \\ \text{keep} \\ \hline \text{move} \\ \text{move} \\ \text{move} \\ \text{move} \end{pmatrix}, \quad \pi^{\text{intense}} = \begin{pmatrix} \text{move} \\ \text{move} \\ \text{move} \\ \text{move} \\ \hline \text{keep} \\ \text{keep} \\ \text{keep} \\ \text{keep} \end{pmatrix}, \quad \pi^* = \begin{pmatrix} \text{keep} \\ \text{keep} \\ \text{keep} \\ \text{move} \\ \hline \text{move} \\ \text{move} \\ \text{keep} \\ \text{keep} \end{pmatrix}.$$

Table 3.1. Discounted values with $\gamma = 0.99$ for different policies from initial state $s_{\text{init}} = (0, 0, \text{normal}, \text{idle})$.

policy	$\left(V^{0.99}\right)^{\pi}(s_{\text{init}})$
π^{normal}	1952.36
π^{intense}	1435.00
π^*	2220.95

The optimal policy was computed by policy iteration with initial policy π^{normal} and converged after 3 iterations. In each iteration a system of linear equations with $|S| = 8$ variables had to be solved and $\sum_{s \in S} |e(s)| = 16$ updates (sets of improving actions) computed. With standard algorithms like Gaussian elimination for solving the system of linear equations the worst-case complexity in each iteration is $O(|S|^3 + |Act||S|)$. Note that the number of all available policies is $|\Pi| = |S|^{|Act|} = 256$. Thus policy iteration is a huge gain in efficiency in contrast to the brute-force method, which computes the values V^π for every policy $\pi \in \Pi$ in order to establish the global maximum $V^*(s) = \max_{\pi \in \Pi} V^\pi(s)$.

The value iteration algorithm with initial value function constantly 0, $\varepsilon = 0.1$ and maximum norm $||.||$ converged after 1067 iterations to the value $V_{1067}^{0.99}(s_{\text{init}}) = 2220.90$ and the value-maximizing policy $\pi_\varepsilon = \pi^*$. In each iteration the values $V_n(s)$ for all states s have to be updated, thus the worst-case complexity for one iteration is $O(|Act||S|)$. Also note that value iteration already finds π^* after only 6 iterations with value $V_6^{0.99}(s_{\text{init}}) = 89.08$ – all the other remaining steps just solve the linear equation $V^{0.99} = R^{\pi^*} + 0.99 P^{\pi^*} V^{0.99}$ for π^* by the iterative procedure $V_{n+1}^{0.99} = R^{\pi^*} + 0.99 P^{\pi^*} V_n^{0.99}$. Nevertheless, value iteration in its presented form has to run until it terminates, in order to find a value function which can be guaranteed to be close to the optimal value (distance measured in maximum norm). Furthermore, it is a priori not known at which iteration step the optimization phase stops, i.e. the actions not improvable anymore. □

Remark 3.3. (i) One of the main theorems in Markov Decision Theory states (at least for the models we consider in this tutorial) that if one searches for an optimal value function within the broader space of randomized and history-dependent policies, then an optimal policy is still stationary Markovian deterministic, i.e. lies within Π (see also Remark 3.1). This is the reason why we stick in this introductory tutorial from beginning to the smaller policy space Π.

(ii) An MDP problem can also be transformed to a linear programming problem, such that methods coming from the area of linear optimization can be applied to solve MDPs [33]. In its generality, a linear optimization formulation also allows to add further constraints to the set of linear constraints induced by the Bellman equation. An MDP model with an additional set of linear constraints is also known as a Constrained MDP [1]. It can be shown that in this case stationary Markovian deterministic policies $\pi \in \Pi$ can indeed be outperformed in their value by randomized and history-dependent policies. Also note that history-dependent policies can also be better then stationary Markovian deterministic policies in the finite-horizon case.

3.3 Stochastic Shortest Paths

We now address the problem of optimizing MDPs with respect to the infinite-horizon total reward measure. In contrast to the discounted case, the existence of an optimal value in general can not be guaranteed. In case it exists, it is difficult to provide convergence criteria for dynamic programming algorithms. Therefore, in literature one finds existence results and convergence criteria only for special classes of MDPs with respect to this measure. In the following, we show only one frequently used application for this type of measure.

Definition 3.4. *A **stochastic shortest path problem (SSP)** is an MDP (S, Act, e, P, R) with an absorbing and reward-free state $goal \in S$, i.e. for all policies $\pi \in \Pi$ it holds*

$$P^\pi(goal, goal) = 1 \quad \text{and} \quad R^\pi(goal) = 0.$$

Typically, in SSPs the goal is to minimize costs and not to maximize rewards (see Definition 3.3). Of course, maximization of rewards can be transformed to a minimization of costs, where costs are defined as negative rewards. Note that we allow for rewards (and therefore also for costs) to have both a positive and a negative sign. In order to be consistent with the rest of this tutorial, we stick in the following to the maximization of rewards. For being able to provide results on the existence of optimal solutions, we have to define the notion of a proper policy.

Definition 3.5. *A policy π is called **proper** if there is $m \in \mathbb{N}$ such that under π the goal state can be reached from every state with positive probability within m steps, i.e.*

$$\exists m \in \mathbb{N}\ \forall s \in S:\ (P^\pi)^m(s, goal) > 0.$$

*A policy is called **improper** if it is not proper.*

Define the Bellman operator $\mathcal{T}\colon \mathbb{R}^S \to \mathbb{R}^S$ by

$$(\mathcal{T}V)(s) := \max_{a \in e(s)} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s')V(s') \right\}.$$

In analogy to the discounted case it holds that $\mathcal{T}$ is monotonic (see Lemma 3.1). But the contraction property with respect to the maximum norm is in general not satisfied. However, Bertsekas and Tsitsiklis proved for typical SSPs the existence and uniqueness of optimal values and the existence of optimal policies [9, 10].

Theorem 3.4 (Existence Theorem). *Consider an SSP $\mathcal{M} = (S, Act, e, P, R)$ with infinite-horizon total reward measure. Further assume that there exists a proper policy $\pi_p \in \Pi$ and for every improper policy π_i there exists $s \in S$ such that $V_\infty^{\pi_i}(s) = -\infty$.*

(i) There exists an optimal value V_∞^* which is the unique fixed point of $\mathcal{T}$, i.e $\mathcal{T}V_\infty^* = V_\infty^*$.
(ii) There exists an optimal policy π^* and it holds that $V_\infty^{\pi^*} = V_\infty^*$.
(iii) Every optimal policy π^* can be derived from the optimal value V_∞^* by

$$\pi^*(s) \in \operatorname*{argmax}_{a \in e(s)} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s')V_\infty^*(s') \right\}.$$

The dynamic programming algorithms value iteration and policy iteration as presented for discounted MDPs (Theorems 3.2 and 3.3) can be applied for SSPs in an analogous way and are shown in Theorems 3.5 and 3.6. The most important difference is the termination criterion in value iteration, which in contrast to the discounted case uses a weighted maximum norm in order to measure the distance between the iterated values and the optimal value. The proofs for both theorems can be found in [8, 10].

Theorem 3.5 (Value Iteration). *Consider an SSP $\mathcal{M}$ with the assumptions from Theorem 3.4. Let $V_0(s)$ be an arbitrary value function with $V_0(goal) = 0$. Define the sequence*

$$V_{n+1}(s) := (\mathcal{T}V_n)(s) = \max_{a \in e(s)} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s')V_n(s') \right\}.$$

(i) V_n converges to V^*.
(ii) If every policy is proper, then there exists $\xi \in \mathbb{R}^S$ with $\xi(s) \geq 1$ for all $s \in S$, such that $\mathcal{T}$ is a contraction with respect to the ξ-weighted maximum norm $||.||_\xi$ defined by

$$||V||_\xi := \max_{s \in S} \frac{|V(s)|}{\xi(s)}.$$

As Lipschitz constant q for contraction of $\mathcal{T}$ choose $q := \max_{s \in S} \frac{\xi(s)-1}{\xi(s)}$. For a given $\varepsilon > 0$ stop value iteration when $||V_{n+1} - V_n||_\xi < \frac{1-q}{2q}\varepsilon$ and choose

$$\pi_\varepsilon(s) \in \operatorname*{argmax}_{a \in e(s)} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s')V_{n+1}(s') \right\}.$$

Then $||V^{\pi_\varepsilon} - V^*||_\xi < \varepsilon$.

In the following, we want to briefly outline, how $\xi(s)$ can be determined. Consider an arbitrary policy π. Since π is proper by assumption, the expected number of transitions $t(s)$ from s to *goal* is finite. If π is the single available policy, then take $\xi(s) := t(s)$. In case there are more policies available, it could be the case, that there exists another policy which enlarges for some state s the expected number of transitions towards *goal*. Thus $\xi(s)$ can be chosen as the maximal expected number of transitions to *goal* among all policies. In order to compute ξ exactly, a modified SSP can be considered: Define a reward $R(s,a)$ which acts as a counter for the number of transitions to *goal*, i.e. each state $s \neq goal$ is rewarded $R(s,a) := 1$ independent of $a \in e(s)$ and the *goal* state is rewarded 0. Choose ξ as the optimal solution to the induced Bellman equation:

$$\xi(s) = 1 + \max_{a \in e(s)} \left\{ \sum_{s' \in S} P(s,a,s')\xi(s') \right\} \text{ for } s \neq goal \quad \text{and} \quad \xi(goal) = 0. \quad (3.8)$$

Note that if we allow improper policies π_i in the termination check of Theorem 3.5 then by definition of π_i there is some state s_i from which *goal* is reached with probability 0. In this case the Bellman equation (3.8) is not solvable, since any solution ξ would imply that $\xi(s_i) = \infty$. The proof for the termination criterion in Theorem 3.5 is completely analogous to the proof of Theorem 3.2 (see Appendix, page 236). It holds that for every policy π the linear operator T^π defined by $T^\pi V = R^\pi + P^\pi V$ is also a contraction with respect to $||.||_\xi$ and Lipschitz constant q as defined in Theorem 3.5.

For completeness we also state the policy iteration algorithm which is directly transfered from the discounted case as in Theorem 3.3 by setting the discount factor $\gamma := 1$. We omit the proof since it is analogous.

Theorem 3.6 (Policy Iteration). *Let $\pi_0 \in \Pi$ an arbitrary initial policy. Define the following iteration scheme.*

1. ***Policy evaluation**: Compute the value V^{π_n} of π_n by solving*

$$(I - P^{\pi_n})V^{\pi_n} = R^{\pi_n} \quad with \quad V^{\pi_n}(goal) = 0$$

and define the set of improving actions

$$A_{n+1}(s) := \operatorname*{argmax}_{a \in e(s)} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s')V^{\pi_n}(s') \right\}.$$

Termination: If $\pi_n(s) \in A_{n+1}(s)$ for all s then π_n is an optimal policy.

2. ***Policy improvement**: Otherwise choose an improving policy $\pi_{n+1}(s)$ such that $\pi_{n+1}(s) \in A_{n+1}(s)$.*

The sequence of values V^{π_n} is non-decreasing and policy iteration terminates in a finite number of iterations.

Example 3.3. Coming back to our queueing model from Example 3.1, we are now interested in the following two SSP problems:

(i) $\mathcal{M}_1$: the total expected profit up to first loss and
(ii) $\mathcal{M}_2$: the total expected number of accomplished jobs up to first loss.

For both models we add to $\mathcal{M}$ from Fig. 3.1 a reward-free *goal* state and redirect from the states representing a full queue into *goal* the probability mass for the job loss event (i.e. $q\overline{p_{d,n}}$ respectively $q\overline{p_{d,i}}$). Since by Definition 3.1 every state must have at least one action, we add an artificial action *idle* for looping in *goal* with probability 1.0. For model $\mathcal{M}_1$ and $s \neq$ *goal* the rewards $R_1^{\text{keep}}(s)$ and $R_1^{\text{move}}(s)$ are given as in (3.1) with $C_{\text{loss}} = \$0$. For model $\mathcal{M}_2$ the rewards R_2^{move} and R_2^{keep} are independent of the action and equal to $1 \cdot p_{d,n}$ in normal mode and $1 \cdot p_{d,i}$ in intense mode. We set all the rewards for both models to 0 in state *goal* when taking action *idle*.

If we set the completion probability $p_{d,i} = 1.0$ in the intense mode, it is obvious that going to intense mode would be optimal, since the expected reward for accomplishing a job is greater than the running costs in intense mode. Furthermore, no jobs would be lost in intense mode and thus the total value would diverge to ∞ for all states. Comparing this fact to the assumptions of Theorem 3.4, it holds that $p_{d,i} = 1.0$ implies the existence of an improper policy (moving respectively keeping in intense mode) which does not diverge to $-\infty$. Therefore, we set in the following $p_{d,i} = 0.6$. Now, every policy is proper, since the probability to be absorbed in *goal* is positive for all states and all policies. The following optimal policies π_i^* for model $\mathcal{M}_i$ and values V_i^* were computed by policy iteration.

$$\pi_1^* = \begin{pmatrix} \text{keep} \\ \text{keep} \\ \text{move} \\ \text{move} \\ \hline \text{keep} \\ \text{keep} \\ \text{keep} \\ \text{keep} \\ \hline \text{idle} \end{pmatrix}, V_1^* = \begin{pmatrix} 11447.5 \\ 11447.5 \\ 11047.5 \\ 8803.75 \\ \hline 11450.0 \\ 11490.0 \\ 11170.0 \\ 9230.0 \\ \hline 0.0 \end{pmatrix}, \quad \pi_2^* = \begin{pmatrix} \text{move} \\ \text{move} \\ \text{move} \\ \text{move} \\ \hline \text{keep} \\ \text{keep} \\ \text{keep} \\ \text{keep} \\ \hline \text{idle} \end{pmatrix}, V_2^* = \begin{pmatrix} 193.5 \\ 193.25 \\ 186.13 \\ 148.06 \\ \hline 193.5 \\ 193.5 \\ 187.5 \\ 154.5 \\ \hline 0.0 \end{pmatrix}$$

Note also that if value iteration is applied, then the maximal expected number $\xi(s)$ of transitions from s to *goal* is given as the optimal solution of the SSP in (3.8) by

$$\xi = (790.0, 785.0, 752.5, 596.25, 790.0, 786.0, 758.0, 622.0, 0.0)\,.$$

Therefore, the Lipschitz constant $q = 0.9988$ can be chosen, which is very high and makes value iteration in the ξ-weighted maximum norm terminating very late. Thus, the termination criterion in the ξ-weighted norm is a theoretical guarantee, but in general not applicable in practice. □

We finish the section on the infinite-horizon total reward measure with an outlook to other typical total value problems discussed in [33].

Remark 3.4. Beside SSPs one also often considers the following model types:

(i) ***Positive models***: For each state s there exists $a \in e(s)$ with $R(s,a) \geq 0$ and for all policies π the value $V^\pi(s) < \infty$ (this assumption can be relaxed).
(ii) ***Negative models***: For each state s all rewards $R(s,a) \leq 0$ and there exists a policy π with $V^\pi(s) > -\infty$ for all s.

In a positive model the goal is to maximize the accumulation of positive rewards towards ∞. The model assumptions make this possible, since in every state there is at least one non-negative reward available and the accumulation does not diverge to ∞. In contrast, the goal in a negative model is to maximize negative rewards, i.e. to minimize costs towards 0. The value iteration algorithm for both models is the same as in Theorem 3.5, but without convergence criteria. The policy iteration algorithm however differs from the policy iteration for SSPs. In a positive model, the initial policy π_0 has to be chosen suitable with $R^{\pi_0}(s) \geq 0$. Furthermore, in the policy evaluation phase the solutions to the linear equation $V = R^\pi + P^\pi V$ for a policy π spans up a whole subspace of $\mathbb{R}^S$. In this subspace a minimal solution $V_{\min} \geq 0$ has to be chosen in order to perform the policy improvement phase on $V_{\min}$. In contrast, in a negative model, the initial policy π_0 has to fulfill $V^{\pi_0} > -\infty$ and in the policy evaluation phase the maximal negative solution has to be computed. Puterman shows in [33] that for both model types value iteration converges for $V_0 = 0$, but convergence of policy iteration is only assured for positive models.

3.4 Average Reward Measure

We now come to the final part of the MDP section, which is devoted to the optimization of the average reward measure. As a reminder, let us consider for the moment a discrete-time Markov Reward Model $\mathcal{M} = (S, P, R)$. From Theorem 2.4 we know that if g is the average reward of $\mathcal{M}$ then $g = Pg$. On the other hand, the average reward cannot be uniquely determined by this single equation. Any solution u to $u = Pu$ defines a further linear equation $u + (I - P)v = R$ in v. If this additional equation is solvable, then and only then $u = g$ is the average reward. In this case the bias h is one of the possible solutions to the second equation (and unique modulo $\ker(I - P)$). We write both of these equations in the fixed-point form

$$g = Pg \quad \text{and} \quad h = (R + Ph) - g. \tag{3.9}$$

Also remember, that if $\mathcal{M}$ is unichain, then by Corollary 2.1 the equation $g = Pg$ can be simplified to "g is constant".

We are concerned in this section with the optimization of the average reward for an MDP $\mathcal{M} = (S, Act, e, P, R)$. Surely, every policy $\pi \in \Pi$ of $\mathcal{M}$ induces a DTMRM $\mathcal{M}^\pi = (S, P^\pi, R^\pi)$ and we can compute for each π the average reward

g^π as described above. By Definition 3.3 the optimal average value function is $g^*(s) = \sup_{\pi \in \Pi} g^\pi(s)$ and in case an optimal policy π^* exists then $g^{\pi^*}(s) \geq g^\pi(s)$ for all s and all π. In analogy to the previous sections, it is possible to establish Bellman equations which reduce the global optimization task over all policies $\pi \in \Pi$ to a local state-wise search over actions $e(s)$ for all states s. Since the formal derivation of these Bellman equations is slightly involved, we just state them and refer for proof to [33]. Define for each of the two linear fixed-point equations in (3.9) the Bellman operators $\mathcal{T}_{\text{av}} \colon \mathbb{R}^S \to \mathbb{R}^S$ and $\mathcal{T}^g_{\text{bias}} \colon \mathbb{R}^S \to \mathbb{R}^S$ (parametrized by $g \in \mathbb{R}^S$) as follows:

$$(\mathcal{T}_{\text{av}} g)(s) := \max_{a \in e(s)} \left\{ \sum_{s' \in S} P(s,a,s') g(s') \right\} \tag{3.10}$$

$$(\mathcal{T}^g_{\text{bias}} h)(s) := \max_{a \in e^g(s)} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s') h(s') \right\} - g(s) \tag{3.11}$$

$$\text{where } e^g(s) := \left\{ a \in e(s) \mid g(s) = \sum_{s' \in S} P(s,a,s') g(s') \right\}.$$

The corresponding Bellman optimality equations for the average reward are just the fixed-point equations of these operators and read as

$$g(s) = \max_{a \in e(s)} \left\{ \sum_{s' \in S} P(s,a,s') g(s') \right\} \tag{3.12}$$

$$h(s) = \max_{a \in e^g(s)} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s') h(s') \right\} - g(s). \tag{3.13}$$

Equations (3.12) and (3.13) are referred to as the first and the second optimality equation. In order to provide an intuition for these equations, assume for the moment that there is an optimal policy π^* with $g^{\pi^*} = g^*$ and moreover that the MDP $\mathcal{M}$ is unichain. In this case, the average reward $g^\pi(s)$ is a constant function for any policy π and thus the first optimality equation does not yield any further restriction since it is satisfied for every constant function g. Only the second equation takes the reward values $R(s,a)$ into account that are needed in order determine their average. For each policy π the average reward g^π and the bias h^π satisfy

$$h^\pi(s) = (R^\pi(s) - g^\pi(s)) + \sum_{s' \in S} P^\pi(s,s') h^\pi(s'),$$

i.e. the bias h^π is the total value function for the DTMRM with average-corrected rewards $R^\pi(s) - g^\pi(s)$. Since this holds especially for $\pi = \pi^*$, the second optimality equation can be seen as a Bellman equation for maximizing the total value function for the MDP model with rewards $R(s,a) - g^*(s)$. In other words, if π is an arbitrary policy then the DTMRM with rewards $R(s,\pi(s)) - g^*(s)$

has a total value function if and only if $g^*(s)$ is the average reward for the DTMRM (S, P^π, R^π) and this holds especially for $\pi = \pi^*$. In case $\mathcal{M}$ is multichain, then $g^{\pi^*}(s)$ is only constant on recurrent classes of $\mathcal{M}^{\pi^*}$, whereas if s is transient then $g^{\pi^*}(s)$ is a weighted sum over all those average rewards on recurrent classes. This means that g^{π^*} has to fulfill $g^{\pi^*} = P^{\pi^*} g^{\pi^*}$ in addition and thus g^{π^*} is a solution to the first optimality equation. Since both Bellman equations are nested and have to be satisfied simultaneously, it is possible to reduce the set of actions $e(s)$ in the second equation to the maximizing actions $e^g(s) = \operatorname{argmax}_{a \in e(s)} \left\{ \sum_{s' \in S} P(s,a,s') g(s') \right\}$ for a solution g of the first equation. Note that in case of unichain models it holds that $e^g(s) = e(s)$ for all s.

The following theorem formalizes the explanations in the motivation above and connects the Bellman equations to the optimal value and optimal policies. We refer for proof to [33].

Theorem 3.7 (Existence Theorem). *Consider an MDP $\mathcal{M}=(S, Act, e, P, R)$ with average reward measure.*

(i) The average optimal value function g^* is a solution to (3.12)*, i.e $g^* = \mathcal{T}_{\mathrm{av}} g^*$. For $g = g^*$ there exists a solution h to* (3.13)*, i.e. $h = \mathcal{T}_{\mathrm{bias}}^{g^*} h$. If g and h are solutions to* (3.12) *and* (3.13) *then $g = g^*$.*

(ii) There exists an optimal policy π^* and it holds that $g^{\pi^*} = g^*$.

(iii) For any solution h to (3.13) *with $g = g^*$ an optimal policy π^* can be derived from*

$$\pi^*(s) \in \operatorname*{argmax}_{a \in e^{g^*}(s)} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s') h(s') \right\}.$$

As a special case of part (iii) in this theorem it holds that if for a policy π the average reward g^π and the bias h^π solve the Bellman optimality equations, then π is optimal. In contrast to the discounted and total reward cases, the converse does not hold. This means that if a policy π is optimal then g^π and h^π are not necessary solutions to the Bellman equations.
The following policy iteration algorithm, can be applied in order to compute the optimal average reward g^* as well as an optimal policy π^*. A proof can be found in [33].

Theorem 3.8 (Policy Iteration). *Let $\pi_0 \in \Pi$ be an initial policy. Define the following iteration scheme:*

1. ***Policy evaluation**: Compute a solution $(g^{\pi_n}, h^{\pi_n}, w)^T$ to*

$$\begin{pmatrix} I - P^{\pi_n} & 0 & 0 \\ I & I - P^{\pi_n} & 0 \\ 0 & I & I - P^{\pi_n} \end{pmatrix} \begin{pmatrix} g^{\pi_n} \\ h^{\pi_n} \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ R^{\pi_n} \\ 0 \end{pmatrix}. \tag{3.14}$$

Define

$$G_{n+1}(s) := \operatorname*{argmax}_{a \in e(s)} \left\{ \sum_{s' \in S} P(s,a,s') g^{\pi_n}(s') \right\}.$$

2. ***Policy improvement***: *If* $\pi_n(s) \notin G_{n+1}(s)$ *for some* s *then choose an improving policy* π_{n+1} *with* $\pi_{n+1}(s) \in G_{n+1}(s)$ *and go to the policy evaluation phase. Otherwise if* $\pi_n(s) \in G_{n+1}(s)$ *for all* s *then define*

$$H_{n+1}(s) := \underset{a \in e^{g^{\pi_n}}(s)}{\operatorname{argmax}} \left\{ R(s,a) + \sum_{s' \in S} P(s,a,s') h^{\pi_n}(s') \right\}.$$

If $\pi_n(s) \notin H_{n+1}(s)$ *for some* s *then choose an improving policy* π_{n+1} *such that* $\pi_{n+1}(s) \in H_{n+1}(s)$ *and go to the policy evaluation phase.*
Termination: If $\pi_n(s) \in H_{n+1}(s)$ *for all* s *then* π_n *is an optimal policy.*

The values g^{π_n} *are non-decreasing and policy iteration terminates in a finite number of iterations with an optimal policy* π_n *and optimal average value* g^{π_n}.

Note that (3.14) corresponds to (2.19) instead of (2.18) for the following reason: For each policy π (2.18) provides a unique solution g^π to the average reward, but in general the bias h^π cannot be uniquely determined. Policy iteration can be assured to converge if the bias h^{π_n} is computed for each iterated policy π_n [33]. As described in Remark 2.4(iii) this can be done by solving (2.19), i.e. the equation $v + (I - P^{\pi_n})w = 0$ in addition to (2.18) for which $v = h^{\pi_n}$ is the unique solution. There are also other possibilities to assure convergence of policy iteration by solving only (2.18) and fixing a scheme that chooses a solution v to $g+(I-P)v = R$ in order to prevent cycles in policy iteration (see Remark 2.4(ii)).

Before showing the application of policy iteration on our queueing model running example, we first state the following remark regarding some algorithmic aspects.

Remark 3.5. (i) During policy iteration the action set $e^{g^{\pi_n}(s)}$ can be replaced by the whole action set $e(s)$ – this leads to the so-called modified optimality equations. The convergence and the optimality of the solution in policy iteration are not influenced by this replacement.

(ii) In the policy improvement phase, there are two jumps to the policy evaluation phase, which represent two nested cycles of evaluation and improvement phases. First, a policy π_n has to be found, which solves the first optimality equation. Then in a nested step, π_n is tested on the second optimality equation. If π_n can be improved by a better policy π_{n+1} with actions from H_{n+1} then π_{n+1} has to be sent back to the first evaluation and improvement cycle until it again solves the first optimality equation, and so on.

(iii) As already mentioned in the introducing motivation, if it is a priori known that the MDP is unichain, i.e. for all policies there is only one closed recurrent class of states, then the optimal average reward is constant and the first optimality equation is automatically satisfied (see Corollary 2.1). This reduces the complexity of policy iteration, since only the second optimality equation has to be considered for optimization.

(iv) We skip the value iteration algorithm in this tutorial since it is exactly the same as for the discounted case (Theorem 3.2) with $\gamma := 1$. It can be

proven that the sequence $V_{n+1} - V_n$ converges to the optimal average reward g^*, if for every (optimal) policy the transition matrix is aperiodic [33]. The aperiodicity constraint is not a restriction, since every periodic DTMRM can be made aperiodic, by inserting self-loops with strictly positive probability for every state. (The reward function has to be transformed accordingly.) However, [33] presents a termination criterion for value iteration only for models with $g^*(s)$ constant for all s (e.g. unichain models).

Example 3.4. Consider the queueing MDP model from Example 3.1. We want to compute the optimal average value function for the queueing model with parameters $q = 0.25$, $p_{d,n} = 0.5$ and $p_{d,i} = 1.0$ and the reward structure as specified in (3.1). Note that the model is multichain, since the policy that takes the action keep in every state induces a DTMRM with two recurrent classes. Policy iteration converges after three iterations (with initial policy π_0 which keeps in normal mode or moves to it from intense mode) and results in the following optimal policy π^*, optimal average value $g^* = g^{\pi^*}$ and bias h^{π^*}:

$$\pi^* = \begin{pmatrix} \text{keep} \\ \text{keep} \\ \text{keep} \\ \text{move} \\ \hline \text{move} \\ \text{move} \\ \text{keep} \\ \text{keep} \end{pmatrix}, \quad g^{\pi^*} = \begin{pmatrix} 22.7 \\ 22.7 \\ 22.7 \\ 22.7 \\ \hline 22.7 \\ 22.7 \\ 22.7 \\ 22.7 \end{pmatrix}, \quad h^{\pi^*} = \begin{pmatrix} -49.4 \\ 41.3 \\ 95.3 \\ 38.7 \\ \hline -59.4 \\ 40.6 \\ 130.3 \\ 220.0 \end{pmatrix}.$$

Thus, the optimal policy induces a DTMRM that is unichain with constant optimal average reward 22.7. The finite-horizon total value function $V_N^{\pi^*}$ from state $(0, 0, \text{normal}, \text{idle})$ increases asymptotically with $22.7 \cdot N - 49.4$ as $N \to \infty$. □

We conclude the MDP section by a further remark, which presents a short outlook on other optimization criteria that are applicable for MDPs.

Remark 3.6. The average reward considers reward accumulation in the long-run. Therefore, it is not very sensitive in the selection between policies with the same average reward: If two policies have the same long-run average reward but different short-run rewards, then one would prefer among all policies with the same average reward such a policy that also maximizes the short-run reward accumulation. This idea leads to the so-called bias (or more general ***n-bias*** or ***n-discount***) optimization criteria, which belongs to the class of ***sensitive discount optimality*** criteria. In more detail, policies $\pi \in \Pi$ are compared regarding their Laurent series expansions (2.15)

$$(V^\gamma)^\pi = a_{-1} g^\pi + a_0 H^\pi R^\pi + a_1 (H^\pi)^2 R^\pi + \ldots,$$

where a_i are constants (depending only on γ) and $H^\pi R^\pi = h^\pi$ is the bias, which represents the excess in reward accumulation up to steady-state. Now if a subset of policies $\Pi^*_{-1} \subseteq \Pi$ maximize g^π then this subset can be further refined to a reduced subset $\Pi^*_0 \subseteq \Pi^*_{-1}$ by maximizing the bias h^π for $\pi \in \Pi^*_{-1}$ in addition to the average reward g^π. If Π^*_0 still consists of more then one policy, then one can proceed iteratively and compare the higher order n-bias terms sequentially. Note that the n-bias reward measure specifies the vector space $\mathcal{V}$ of value functions as $\mathcal{V} = \left\{V : S \to \mathbb{R}^{n+1}\right\}$ and values can be compared in Definition 3.3 by lexicographic order.

The most sensitive optimization criterion is the ***Blackwell optimality criterion***, which selects a policy π^*, such that the entire Laurent series expansion (i.e. the complete discounted value) is maximal among the discounted values for all policies and for all discount factors high enough, i.e.

$$\exists \gamma^* \in [0,1)\ \forall \gamma \in [\gamma^*, 1)\ \forall \pi \in \Pi :\ (V^\gamma)^{\pi^*} \geq (V^\gamma)^\pi .$$

It can be shown, that Blackwell optimal policies exist and a Blackwell optimal policy is n-bias optimal for all n [33]. Furthermore, such a policy can be computed by proceeding with the policy space reduction as described above until some Π^*_n consists of only one policy, which is Blackwell optimal (or if $n \geq |S| - 2$ then Π^*_n is a set of Blackwell optimal policies).

4 Continuous Time Markov Reward Models

In both DTMRMs and MDPs time is discrete, i.e. it proceeds in a step by step manner. However, this is unrealistic for many applications – one rather wishes to work with a continuous notion of time. Therefore, in this and the following section, we study continuous-time models. Since we stick to the principle of memorylessness, it will turn out that the state sojourn times follow an exponential distribution (as opposed to the geometrically distributed sojourn times in the discrete-time setting).

4.1 Preliminaries

Definition 4.1. *A **continuous-time Markov chain (CTMC)** is a structure $\mathcal{M} = (S, Q)$ with finite state space S and generator function $Q\colon S \times S \to \mathbb{R}$ such that $Q(s, s') \geq 0$ for $s' \neq s$ and $\sum_{s' \in S} Q(s, s') = 0$ for all s. A **continuous-time Markov Reward Model (CTMRM)** is a structure (S, Q, i, r) which enhances a CTMC (S, Q) by a reward structure consisting of an impulse reward function $i\colon S \times S \to \mathbb{R}$ for transitions with $i(s, s) = 0$ for all $s \in S$ and a rate reward function $r\colon S \to \mathbb{R}$ for states.*

From state $s \in S$ each quantity $Q(s, s')$ with $s' \neq s$ defines an event which occurs after a random amount of time $\tau_{s,s'} \in \mathbb{R} \cup \{\infty\}$ to trigger. If $Q(s, s') > 0$ then $\tau_{s,s'}$ is exponentially distributed with rate $Q(s, s')$ and otherwise if $Q(s, s') = 0$ then we set $\tau_{s,s'} := \infty$. For a fixed $s \in S$ all these times $\tau_{s,s'}$ are independent and

concurrently enabled. Therefore, they define a race among each other and only that τ_{s,s'_0} which triggers first, within a finite amount of time (i.e. $\tau_{s,s'_0} \leq \tau_{s,s'}$ for all $s' \in S$), wins the race. In this case the system performs a transition from s to $s'_0 \neq s$ and collects the impulse reward $i(s, s'_0) \in \mathbb{R}$. The time τ_s that the system resides in state s (up to transition) is called the ***sojourn time*** and fulfills $\tau_s = \min\{\tau_{s,s'} \mid s' \neq s\}$. While the system is in state s the rate reward $r(s)$ is accumulated proportionally to the sojourn time τ_s. Thus, the accumulated reward in s for the sojourn time including the transition to state s'_0 is given by $R(s) := i(s, s'_0) + r(s)\tau_s$. The quantity $E(s) := \sum_{s' \neq s} Q(s, s')$ is called the ***exit rate*** in state s and by definition of the generator function Q it holds that $E(s) = -Q(s, s) \geq 0$. If $E(s) > 0$ then there is some state s' with $Q(s, s') > 0$ and due to the race condition it holds that τ_s is exponentially distributed with rate $E(s)$. The ***transition probability*** $P(s, s')$ is the probability that $\tau_{s,s'}$ wins the race and is given by $P(s, s') = P(\tau_{s,s'} = \tau_s) = \frac{Q(s,s')}{E(s)}$. Otherwise, if $E(s) = 0$ (or all $Q(s, s') = 0$) then $\tau_s = \infty$ and state s is absorbing. In this case we set $P(s, s') := \delta_{s,s'}$, i.e. $P(s, s') = 1$ if $s' = s$ and $P(s, s') = 0$ if $s' \neq s$. The function $P : S \to \mathcal{D}(S)$ with $(P(s))(s') := P(s, s')$ is called the ***embedded transition probability*** function. The model (S, P) can be considered as a discrete-time Markov chain, which models the transitions of the underlying CTMC and abstracts from the continuous time information. Similarly to DTMRMs, $i(s) := \sum_{s' \neq s} P(s, s')i(s, s')$ will denote the state-based version of the transition-based impulse reward $i(s, s')$, i.e. $i(s)$ is the expected impulse reward from state s to some other state $s' \neq s$.

Example 4.1 (WSN node model). A wireless sensor network (WSN) consists of nodes that have to observe their environment by sensing activities and transmit information towards a destination. Each node consists of a battery unit with some initial capacity, a sensor and a transmitter. Furthermore, environmental events occur randomly. For the purposes of this section and in order to show how CTMRMs can be modelled, we assume a very simple WSN node model (see Fig. 4.1), which consists only of

- one sensor node, which randomly switches between "idle" and "listen" states after an exponentially distributed time and does not transmit any information and
- the environment, in which activities occur and endure in a memoryless way.

For simplicity we further assume that the node has infinite energy supply and does not consume any energy in idle mode. In case an environmental activity takes place and the node is listening, it must observe the activity at least until it stops. When the sensor switches from idle to listen it consumes instantaneously 5 energy units. While if the sensor is listening it consumes energy with rate 10. We want to measure the energy consumption in this model. Suitable measures could be the average energy consumption (per time unit) or some discounted energy consumption. □

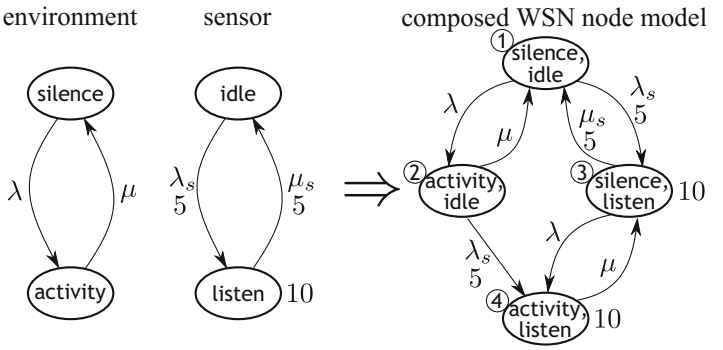

Fig. 4.1. A simple WSN node model, which consists of a single node which can listen to activities in the environment. The transition rates are $\lambda = 2$, $\mu = 4$, $\lambda_s = 4$, $\mu_s = 30$. If the sensor is listening, it uses 10 energy units per time. For every activation to "listen" or deactivation to "idle" an impulse energy of 5 units is employed.

Remark 4.1. Puterman [33] allows impulse rewards which are state-based and are gained in a state s if s is the initial state or when s is reached after some transition from s' to s ("arrival" point of view). In contrast, we have defined transition-based impulse rewards $i(s, s')$ that are gained when state s is left, i.e. a transition from s to s' is performed ("departure" point of view). Therefore, the impulse reward can be considered state-based as the expectation $i(s) = \sum_{s' \neq s} i(s, s')P(s, s')$ over transition probabilities. When considering the infinite-horizon total reward measure or the average reward measure, then both points of view lead to the same value functions and thus their distinction doesn't matter in this case. However, this difference is important when we are dealing with the finite-horizon total reward measure and the discounted reward measure.

Before being able to define and evaluate reward measures for the continuous-time case, we have to provide more theoretical background. The next section is devoted to this.

4.2 Probability Space for CTMCs

In the following, we want to formalize the transition behavior of a CTMC $\mathcal{M} = (S, Q)$ that we have informally introduced in Sect. 4.1. For this reason, we first define a suitable sample space Ω together with a Borel σ-algebra $\mathcal{B}(\Omega)$ consisting of those measurable events for which we will assign probabilities. Subsequently, we will define several stochastic processes upon Ω that are all induced by the CTMC $\mathcal{M}$. These processes will allow us to define a time-dependent transition probability matrix, which in turn will play an important role for the definition of reward measures for a CTMRM.

4.2.1 Sample Space

Since continuous time plays a role for these measures, we put this information along with the state space S into the sample space. Define the sample space $\Omega \subseteq$

$(S \times (0,\infty])^{\mathbb{N}}$ as the set of infinite paths of the form $\omega = (s_0, t_0, s_1, t_1, s_2, t_2, \dots)$ such that for all $i \in \mathbb{N}$:

$$(E(s_i) > 0 \Rightarrow Q(s_i, s_{i+1}) > 0 \wedge t_i < \infty) \vee (E(s_i) = 0 \Rightarrow s_{i+1} = s_i \wedge t_i = \infty)\,.$$

Roughly speaking, ω represents a sample path, where $t_i < \infty$ is the finite sojourn time in a non-absorbing state s_i or otherwise if s_i is absorbing then for all $j \geq i$ it holds that $t_j = \infty$ and $s_j = s_i$. A sample path $\omega = (s_0, t_0, s_1, t_1, \dots)$ can also be considered as a jump function $\omega : [0,\infty) \to S$ that is constantly s_0 for all $t \in [0, t_0)$ and if $t_0 \neq \infty$ then ω jumps to state $s_1 \neq s_0$ at t_0 and $\omega(t) = s_1$ for all $t \in [t_0, t_0 + t_1)$. If $t_1 \neq \infty$ then ω has a next jump to state $s_2 \neq s_1$ at $t_0 + t_1$ and so on, until there is eventually a first index i with $t_i = \infty$ and therefore $\omega(t) = s_i$ for all $t \geq \sum_{k=0}^{i-1} t_k$. In order to define a probability space over Ω we transform Ω to the set $Path$ of finite absorbing and infinite paths as defined in [3]. Let $\psi\colon \Omega \to Path$ be the transformation that drops the artificial repetitions of absorbing states, i.e.

$$\psi(s_0, t_0, s_1, t_1, \dots) := \begin{cases} (s_0, t_0, s_1, t_1, \dots), & \text{if } \forall k \in \mathbb{N} : t_k < \infty \\ (s_0, t_0, s_1, t_1, \dots, s_l), & l := \min\{k \mid t_k = \infty\} < \infty \end{cases}$$

where $\min \emptyset := \infty$. Note that in the definition of ψ the two cases are disjoint. Since ψ is bijective the probability space $(Path, \mathcal{F}(Path), Pr_\alpha)$ as defined in [3] (where $\alpha \in \mathcal{D}(S)$ is a distribution over initial states) induces for each $s \in S$ a probability space $(\Omega, \mathcal{B}(\Omega), P_s)$ in a canonical way:

$$\mathcal{B}(\Omega) := \{A \subseteq \Omega \mid \psi(A) \in \mathcal{F}(Path)\} \quad \text{and} \quad P_s := Pr_{\delta_s} \circ \psi,$$

where we choose $\alpha := \delta_s$ with $\delta_s(s') := \delta_{s,s'}$ (i.e. s is the initial state). Before moving on, we want to mention that both sample spaces Ω and $Path$ are equivalent, since ψ is bijective (and measurable by definition of $\mathcal{B}(\Omega)$). The sample space $Path$ allows for an intuitive interpretation of sample paths ω regarded as jump functions $\omega : [0,\infty) \to S$ as described above. Every jump function that is constant on intervals of positive length has at most a finite or countably infinite number of jumps – this distinction is encoded in the sample paths of $Path$. However, this differentiation of cases would directly be transferred to a corresponding differentiation in the definition of stochastic processes that we will introduce in the sequel. For this reason, we have chosen Ω as the sample space which embedds these cases already in its definition and thus does not lead to an overload of notation in the definition of these processes.

4.2.2 Induced Stochastic Processes

The CTMC $\mathcal{M} = (S, Q)$ induces a number of stochastic processes over Ω. For $\omega = (s_0, t_0, s_1, t_1, \dots) \in \Omega$ define the

(i) discrete-time state process $(X_n)_{n \in \mathbb{N}}$ by

$$X_n(\omega) := s_n$$

(ii) sojourn time $(\tau_n)_{n\in\mathbb{N}}$, where

$$\tau_n(\omega) := t_n \leq \infty$$

(iii) total elapsed time $(T_n)_{n\in\mathbb{N}}$ for the first n transitions as

$$T_n(\omega) := \sum_{i=0}^{n-1} \tau_i(\omega)$$

(iv) number of transitions $(N_t)_{0\leq t<\infty}$ up to time t as

$$N_t(\omega) := \max\{n \mid T_n(\omega) \leq t\} \in \mathbb{N}$$

(note that with probability 1 the maximum is taken over a finite set and thus N_t is almost surely finite, i.e. $P(N_t < \infty) = 1$)

(v) continuous-time state process $(Z_t)_{0\leq t<\infty}$, where

$$Z_t(\omega) := X_{N_t(\omega)}(\omega),$$

i.e. Z_t is the state of the system at point in time $t \geq 0$.

Remark 4.2. For all $t \in [0, \infty)$ and $n \in \mathbb{N}$ the following equalities of events hold:

$$\{N_t = n\} = \{T_n \leq t < T_{n+1}\} \quad \text{and} \quad \{N_t \geq n\} = \{T_n \leq t\}.$$

The discrete-time state process X_n represents the n-th visited state (or an absorbing state) and it fulfills the discrete-time Markov property as in (2.4), i.e. for all $s, s_0, s_1, \ldots s_k \in S$ and $0 < n_1 < \cdots < n_k < n$

$$P_{s_0}(X_n = s \mid X_{n_1} = s_1, \ldots, X_{n_k} = s_k) = P_{s_0}(X_n = s \mid X_{n_k} = s_k).$$

From $Z_t(\omega) = X_{N_t(\omega)}(\omega)$ and $N_t(\omega)$ non-decreasing for all ω it follows that the continuous-time state process Z_t also fulfills the Markov property, which reads as a continuous time version:

$$P_{s_0}(Z_t = s \mid Z_{t_1} = s_1, \ldots, Z_{t_k} = s_k) = P_{s_0}(Z_t = s \mid Z_{t_k} = s_k)$$

for all $s, s_0, s_1, \ldots s_k \in S$ and $0 \leq t_1 < \cdots < t_k < t$. Thus given knowledge about the state $Z_{t_k} = s_k$ of the process for any arbitrary point in time $t_k < t$, then the process Z_t does not depend on its history comprising the visited states before time t_k. It further holds that Z_t is homogeneous in time, i.e. the following property holds:

$$P_{s_0}(Z_{t+t'} = s' \mid Z_t = s) = P_s(Z_{t'} = s').$$

As in Sect. 2 we fix a representation of the state space S through indices $\{1, 2, \ldots, n\}, n := |S|$ such that functions $S \to \mathbb{R}$ can be represented by vectors in $\mathbb{R}^n$ and functions $S \times S \to \mathbb{R}$ as matrices in $\mathbb{R}^{n\times n}$. Define the ***transient probability matrix*** $P(t)$ as

$$P(t)(s, s') := P_s(Z_t = s'). \tag{4.1}$$

The matrix $P(t)$ is stochastic for all $t \geq 0$ and fulfills the property

$$P(t+t') = P(t)P(t') \quad \forall t, t' \geq 0,$$

which reads componentwise as $P(t+t')(s,s') = \sum_u P(t)(s,u) \cdot P(t')(u,s')$. This means that from state s the probability to be in state s' after $t+t'$ time units is the probability to be in some arbitrary state $u \in S$ after t time units and traverse from there within further t' time units to state s'. It can be shown that all entries of $P(t)$ are differentiable for all $t \geq 0$ and $P(t)$ is related to the generator matrix Q of the CTMC by the ***Kolmogorov differential equations***

$$\frac{d}{dt}P(t) = QP(t) \quad \text{and} \quad \frac{d}{dt}P(t) = P(t)Q, \tag{4.2}$$

which read in componentwise notation as

$$\frac{d}{dt}\left(P(t)(s,s')\right) = \sum_u Q(s,u) \cdot P(t)(u,s') = \sum_v P(t)(s,v) \cdot Q(v,s').$$

All solutions to these equations are of the form $P(t) = e^{Qt}$ since $P(0) = I$ is the identity matrix, where for a matrix A the quantity e^A denotes the matrix exponential that is given by $e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$.

4.2.3 State Classification

As in Sect. 2.1.2 there is also a classification of states in case of continuous time Markov chains. Since this taxonomy is almost the same as in the discrete-time case, we only present it very briefly. The most important difference is that in the continuous-time setup there is no notion for periodicity of states and it can be shown that the matrix $P(t)$ converges as $t \to \infty$ (for finite state spaces). We denote the limit by $P^* := \lim_{t\to\infty} P(t)$. Note that in Definition 2.9 we denoted the corresponding discrete-time limiting matrix as P^∞ and its time-averaged version as P^* and mentioned in Proposition 2.2 that they both coincide if P^∞ exists. Since the existence of this limit in the continuous-time case is always guaranteed, we call this limit directly P^* instead of P^∞ in order to use similar notation. One can show that P^* is stochastic and fulfills the invariance conditions

$$P^*P(t) = P(t)P^* = P^*P^* = P^*.$$

Therefore, the probability distribution $P^*(s,\cdot) \in \mathcal{D}(S)$ in each row of P^* is a ***stationary distribution*** and since $P(t)(s,\cdot) \to P^*(s,\cdot)$ as $t \to \infty$ it is also the ***limiting distribution*** from state s. Furthermore, it holds that

$$P^*Q = QP^* = 0,$$

which can be derived from (4.2) and $\frac{d}{dt}P(t) \to 0$ as $t \to \infty$.

Let the random variable $M_s \in (0, \infty]$ denote the point in time when the state process Z_t returns to s for the first time (given $Z_0 = s$). A state s is ***transient*** if $P_s(M_s = \infty) > 0$ or equivalently $P^*(s,s) = 0$. In the other case,

if $P_s(M_s < \infty) = 1$ then s is called ***recurrent*** and it holds equivalently that $P^*(s,s) > 0$. It can be shown that there is always at least one recurrent state if the state space is finite. A state s' is ***reachable*** from s if $P(t)(s,s') > 0$ for some $t \geq 0$. The states s and s' are communicating if s' is reachable from s and s is reachable from s'. This communication relation is an equivalence relation and partitions the set of recurrent states into ***closed recurrent classes***. Therefore, the state space partitions into $S = \bigcup_{i=1}^{k} S_i^r \cup S^t$, where S^t denotes the set of transient states and S_i^r is a closed recurrent class for all $i = 1, \ldots, k$. For $s, s' \in S_i^r$ in the same recurrent class it holds that $P^*(s,s') > 0$. As in the discrete-time case P^* can be represented by

$$P^* = \begin{pmatrix} P_1^* & 0 & 0 & \ldots & 0 & 0 \\ 0 & P_2^* & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & 0 & \ldots & P_k^* & 0 \\ \widetilde{P}_1^* & \widetilde{P}_2^* & \widetilde{P}_3^* & \ldots & \widetilde{P}_k^* & 0 \end{pmatrix} \tag{4.3}$$

where P_i^* has identical rows for the stationary distribution in class S_i^r and $\widetilde{P}_i^*$ contains the trapping probabilities from transient states S^t into S_i^r. If a closed recurrent class consists of only one state s, then s is called ***absorbing***. A CTMC is ***unichain*** if $k = 1$ and ***multichain*** if $k \geq 2$. A unichain CTMC is called ***irreducible*** or ***ergodic*** if $S^t = \emptyset$.

4.3 Model Transformations

In this section we present a set of model transformations, which will allow us to

- unify the different types of rewards (impulse reward and rate reward) in the reward accumulation process ("Continuization") and
- relate some continuous-time concepts to discrete-time Markov Reward Models from Sect. 2 ("Embedding" and "Uniformization").

These transformations simplify the evaluation process of all the reward measures and map the computation of the value functions for continuous-time models to the discrete-time case.

4.3.1 Embedding

As mentioned in Sect. 4.1, a CTMC (S, Q) defines for all states $s, s' \in S$ the embedded transition probabilities $P(s,s')$. The structure (S,P) can be considered as a discrete-time Markov chain and it induces on the sample space $\Omega' := \left\{(s_0, s_1, s_2, \ldots) \in S^{\mathbb{N}} \mid P(s_{i-1}, s_i) > 0 \text{ for all } i \geq 1\right\}$ as in (2.3) the state process X'_n (by Definition 2.3) given by $X'_n(s_0, s_1, \ldots) = s_n$. This stochastic process is related to the discrete-time state process $X_n : \Omega \to S$ by abstracting away from the time information, i.e. for all $n \in \mathbb{N}$

$$X_n(s_0, t_0, s_1, t_1, \ldots) = X'_n(s_0, s_1, \ldots).$$

This equation establishes the connection to DTMCs and thus X_n can be considered as the state process of the DTMC (S, P). Therefore, (S, P) is also called the ***embedded discrete-time Markov chain*** and X_n is the ***embedded state process*** of the CTMC (S, Q).

Now consider a CTMRM (S, Q, i, r) and define a function $R : S \times S \to \mathbb{R}$ where $R(s, s')$ denotes the expected accumulated rate reward $r(s)$ in state s over time including the impulse reward $i(s, s')$ gained for transition from s to some other state $s' \neq s$ (as in Sect. 4.1). If s is non-absorbing, then the sojourn time τ_s in s is exponentially distributed with rate $E(s) > 0$ and $R(s, s')$ is given by

$$R(s, s') := i(s, s') + \frac{r(s)}{E(s)}. \tag{4.4}$$

Otherwise, if s is absorbing, the embedding is only possible if $r(s) = 0$ and in this case we define $R(s, s') := 0$ for all s'.

It is very important to note that if we consider a reward measure on the CTMRM with value function V and a corresponding reward measure on the transformed DTMRM with value function V', then it is of course desirable that $V = V'$, i.e. the transformation should be ***value-preserving***. This allows to compute the value V by applying the theory and algorithms for the discrete-time models as presented in Sect. 2. However, as we will see, such a model transformation needs in general the reward measure itself as input in order to be value-preserving. As an example, the integration of the rate reward $r(s)$ into the reward $R(s, s')$ is performed by total expectation over an infinite time-horizon, which gives the term $\frac{r(s)}{E(s)}$. If one considers a finite horizon for the continuous-time model, then $R(s, s')$ as defined is obviously not the appropriate reward gained in state s in the embedded discrete-time model.

4.3.2 Uniformization

We have seen in Sect. 4.1 that the quantities $Q(s, s')$ for $s' \neq s$ can be regarded as rates of exponentially distributed transition times $\tau_{s,s'}$. All these transition events define a race and only the fastest event involves a transition to another state $s' \neq s$. We can manipulate this race, by adding to the set of events $\{\tau_{s,s'} \mid s' \neq s\}$ of a state s an auxiliary exponentially distributed event $\tau_{s,s}$ with an arbitrary positive rate $L(s) > 0$ that introduces a self-loop (i.e. a transition from s to s), if it wins the race. The time up to transition is $\tau_s := \min\{\tau_{s,s'} \mid s' \in S\}$ and it is exponentially distributed with increased exit rate $\widetilde{E}(s) := E(s) + L(s)$. The probability that $\tau_{s,s}$ wins the race can be computed to $P(\tau_{s,s} \leq \tau_{s,s'} \ \forall s' \in S) = \frac{L(s)}{\widetilde{E}(s)} = 1 + \frac{Q(s,s)}{\widetilde{E}(s)}$ and for all $s'_0 \neq s$ it holds that $P(\tau_{s,s'_0} \leq \tau_{s,s'} \ \forall s' \in S) = \frac{Q(s,s')}{\widetilde{E}(s)}$. We can add such events $\tau_{s,s}$ to a set of states s and thus increase the exit rates for all these states simultaneously. Moreover, we can choose an arbitrary $\mu > 0$ with $\max\{E(s) \mid s \in S\} \leq \mu < \infty$ (called ***uniformization rate***) such that $\widetilde{E}(s) \equiv \mu$ is constant for all $s \in S$. The uniformization rate μ allows to define a transformation to the ***μ-uniformized***

DTMRM $\mathcal{M}^\mu := (S, P^\mu, R^\mu)$ where a transition from s to s' in $\mathcal{M}^\mu$ captures the event that $\tau_{s,s'}$ wins the race and thus

$$P^\mu(s, s') := \delta_{s,s'} + \frac{Q(s, s')}{\mu}. \tag{4.5}$$

Note that the probability to eventually leave state s to a state $s' \neq s$ is exactly the embedded transition probability $P(s, s') = \sum_{i=0}^{\infty} P^\mu(s, s)^i P^\mu(s, s')$. The reward $R^\mu(s, s')$ combines the accumulated rate reward in state s and the impulse reward up to transition to some state s'. In the CTMRM the rate reward $r(s)$ is accumulated for the complete sojourn time in s. Since self-loops are possible in the uniformized DTMRM the accumulation process stops when an arbitrary transition occurs. The expected value of the accumulated rate reward up to transition is given by $r(s) \cdot \frac{1}{\mu}$. Furthermore, the impulse reward $i(s, s')$ is only gained if a transition to another state $s' \neq s$ takes place. But since $i(s, s) = 0$ for all $s \in S$ by Definition 4.1 it follows that for all $s, s' \in S$ the total uniformized reward $R^\mu(s, s')$ is given by

$$R^\mu(s, s') := i(s, s') + \frac{r(s)}{\mu}. \tag{4.6}$$

This equation is similar to (4.4) with the difference that the exit rate $E(s)$ is replaced by the uniformization rate $\mu \geq E(s)$. A further difference comes into the picture when considering the accumulation of these rewards. Both rewards $R(s, s)$ and $R^\mu(s, s)$ for self-loops are possibly non-zero. In case of the embedded DTMRM the probability $P(s, s)$ for self-loops is 0 in non-absorbing states s and thus $R(s, s)$ is not accumulated, in contrast to the uniformized model where $P^\mu(s, s) > 0$ is possible.

So far we have defined the two transformations "Embedding" and "Uniformization" both discretizing the continuous time of a CTMRM and the accumulation of the rate reward over time. In contrast, the upcoming third transformation does not modify the time property itself, but rather merges the impulse rewards into the rate reward. In this way, the CTMRM model has no discrete contributions in the reward accumulation process, which allows to simplify the evaluations of the reward measures (as we will see in the upcoming sections).

4.3.3 Continuization

Let $\mathcal{M} = (S, Q, i, r)$ be a CTMRM and for a non-absorbing state s denote $R(s) := \sum_{s' \neq s} P(s, s') R(s, s')$, where $R(s, s')$ is as in (4.4) and $P(s, s')$ is the embedded transition probability. Thus

$$R(s) = \sum_{s' \neq s} P(s, s') i(s, s') + \frac{r(s)}{E(s)}$$

is the expected accumulated rate reward $r(s)$ in state s including the expected impulse reward $\sum_{s' \neq s} P(s, s') i(s, s') = i(s)$ gained for transition from s to

some other state $s' \neq s$. Consider for the moment that $i(s,s') = 0$ for all s, s', i.e. there are no impulse rewards defined. Then $r(s) = R(s)E(s)$, which means that the rate reward $r(s)$ is the same as the expected reward $R(s)$ accumulated in s weighted by the exit rate $E(s)$. More generally, if the impulse rewards $i(s,s')$ were defined then from $P(s,s') = \frac{Q(s,s')}{E(s)}$ it follows that $R(s)E(s) = \sum_{s' \neq s} i(s,s')Q(s,s') + r(s)$. This means that we can transform the original CTMRM $\mathcal{M}$ with impulse rewards into a CTMRM $\overline{\mathcal{M}} = (S, Q, \overline{r})$ without impulse rewards by integrating the original impulse rewards into a new rate reward

$$\overline{r}(s) := \sum_{s' \neq s} i(s,s')Q(s,s') + r(s).$$

We call $\overline{r}$ the ***continuized*** rate reward since in the continuized model $\overline{\mathcal{M}}$ there is no discrete contribution to the reward accumulation process. As we will see in Theorem 4.1 this (rather heuristically deduced) transformation preserves the finite-horizon total reward measure and thus all the reward measures that are derived from the finite-horizon case.

Figure 4.2 shows a diagram with all the presented transformations and also some relations between them. It is interesting to note that this diagram commutes.

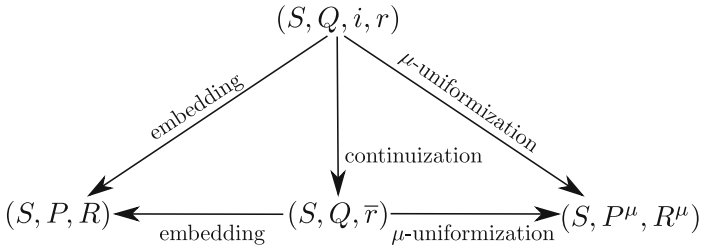

Fig. 4.2. Commuting model transformations

This means that instead of computing the embedded or uniformized DTMRM from the CTMRM (S, Q, i, r) it is possible to continuize the model before performing such a transformation and the resulting DTMRM is the same. We show the commutation of the transformation only for the μ-uniformization, since analogous arguments can be employed for the embedding. When performing the μ-uniformization on (S, Q, i, r) then $R^\mu(s,s') = i(s,s') + \frac{r(s)}{\mu}$ by (4.6). Also denote $\overline{R}^\mu(s)$ as the μ-uniformization of the continuized rate reward $\overline{r}(s)$. Due to the absence of impulse rewards in the continuized model it follows for all $s \in S$ that

$$\overline{R}^\mu(s) = \frac{\overline{r}(s)}{\mu} = \frac{1}{\mu}\left(\sum_{s' \neq s} i(s,s')Q(s,s') + r(s)\right) = \sum_{s' \neq s} i(s,s')P^\mu(s,s') + \frac{1}{\mu}r(s)$$

by definition of P^μ as in (4.5). Furthermore, since $i(s,s) = 0$ it follows

$$\overline{R}^\mu(s) = \sum_{s' \in S} i(s,s) P^\mu(s,s') + \frac{1}{\mu} r(s) = \sum_{s' \in S} R^\mu(s,s') P^\mu(s,s') = R^\mu(s).$$

Thus, the μ-uniformization of the continuized rate reward $\overline{R}^\mu(s)$ is exactly the state-based view on the μ-uniformized reward $R^\mu(s,s')$. Also note that the definition of recurrency and reachability in the discrete-time and continuous-time cases are similar. For this reason the classification of states into closed recurrent classes S_i^r and transient states S^t is invariant under the model transformations, since the directed graph structure of the model does not change.

In the following we are going to provide natural definitions for the value functions of the reward measures that we have also considered in the discrete-time case in Sect. 2. The most important question that we will consider is whether the transformations we have presented in this section are value-preserving. More clearly, let $\mathcal{R}$ be a reward measure with value function V on the CTMRM $\mathcal{M}$. We can also evaluate $\mathcal{R}$ on one of the transformed models, e.g. on $\mathcal{M}^\mu$ which gives a value function V^μ. Under what circumstances is $V = V^\mu$? This question will be answered in the forthcoming sections.

4.4 Total Reward Measure

With all the definitions and tools introduced in the preceding sections we are now set to define the total reward measure. We write $\mathbb{E}_s$ for the expectation operator if $X_0 = s$ (or $Z_0 = s$) is the initial state. For a random variable Y we also write $\mathbb{E}[Y]$ for the function $s \mapsto \mathbb{E}_s[Y] \in \mathbb{R}$, respectively for the vector in $\mathbb{R}^{|S|}$ consisting of the expected values $\mathbb{E}_s[Y]$.

Definition 4.2. *Let $T \in \mathbb{R}$, $T \geq 0$ be some finite real time horizon and N_T the random number of transitions up to time T. The **finite-horizon total value function** is defined as*

$$V_T(s) := \mathbb{E}_s\left[\sum_{k=1}^{N_T} i(X_{k-1}, X_k)\right] + \mathbb{E}_s\left[\int_0^T r(Z_t)\, dt\right], \tag{4.7}$$

*if both expectations exist. If furthermore the expectations $\mathbb{E}_s\left[\sum_{k=1}^{N_T} |i(X_{k-1}, X_k)|\right]$ and $\mathbb{E}_s\left[\int_0^T |r(Z_t)|\, dt\right]$ converge as $T \to \infty$, then we also define the **(infinite-horizon) total value function** as*

$$V_\infty(s) := \lim_{T \to \infty} V_T(s).$$

In (4.7) the rate reward $r(Z_t)$ in state Z_t is continuously accumulated over the time interval $[0,T]$ by integration, whereas the impulse rewards $i(X_{k-1}, X_k)$ for the N_T transitions from states X_{k-1} to X_k for $k = 1, \ldots, N_T$ are discretely accumulated via summation. Note that the upper bound $N_T = \max\{n \mid T_n \leq T\}$

in the summation is random. If $\omega = (s_0, t_0, s_1, t_1, \dots) \in \Omega$ then $N_T(\omega) \in \mathbb{N}$ and the random variable $\sum_{k=1}^{N_T} i(X_{k-1}, X_k)$ takes the value $\sum_{k=1}^{N_T(\omega)} i(s_{k-1}, s_k) \in \mathbb{R}$. Furthermore, N_T has finite expectation (see Lemma A.3 in the Appendix). Since the state space is finite there exists $C \geq 0$ such that $|r(s)| \leq C$ and $|i(s, s')| \leq C$ for all $s, s' \in S$. Therefore

$$\mathbb{E}_s\left[\left|\int_0^T r(Z_t)\,dt\right|\right] \leq C{\cdot}T < \infty \quad \text{and} \quad \mathbb{E}_s\left[\left|\sum_{k=1}^{N_T} i(X_{k-1}, X_k)\right|\right] \leq C{\cdot}\mathbb{E}\left[N_T\right] < \infty$$

such that $V_T(s)$ is defined for all $T \geq 0$. In the prerequisites for the definition of the total value function V_∞ we require a more restrictive absolute convergence. However, this property is quite natural since it is equivalent to the (joint) integrability of the function $r(Z_t) : [0, \infty) \times \Omega \to \mathbb{R}$ with respect to the probability measure P_s on Ω for the expectation $\mathbb{E}_s$ and the Lebesgue measure for the integral over $[0, \infty)$.

Note that $\mathbb{E}_s\left[r(Z_t)\right] = \sum_{s'} P_s(Z_t = s')r(s')$ is the s-th row of the vector $P(t)r$. If we assume that $i(s, s') = 0$ for all $s, s' \in S$ then the finite-horizon total value function $V_T \in \mathbb{R}^S$ regarded as a vector in $\mathbb{R}^{|S|}$ can be computed by

$$V_T = \mathbb{E}\left[\int_0^T r(Z_t)\,dt\right] = \int_0^T P(t)r\,dt. \tag{4.8}$$

The following theorem generalizes this computation for the case with impulse rewards $i(s, s')$. Furthermore, it explains why the continuization transformation as defined in Sect. 4.3.3 preserves the finite-horizon total reward measure. Therefore, this can be considered as the main theorem in the section on CTMRMs.

Theorem 4.1 (Value Preservation of Continuization).
For a CTMRM $\mathcal{M} = (S, Q, i, r)$ let $\overline{\mathcal{M}} = (S, Q, \overline{r})$ be its continuization with

$$\overline{r}(s) = \sum_{s' \neq s} i(s, s')Q(s, s') + r(s).$$

For the finite-horizon total value function it holds that

$$V_T(s) = \mathbb{E}_s\left[\int_0^T \overline{r}(Z_t)\,dt\right].$$

V_T can be computed by

$$V_T = \int_0^T P(t)\overline{r}\,dt,$$

which reads in componentwise notation as

$$V_T(s) = \sum_{s' \in S} \overline{r}(s') \int_0^T P(t)(s, s')\,dt.$$

Proof. In (4.8) we have already shown the statement for the integral term in the definition of V_T in (4.7). It remains to show the statement for the summation term. We have already mentioned that N_T has finite expectation. By Lemma A.1 in the Appendix and the law of total expectation it follows for an arbitrary initial state $s_0 \in S$ that

$$\mathbb{E}_{s_0}\left[\sum_{k=1}^{N_T} i(X_{k-1}, X_k)\right] = \sum_{k=1}^{\infty} \mathbb{E}_{s_0}\left[i(X_{k-1}, X_k)\right] P_{s_0}(N_T \geq k) =$$

$$\sum_{k=1}^{\infty} \sum_{s,s'} i(s,s') P_{s_0}(X_{k-1} = s, X_k = s') P_{s_0}(N_T \geq k) = \sum_{s,s'} i(s,s') n_T(s,s'),$$

where

$$n_T(s,s') := \mathbb{E}_{s_0}\left[\sum_{k=1}^{N_T} \mathbb{1}_{\{X_{k-1}=s, X_k=s'\}}\right] = \sum_{k=1}^{\infty} P_{s_0}(X_{k-1} = s, X_k = s') P_{s_0}(N_T \geq k)$$

is the expected number of transitions from s to s' up to time T from initial state s_0. If we can show that

$$n_T(s,s') = Q(s,s') \cdot \int_0^T P_{s_0}(Z_t = s)\, dt$$

then we are done. The proof for this equation is outsourced to the Appendix. There, in Lemma A.2 we present a proof which uses the uniformization method and in Remark A.2 we sketch a more direct proof without the detour with uniformization which relies on facts from queueing theory. □

Example 4.2. We come back to our WSN node model introduced in Example 4.1 and assume $\lambda = 2$ activities per hour and an average duration of 15 minutes, i.e. $\mu = 4$ and for the sensor $\lambda_s = 4$ and $\mu_s = 30$. Figure 4.3 shows the transient probabilities $P(t)(s_{\text{init}}, s)$ for the initial state $s_{\text{init}} = (\text{silence}, \text{idle})$ and the finite-horizon total value function

$$V_T(s_{\text{init}}) = (1,0,0,0) \int_0^T e^{Qt} \overline{r}\, dt = \tag{4.9}$$

$$\frac{220}{7}T + \frac{10}{49} + \frac{5}{833} e^{-22T} \left(\left(13\sqrt{51} - 17\right) e^{-2\sqrt{51}T} - \left(13\sqrt{51} + 17\right) e^{2\sqrt{51}T}\right)$$

indicating the total energy consumption up to time T. The continuized rate reward is given by

$$\overline{r} = (5\lambda_s, 5\lambda_s, 10 + 5\mu_s, 10)^T = (20, 20, 160, 10)^T. \qquad \square$$

In the following we provide methods for the evaluation of the infinite-horizon total reward measure V_∞. We also show that the model transformations embedding, uniformization and continuization are value-preserving with respect to this

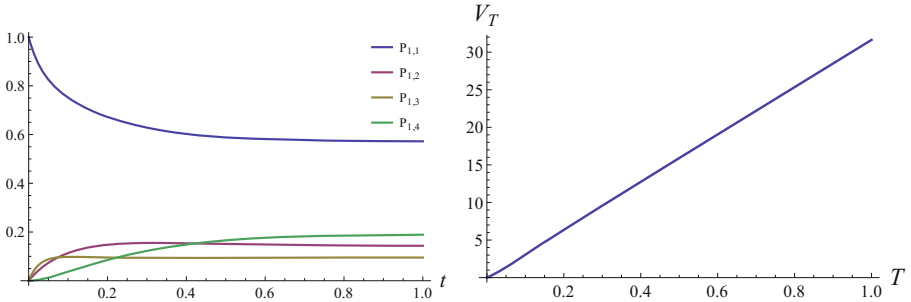

Fig. 4.3. Left: Transient probability functions for initial state $s_{\text{init}} = (\text{silence}, \text{idle})$ converging to the limiting distribution. Right: Total energy consumption during the first hour given by the finite-horizon total value function $V_T(s_{\text{init}})$ as a function of T.

measure. This enables us to provide several methods for the evaluation of V_∞. Before presenting Theorem 4.2, we start with an important proposition about the relation between the existence of the infinite-horizon total value function and the model data Q, i and r of a CTMRM. A proof can be found in the Appendix on page 237.

Proposition 4.1. *For a CTMRM (S, Q, i, r) let $S = \bigcup_{i=1}^{k} S_i^r \cup S^t$ be the partitioning of S into the k closed recurrent classes S_i^r and transient states S^t. The infinite-horizon total value function V_∞ exists if and only if for all $i = 1, \ldots, k$ and for all $s, s' \in S_i^r$ it holds that*

$$r(s) = 0 \quad \textit{and} \quad i(s, s') = 0.$$

Theorem 4.2 (Total Reward Measure – Direct Evaluation and Embedding). *If for a CTMRM (S, Q, i, r) the total value function V_∞ exists, then it is the unique solution to the system of linear equations*

$$\begin{aligned} V_\infty(s) &= R(s) + \sum_{s' \neq s} P(s, s') V_\infty(s') \quad \textit{for } s \in S^t \\ V_\infty(s) &= 0 \qquad\qquad\qquad\qquad\quad \textit{for } s \in S \setminus S^t. \end{aligned} \tag{4.10}$$

Here, $P(s, s')$ are the embedded transition probabilities and $R(s)$ is the state-based embedded reward, i.e. $R(s) = \sum_{s' \in S} R(s, s') P(s, s')$ (see (4.4)). In vector notation this system of equation reads as

$$(I - P) V_\infty = R \quad \textit{with} \quad V_\infty(s) = 0 \ \forall s \in S \setminus S^t.$$

If the impulse reward function i is represented as a matrix with entries $i(s, s')$ then this system of equations can be written in vector notation as

$$-Q V_\infty = \operatorname{diag}(i Q^T) + r, \tag{4.11}$$

where Q^T is the transpose of the matrix Q and $\operatorname{diag}(iQ^T)$ is the diagonal of the matrix iQ^T. This equation represents the direct evaluation of the total reward measure (without performing the embedding).

Proof. In Sect. 4.5 on the discounted reward measure we establish similar equations to (4.10) which involve a discount rate parameter α. By setting α to 0 and noting that all occuring expectations exist (4.10) can be derived analogously. By multiplying (4.10) with $E(s)$ and rearranging terms one can directly deduce (4.11). The uniqueness holds since the $(|S^t| \times |S^t|)$-submatrix of Q with entries $Q(s, s')$ for transient states $s, s' \in S^t$ has full rank. □

Corollary 4.1 (Total Reward Measure – Continuization).
Let $\mathcal{M} = (S, Q, i, r)$ be a CTMRM and $\overline{\mathcal{M}} = (S, Q, \overline{r})$ its continuization. If the total value function V_∞ for $\mathcal{M}$ exists then it also exists for $\overline{\mathcal{M}}$ and in this case they are equal, i.e. V_∞ is the unique solution to the system of linear equations

$$QV_\infty = -\overline{r} \tag{4.12}$$

with $V_\infty(s) = 0$ for all recurrent states s.

Proof. Let $S_i^r \subseteq S$ be a closed recurrent class of S and consider a recurrent state $s \in S_i^r$. If V_∞ exists for $\mathcal{M}$, then by Proposition 4.1 it holds that $r(s) = 0$ and $i(s, s') = 0$ for all $s' \in S_i^r$ in the same recurrent class. Furthermore, if $s' \in S \setminus S_i^r$ then $Q(s, s') = 0$ and therefore $\overline{r}(s) = \sum_{s' \neq s} i(s, s')Q(s, s') + r(s) = 0$. Thus, the total value function for $\overline{\mathcal{M}}$ denoted by $\overline{V}_\infty$ is also defined and it solves (4.11) which reads as $-Q\overline{V}_\infty = \overline{r}$. In order to show that $V_\infty = \overline{V}_\infty$ note that the s-th diagonal entry of the matrix iQ^T is $\sum_{s' \in S} i(s, s')Q(s, s') = \sum_{s' \neq s} i(s, s')Q(s, s')$ since $i(s, s) = 0$. Therefore, $\mathrm{diag}(iQ^T) + r = \overline{r}$ is the continuized rate reward and the conclusion follows since both V_∞ and $\overline{V}_\infty$ solve $-QX = \overline{r}$ and the solution is unique (with the property that both are 0 on recurrent states). □

Corollary 4.2 (Total Reward Measure – Uniformization).
Let $\mathcal{M} = (S, Q, i, r)$ be a CTMRM and $\mathcal{M}^\mu = (S, P^\mu, R^\mu)$ the μ-uniformized DTMRM. If the total value function V_∞ for $\mathcal{M}$ exists then it also exists for $\mathcal{M}^\mu$ and in this case they are equal, i.e. V_∞ is the unique solution to

$$(I - P^\mu)V_\infty = R^\mu$$

with $V_\infty(s) = 0$ for all recurrent states s.

Proof. From (4.5) and (4.6) it holds that $R^\mu(s, s') = i(s, s') + \frac{r(s)}{\mu}$ and $P^\mu = I + \frac{1}{\mu}Q$. If s and s' are communicating recurrent states (i.e. in the same closed recurrent class) then $r(s) = 0$ and $i(s, s') = 0$ by Proposition 4.1 and therefore $R^\mu(s, s') = 0$. If V_∞^μ denotes the total value function for the μ-uniformized model $\mathcal{M}^\mu$ then V_∞^μ exists by Proposition 2.1 since $R^\mu(s, s') = 0$ for all states s and s' in the same closed recurrent class and by Theorem 2.1 V_∞^μ is also a solution of $(I - P^\mu)V_\infty^\mu = R^\mu$. It follows from (4.12) that

$$(I - P^\mu)V_\infty = -\frac{1}{\mu}QV_\infty = \frac{1}{\mu}\overline{r} = R^\mu$$

and since V_∞ and V_∞^μ are 0 on recurrent states, it follows that $V_\infty = V_\infty^\mu$. □

4.5 Horizon-Expected and Discounted Reward Measure

In analogy to Sect. 2.3 we want to introduce the discounted reward measure in the continuous-time case. This reward measure can be formally deduced from the horizon-expected reward measure, which we are going to define first.

Definition 4.3. *Let $\mathcal{M} = (S, Q, i, r)$ be a CTMRM and consider a random horizon length T for $\mathcal{M}$, i.e. T is a non-negative continuous random variable that is independent of the state process Z_t of $\mathcal{M}$. Let $V_{(T)}$ denote the random finite-horizon total value function that takes values in $\{V_t \in \mathbb{R}^S \mid t \in [0, \infty)\}$. Define the* ***horizon-expected value function*** *as*

$$V(s) := \mathbb{E}\left[V_{(T)}(s)\right],$$

if the expectation exists for all $s \in S$, i.e. $|V_{(T)}(s)|$ has finite expectation.

The random variable $V_{(T)}(s)$ can be regarded as the conditional expectation

$$V_{(T)}(s) = \mathbb{E}_s\left[\sum_{k=1}^{N_T} i(X_{k-1}, X_k) + \int_0^T r(Z_t)\,dt \mid T\right] = \mathbb{E}_s\left[\int_0^T \overline{r}(Z_t)\,dt \mid T\right]$$

that takes the value $V_t(s)$ if $T = t$. Let P_T denote the probability measure of T. Due to the law of total expectation the horizon-expected value function $V(s)$ is the joint expectation with respect to the probability measures P_T of T and P_s of all the Z_t, i.e.

$$V(s) = \mathbb{E}\left[\mathbb{E}_s\left[\int_0^T \overline{r}(Z_t)\,dt \mid T\right]\right] = \mathbb{E}_s\left[\int_0^T \overline{r}(Z_t)\,dt\right], \tag{4.13}$$

where $\mathbb{E}_s$ on the right hand side denotes the joint expectation.

Lemma 4.1. *Let T be a random horizon length with $\mathbb{E}[T] < \infty$ and probability measure P_T. Then the horizon-expected value function $V(s)$ exists and is given by*

$$V(s) = \mathbb{E}_s\left[\int_0^\infty \overline{r}(Z_t) P_T(T \geq t)\,dt\right] = \mathbb{E}_s\left[\sum_{n=0}^{\infty} \overline{r}(X_n) \int_{T_n}^{T_{n+1}} P_T(T \geq t)\,dt\right].$$

The proof can be found in the Appendix on page 237. Note that $V(s)$ can also be represented directly in terms of the impulse reward i and rate reward r (instead of the continuized rate reward $\overline{r}$) as

$$V(s) = \mathbb{E}_s\left[\sum_{n=0}^{\infty}\left(i(X_n, X_{n+1}) \cdot P_T(T \geq T_{n+1}) + r(X_n) \int_{T_n}^{T_{n+1}} P_T(T \geq t)\,dt\right)\right].$$

In this equation, one can also see that an impulse reward $i(X_n, X_{n+1})$ for the $(n+1)$-st transition is only accumulated if the time horizon T is not exceeded by the total elapsed time T_{n+1} up to this transition.

Definition 4.4. *Let the horizon length T be exponentially distributed with rate $\alpha > 0$. In this case the horizon-expected reward measure is called* ***discounted reward measure*** *with* ***discount rate*** *α (or just α-discounted reward measure) and its value function will be denoted by $V^\alpha(s)$.*

The discounted value function V^α represented as a vector in $\mathbb{R}^{|S|}$ is given by

$$V^\alpha = \int_0^\infty e^{-\alpha t} P(t)\overline{r}\, dt. \tag{4.14}$$

This follows directly from Lemma 4.1 together with $P_T(T \geq t) = e^{-\alpha t}$ and $\mathbb{E}[\overline{r}(Z_t)] = P(t)\overline{r}$. As in Sect. 2.3 in the discrete-time setting we can also derive a system of linear equations which allows to compute the discounted value function V^α. The proof can be found in the Appendix on page 238.

Theorem 4.3 (Discounted Reward Measure – Continuization). *The discounted value function with discount rate $\alpha > 0$ is the unique solution to the system of linear equations*

$$V^\alpha(s) = \frac{\overline{r}(s)}{\alpha + E(s)} + \sum_{s' \neq s} \frac{Q(s,s')}{\alpha + E(s)} V^\alpha(s'). \tag{4.15}$$

In vector notation this system of equations reads as

$$(Q - \alpha I)V^\alpha = -\overline{r}. \tag{4.16}$$

Note that in case the total value function V_∞ exists (and is thus finite) it is the limit of the α-discounted value function as α decreases to 0, i.e. for all $s \in S$ it holds that

$$V_\infty(s) = \lim_{\alpha \searrow 0} V^\alpha(s). \tag{4.17}$$

Example 4.3. Figure 4.4 shows the α-discounted value function $V^\alpha(s_{\text{init}})$ for the initial state $s_{\text{init}} = (\text{silence}, \text{idle})$ of the WSN node model from Example 4.1 dependent on the discount rate α. By solving (4.16) we get

$$V^\alpha(s_{\text{init}}) = \frac{20\left(\alpha^2 + 72\alpha + 440\right)}{\alpha\left(\alpha^2 + 44\alpha + 280\right)}.$$

Clearly, for increasing α the expected horizon length $\mathbb{E}[T] = \frac{1}{\alpha}$ decreases and thus the discounted value representing the expected energy consumption up to time T also decreases. On the other hand, if α decreases towards 0, then the discounted value increases and in our case it diverges to ∞. Note that the total value function V_∞ does not exist for this model. □

Remember that one of our main goals in this section is to check, whether all the model transformations in Fig. 4.2 are value-preserving. For a CTMRM $(S, Q, \overline{r})$ with α-discounted value function V^α consider its μ-uniformization (S, P^μ, R^μ)

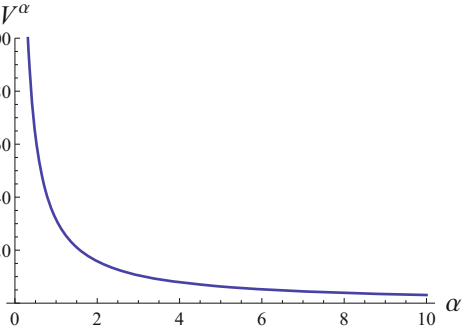

Fig. 4.4. The discounted value $V^\alpha(s_{\text{init}})$ for the initial state $s_{\text{init}} = (\text{silence}, \text{idle})$ as a function of the discount rate α

with γ-discounted value function V^γ. We show that there is no choice of $\gamma \in (0,1)$ such that $V^\alpha = V^\gamma$. Thus the μ-uniformization is not value-preserving with respect to the discounted reward measure. (As a special case it also follows, that the embedding is not value-preserving as well.)
Assume that there exists $\gamma \in (0,1)$ such that $V^\alpha = V^\gamma$. On the one hand V^α is the unique solution to $(Q - \alpha I)V^\alpha = -\overline{r}$ and thus $QV^\gamma = \alpha V^\gamma - \overline{r}$. On the other hand, V^γ is by Theorem 2.2 the unique solution to $(I - \gamma P^\mu)V^\gamma = R^\mu$ where $P^\mu = I + \frac{1}{\mu}Q$ and $R^\mu = \frac{1}{\mu}\overline{r}$. Thus

$$R^\mu = (I - \gamma P^\mu)\, V^\gamma = \left(I - \gamma\left(I + \frac{1}{\mu}Q\right)\right) V^\gamma = (1-\gamma)\, V^\gamma - \gamma\frac{1}{\mu}\left(\alpha V^\gamma - \overline{r}\right).$$

By rearranging terms it follows that

$$\left(1 - \frac{\gamma\alpha}{(1-\gamma)\mu}\right) V^\gamma = R^\mu,$$

which means that V^γ is a multiple of R^μ and is thus independent of the transition probabilities P^μ! However, we can save the value-preserving property by observing the following link between the discounted and the total reward measure in analogy to the discrete-time case as shown in Remark 2.1.

Remark 4.3. If $\mathcal{M} = (S, Q, i, r)$ is a CTMRM then extend $\mathcal{M}$ to a CTMRM $\mathcal{M}' = (S', Q', i', r')$ with an artificial absorbing reward-free state *abs* that is reachable with rate $\alpha > 0$ from every other state in S, i.e.

$$S' := S \cup \{abs\}, \quad Q' := \begin{pmatrix} Q - \alpha I & \alpha\mathbf{1} \\ 0 & 0 \end{pmatrix}, \quad i' := \begin{pmatrix} i & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad r' := \begin{pmatrix} r \\ 0 \end{pmatrix}.$$

Since *abs* is the single recurrent state in $\mathcal{M}'$ it follows that the total value function V'_∞ for $\mathcal{M}'$ with $V'_\infty(abs) = 0$ is a solution to (4.12), i.e.

$$Q'V'_\infty = -\overline{r'},$$

where $\overline{r'}$ is the continuized rate reward of $\mathcal{M}'$. By definition of $\mathcal{M}'$ it holds for all $s \in S' \setminus \{abs\} = S$ that $i'(s, abs) = 0$ and $r'(s) = r(s)$ and it follows that

$$\overline{r'}(s) = \sum_{\substack{s' \in S' \\ s' \neq s}} i'(s,s')Q'(s,s') + r'(s) = \sum_{\substack{s' \in S \\ s' \neq s}} i(s,s')Q(s,s') + r(s) = \overline{r}(s).$$

Since V^α is the α-discounted value function for $\mathcal{M}$ it is the unique solution to $(Q - \alpha I)V^\alpha = -\overline{r}$ and thus

$$Q' \begin{pmatrix} V^\alpha \\ 0 \end{pmatrix} = \begin{pmatrix} Q - \alpha I & \alpha \mathbf{1} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V^\alpha \\ 0 \end{pmatrix} = - \begin{pmatrix} \overline{r} \\ 0 \end{pmatrix} = -\overline{r'}.$$

Since V'_∞ is also a solution to $Q'V'_\infty = -\overline{r'}$ and also unique with the property $V'_\infty(abs) = 0$ it follows that $V^\alpha(s) = V'_\infty(s)$ for all $s \in S$.

This remark allows to provide a further method for the evaluation of the discounted value function by means of uniformization. Note that if $\mu \geq E(s)$ for all $s \in S$ is a uniformization rate for the original model $\mathcal{M}$ then $\mu + \alpha$ is a uniformization rate for the extended model $\mathcal{M}' = (S', Q', i', r')$. The following theorem states that the rates and the rewards have to be uniformized differently in order to be able to establish a connection between the α-discounted value function and a γ-discounted value function for some suitable DTMRM. For this reason, we refer to the transformation to that DTMRM as ***separate uniformization***.

Theorem 4.4 (Discounted Reward Measure – separate Uniformization). *Let $\mathcal{M} = (S, Q, i, r)$ be a CTMRM, $\mu > 0$ a uniformization rate for $\mathcal{M}$ and $\alpha > 0$ a discount rate. Then V^α is the unique solution to the system of linear equations*

$$(I - \gamma P^\mu)\, V^\alpha = R^{\mu+\alpha},$$

where $P^\mu = I + \frac{1}{\mu}Q$ is the μ-uniformized transition probability matrix, $R^{\mu+\alpha} = \frac{1}{\mu+\alpha}\overline{r}$ is the $(\mu+\alpha)$-uniformized reward vector and $\gamma = \frac{\mu}{\mu+\alpha} \in (0,1)$ is a discount factor. In other words the α-discounted value function V^α for the CTMRM $\mathcal{M}$ is precisely the γ-discounted value function for the DTMRM $\widetilde{\mathcal{M}} := (S, P^\mu, R^{\mu+\alpha})$ denoted by $\widetilde{V}^\gamma$, i.e.

$$V^\alpha = \widetilde{V}^\gamma.$$

The proof is straightforward and integrated in the following discussion on several relationships between models and value functions that can occur by the model transformations. Figure 4.5 shows transformations between Markov chains without rewards. A CTMC (S, Q) is uniformized into a DTMC (S, P^μ) and afterwards the model is extended with an auxiliary absorbing state *abs* as described in Remark 2.1 which leads to a DTMC $(S', (P^\mu)')$ with $S' = S \cup \{abs\}$ (γ-extension). On the other hand, (S, Q) can be directly extended with *abs* as described in Remark 4.3 to the model (S', Q') and then uniformized with rate $\mu + \alpha$ (α-extension). This diagram commutes, since

$$(P^\mu)' = \begin{pmatrix} \gamma P^\mu & (1-\gamma)\mathbf{1} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{\mu}{\mu+\alpha}(I + \frac{1}{\mu}Q) & \frac{\alpha}{\mu+\alpha}\mathbf{1} \\ 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} I & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{\mu+\alpha} \begin{pmatrix} Q - \alpha I & \alpha\mathbf{1} \\ 0 & 0 \end{pmatrix} = (P')^{\mu+\alpha}.$$

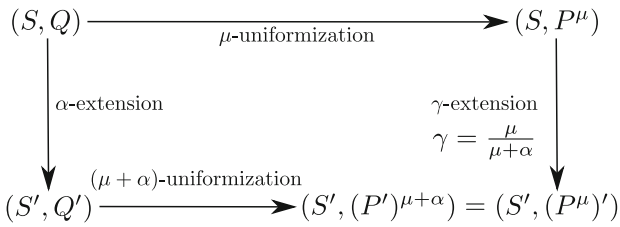

Fig. 4.5. Commuting model transformations on discrete-time and continuous-time Markov chains

In contrast, Fig. 4.6 shows the same transformations applied to (continuized) Markov reward models. This diagram does not commute, since in general

$$(R^\mu)' = \frac{1}{\mu}\begin{pmatrix} \overline{r} \\ 0 \end{pmatrix} \neq \frac{1}{\mu+\alpha}\begin{pmatrix} \overline{r} \\ 0 \end{pmatrix} = (R')^{\mu+\alpha}.$$

However, due to $(P')^{\mu+\alpha} = (P^\mu)'$ it is possible to compute the infinite-horizon total value function V_∞ on the DTMRM $(S', (P')^{\mu+\alpha}, (R')^{\mu+\alpha})$. Let us call its restriction on S as $\widetilde{V}^\gamma$. Since the uniformization is value-preserving with respect to the infinite-horizon total reward measure (see Corollary 4.2) and due to Remark 4.3 it follows that $V^\alpha = \widetilde{V}^\gamma$, which concludes the proof of Theorem 4.4.

4.6 Average Reward Measure

In Sect. 2.4 we defined the discrete-time average reward by considering a sequence of finite-horizon value functions V_N which were averaged over the horizon length N and the limit as $N \to \infty$ was considered. In complete analogy we define the average reward in the continuous-time case.

Definition 4.5. *Let $\mathcal{M} = (S, Q, i, r)$ be a CTMRM with finite-horizon total value function V_T. The* ***average reward*** *value function is defined as*

$$g(s) = \lim_{T\to\infty} \frac{1}{T} V_T(s),$$

if the limit exists for all $s \in S$.

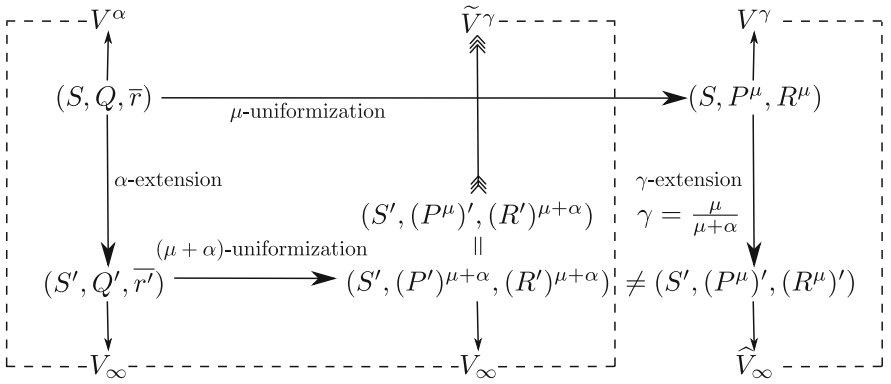

Fig. 4.6. Model transformations (big arrows) on discrete-time and continuous-time reward models that do not commute. A small arrow indicates an evaluation of a reward measure on a model. The dashed lines connect value functions that are related by equality. The value function $\widetilde{V}^\gamma$ is not directly evaluated on $(S', (P^\mu)', (R')^{\mu+\alpha})$ but is induced by V_∞ (feathered arrow) as a restriction from S' to S and it holds $V^\alpha = \widetilde{V}^\gamma$.

Example 4.4. In the WSN node model from Example 4.1 we saw in (4.9) that $V_T(s_{\text{init}}) = \frac{220}{7}T + f(T)$ with some function $f(T)$ such that $\frac{f(T)}{T} \to 0$ as $T \to \infty$. This result means, that on average over infinite time the energy consumption is $g(s) = \frac{220}{7}$ per hour (compare this with the slope of $V_T(s_{\text{init}})$ in Fig. 4.3). □

In the following we want to provide methods for the computation of the average reward that do not rely on an explicit representation of V_T which is computed by integration over the transient probability matrix $P(t) = e^{Qt}$ as in Theorem 4.1. In Sect. 4.2.3 we mentioned that $P(t)$ converges to the limiting matrix P^*. Remind that P^* fulfills the properties

$$P(t)P^* = P^*P(t) = P^*P^* = P^* \quad \text{and} \quad P^*Q = QP^* = 0.$$

Proposition 4.2. *Let $\overline{r}$ be the continuized rate reward of a CTMRM (S, Q, i, r). Then the average reward can be computed by*

$$g = P^*\overline{r}.$$

Proof. By Theorem 4.1 it holds that

$$g = \lim_{T\to\infty} \frac{1}{T} V_T = \lim_{T\to\infty} \frac{1}{T} \int_0^T P(t)\overline{r}\, dt.$$

Fix two states s and s' and consider the monotonically increasing function $h(T) := \int_0^T P(t)(s, s')\, dt \geq 0$. If $h(T)$ is unbounded it follows by the rule of l'Hospital that

$$\lim_{T\to\infty} \frac{h(T)}{T} = \lim_{T\to\infty} P(T)(s, s') = P^*(s, s').$$

In the other case if $h(T)$ is bounded then clearly $P(t)(s,s')$ converges to 0. But this is only the case if either s and s' are in different closed recurrent classes or s' is transient and in both cases it holds that $P^*(s,s') = 0$. Thus from

$$\lim_{T\to\infty} \frac{h(T)}{T} = 0 = P^*(s,s')$$

the conclusion follows. □

As in Sect. 2.4 we also show another possibility to evaluate the average reward which does not rely on the computation of P^* and will be used in the subsequent section on CTMDPs. For this reason we define the notion of a deviation matrix H and a bias h in the continuous-time case.

Definition 4.6. *For a CTMRM $\mathcal{M} = (S, Q, i, r)$ define the **deviation matrix** H as*

$$H := \int_0^\infty (P(t) - P^*)\, dt,$$

where integration is performed componentwise. Further define

$$h := H\overline{r} = \int_0^\infty (P(t)\overline{r} - g)\, dt$$

*as the **bias** of $\mathcal{M}$.*

Note that Q, H and P^* satisfy the following equations:

$$QH = HQ, \quad P^* = I + QH \quad \text{and} \quad HP^* = P^*H = 0. \tag{4.18}$$

That can be easily derived by the Kolmogorov equations (4.2).
In the following, we connect the discounted and the average reward measures. Consider for a fixed $s \in S$ the discounted value $V^\alpha(s)$ as a function of $\alpha \geq 0$. Then $V^\alpha(s)$ might have a pole at $\alpha = 0$ and can be extended as a Laurent series in α. For more information on the Laurent series expansion in continuous time we refer to Theorem A.1 in the Appendix. This theorem directly induces the Laurent series decomposition of the α-discounted value function as is stated in the following corollary.

Corollary 4.3 (Laurent Series of the Discounted Value Function). *The Laurent series expansion of V^α is given by*

$$V^\alpha = \alpha^{-1} g + \sum_{n=0}^{\infty} \alpha^n H^{n+1} \overline{r}.$$

Recall (4.17): In case the infinite-horizon total value V_∞ exists it follows for the average reward g and the bias h from the Laurent expansion for $\alpha \to 0$ that $g = 0$ and $h = V_\infty$. Thus on average no reward is gained over the infinite horizon which can also be seen by Proposition 4.1 since there are no rewards in recurrent states. By Definition 4.6 the bias h measures the total long-term deviation of the

accumulated rewards from the average reward, i.e. $h = \lim_{T\to\infty}(V_T - g \cdot T)$. As in the discrete-time setting, the bias can also be seen as the excess of rewards $\overline{r}$ until the system reaches its steady-state. Moreover, h is also the infinite-horizon total value function for the CTMRM with average-corrected rewards $\overline{r} - g$ (if we also allow for non-absolute convergence in Definition 4.2). Thus, if $g = 0$ it follows that $h = V_\infty$.

Example 4.5. If we decompose the rational function for $V^\alpha(s_{\text{init}})$ in Example 4.3 into a Laurent series at $\alpha = 0$ then

$$V^\alpha(s_{\text{init}}) = \frac{20\left(\alpha^2 + 72\alpha + 440\right)}{\alpha\left(\alpha^2 + 44\alpha + 280\right)} = \frac{220}{7\alpha} + \frac{10}{49} - \frac{25\alpha}{343} + \frac{103\alpha^2}{9604} + O\left(\alpha^3\right).$$

We see that the average reward g for s_{init} is $\frac{220}{7}$ and the bias h is $\frac{10}{49}$. Compare these values also with (4.9). □

In the following we show two possibilities to compute the average reward by a system of linear equations. The first is a direct evaluation which uses the CTMRM model data Q, r and i and the second system of linear equations relies on the uniformized DTMRM.

Theorem 4.5 (Average Reward Measure – Direct Evaluation). *The average reward g and the bias h fulfill the following system of linear equations:*

$$\begin{pmatrix} -Q & 0 \\ I & -Q \end{pmatrix} \begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} 0 \\ \overline{r} \end{pmatrix}. \tag{4.19}$$

Furthermore, a solution (u, v) to this equation implies that $u = P^*\overline{r} = g$ is the average reward and there exists $w \in \ker(I - P^*)$ such that $v - w = h$ is the bias.

Proof. Let $g = P^*\overline{r}$ be the average reward and $h = H\overline{r}$ the bias. From $QP^* = 0$ it follows that $Qg = 0$ and by using the Kolmogorov equations (4.2) it holds that

$$Qh = Q\int_0^\infty (P(t) - P^*)\overline{r}\,dt = \int_0^\infty P'(t)\overline{r}\,dt = (P^* - I)\overline{r} = g - \overline{r}$$

and thus (4.19) follows. Now let (u, v) be a solution to (4.19). Then clearly $0 = Qu = P(t)Qu = P'(t)u$ and by integrating $\int_0^t P'(\tau)u\ d\tau = 0$ and using $P(0) = I$ it follows that $P(t)u = u$ for all $t \geq 0$. Therefore, if $t \to \infty$ together with $u = \overline{r} + Qv$ and $P^*Q = 0$ it follows $u = P^*u = P^*(\overline{r} + Qv) = P^*\overline{r} = g$. Now

$$\begin{aligned}
&(I - P^*)v = -\int_0^\infty P'(t)v\,dt && P(0) = I \\
&= -\int_0^\infty P(t)Qv\,dt = -\int_0^\infty P(t)(g - \overline{r})\,dt && (4.2), u - Qv = \overline{r}, u = g \\
&= \int_0^\infty (P(t)\overline{r} - g)\,dt = \int_0^\infty (P(t) - P^*)\overline{r}\,dt && P(t)g = g\ \forall t \geq 0 \\
&= H\overline{r} = (H - P^*H)\overline{r} = (I - P^*)h. && (4.18), h = H\overline{r}
\end{aligned}$$

Therefore, $v = h + w$ for some $w \in \ker(I - P^*)$. □

In the special case, if $\mathcal{M}$ is unichain then $\ker(Q) = \mathbf{1}\mathbb{R}$ is one-dimensional and therefore $g = g_0\mathbf{1} \in \ker(Q)$ is constant with $g_0 \in \mathbb{R}$. This value can be computed by finding a solution to $g_0\mathbf{1} - Qh = \overline{r}$. Alternatively, in a unichain CTMRM the unique stationary distribution ρ fulfills $\rho Q = 0$ and $\rho\mathbf{1} = 1$ and thus $g_0 = \rho\overline{r}$.

Theorem 4.6 (Average Reward Measure – Uniformization). *Consider a CTMRM $\mathcal{M} = (S, Q, i, r)$ with average reward g and let μ be a uniformization rate. Then g is the unique solution to the system of equations*

$$\begin{pmatrix} I - P^\mu & 0 \\ \mu I & I - P^\mu \end{pmatrix} \begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} 0 \\ R^\mu \end{pmatrix}.$$

If $\mathcal{M}^\mu = (S, P^\mu, R^\mu)$ is the μ-uniformized model and g^μ the average reward of the DTMRM $\mathcal{M}^\mu$ then

$$g = \mu g^\mu.$$

The statement of this theorem can be interpreted as follows: In the continuous-time model $g(s)$ is the average reward per time from initial state s, while in the corresponding μ-uniformized discrete-time model $g^\mu(s)$ is the average reward per transition. In the uniformized model the expected number of transitions per time unit is exactly the rate μ (which corresponds to Little's law) and thus $g(s) = \mu g^\mu(s)$. Note also that one can assume without loss of generality that all exit rates $E(s)$ satisfy $E(s) \leq 1$ by changing the time scale. In this case, one can choose $\mu := 1$ and it follows that $g = g^\mu$. For this reason, the uniformization transformation (with $\mu = 1$ expected number of transitions per time unit) preserves the average reward measure.

Proof. We first show that $g = \mu g^\mu$. Theorem 4.4 allows to link for each discount rate $\alpha > 0$ the α-discounted continuous-time value V^α to the γ-discounted discrete-time value $\widetilde{V}^\gamma$ of the separate uniformized DTMRM $(S, P^\mu, R^{\mu+\alpha})$ with discount factor $\gamma = \frac{\mu}{\mu+\alpha}$. From the continuous-time Laurent series in Corollary 4.3 it follows that $g = \lim_{\alpha\to 0} \alpha V^\alpha$. On the other hand, since $\lim_{\alpha\to 0} R^{\mu+\alpha} = R^\mu$ it follows from the discrete-time Laurent series in Theorem 2.3 that $g^\mu = \lim_{\rho\to 0} \frac{\rho}{1+\rho}\widetilde{V}^\gamma$, where $\rho = \frac{1-\gamma}{\gamma} = \frac{\alpha}{\mu}$. Combining both gives

$$g^\mu = \lim_{\rho\to 0} \frac{\rho}{1+\rho}\widetilde{V}^\gamma = \lim_{\alpha\to 0} \frac{\alpha}{\mu+\alpha} V^\alpha = \frac{1}{\mu} g$$

and the conclusion follows. The system of the linear equations can be directly established from Theorem 2.4 with $P^\mu = I + \frac{1}{\mu}Q$ and $R^\mu = \frac{1}{\mu}\overline{r}$. □

4.7 Big Picture – Model Transformations

We summarize all the transformations and evaluation methods presented in this section in Fig. 4.7. Theorem 4.1 allows to continuize a CTMRM (S, Q, i, r) into a CTMRM $(S, Q, \overline{r})$ and hereby preserving all considered value functions. For

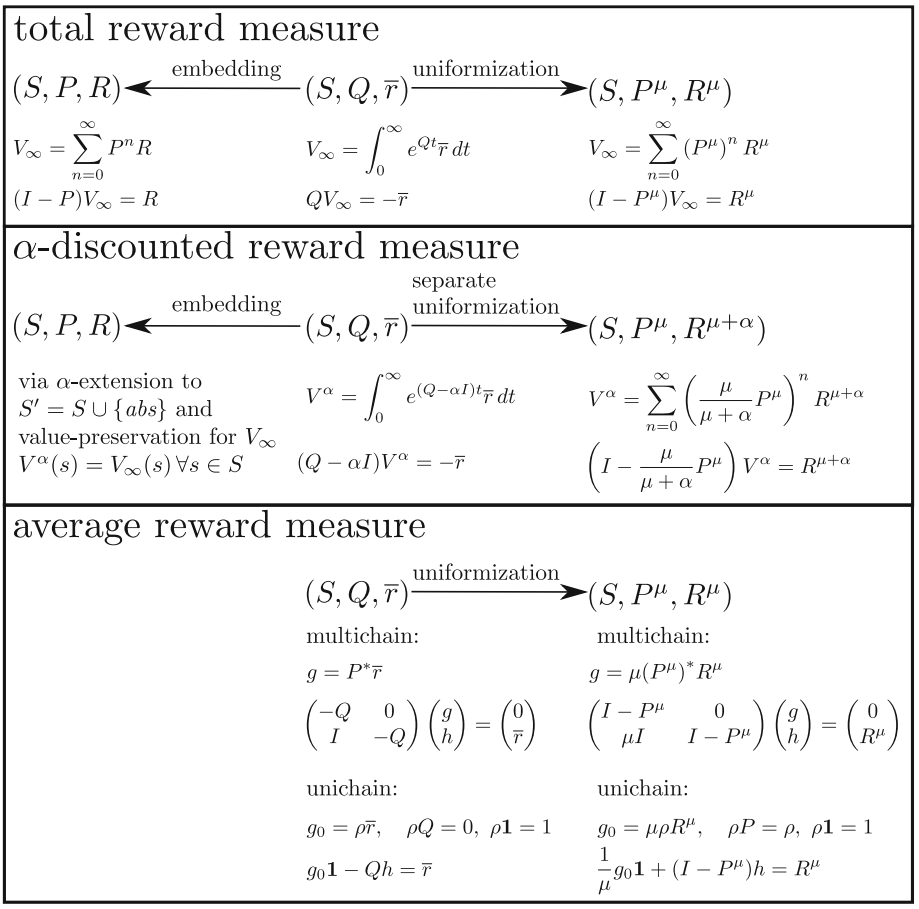

Fig. 4.7. Big Picture: Value-preserving transformations from the continuization $(S, Q, \overline{r})$ of a CTMRM (S, Q, i, r)

this reason, we omit the model (S, Q, i, r) in the figure. The embedded DTMRM (S, P, R) is defined by

$$P = I + E^{-1}Q \quad \text{and} \quad R = \text{diag}(iP^T) + E^{-1}r \in \mathbb{R}^{|S|},$$

where E^{-1} is defined as a diagonal matrix with entries $\frac{1}{E(s)}$ if $E(s) \neq 0$ and 0 otherwise. The vector $\text{diag}(iP^T)$ is the state-based view on the impulse rewards $i(s, s')$ collected in a matrix i. The μ-uniformized DTMRM (S, P^μ, R^μ) is defined by

$$P^\mu = I + \frac{1}{\mu}Q \quad \text{and} \quad R^\mu = \text{diag}\left(i\,(P^\mu)^T\right) + \frac{1}{\mu}r \in \mathbb{R}^{|S|}.$$

The total reward measure is value-preserving for both transformations embedding and uniformization. Therefore, all presented methods for computation of

V_∞ in continuous and discrete time can be used. In order to transform the discounted reward measure with discount rate α we need to consider an extended model (see Remark 4.3). The evaluation of the total reward measure on the extended model is equivalent to the evaluation of the discounted reward measure on the original model. For the average reward model, there is in general no simple direct method to compute the average reward g via embedding, since continuous time and the transition-counting time are not compatible when building averages over time.

We want to conclude this section with a remark on more general reward structures. Beyond impulse rewards or rate rewards as we defined, the authors of [22], [23] and [35] also analyze rewards that can vary over time. This variation can be homogeneous (depending on the length of a time interval) or non-homogeneous (depending on two points in time). These reward structures are mostly accompanied by the more general model class of Semi-Markov Reward Processes. Furthermore, [30] defines path-based rewards which can be analyzed by augmenting the model with special reward variables, such that the state space does not need to be extended for path information.

5 Continuous Time Markov Decision Processes

In this section we merge both model types MDP and CTMRM together into a CTMDP model. This section is rather short, because all of the necessary work has been already done in the preceding sections. For this reason, we establish connections to the previous results. Moreover, we also present an additional method for the computation of the average reward which directly works on CTMDPs.

5.1 Preliminaries and Retrospection

Definition 5.1. *A **continuous-time Markov Decision Process (CTMDP)** is a structure $\mathcal{M} = (S, Act, e, Q, i, r)$, where S is a finite state space, Act a finite set of actions, $e\colon S \to 2^{Act} \setminus \emptyset$ the action-enabling function, $Q\colon S \times Act \times S \to \mathbb{R}$ an action-dependent generator function, $i\colon S \times Act \times S \to \mathbb{R}$ the action-dependent impulse reward function with $i(s,a,s) = 0$ for all $a \in e(s)$ and $r\colon S \times Act \to \mathbb{R}$ the action-dependent rate reward function.*

Completely analogous to Sect. 3 we define the set of policies

$$\Pi := \{\pi\colon S \to Act \mid \pi(s) \in e(s)\}\,.$$

Applying π to a CTMDP $\mathcal{M}$ induces a CTMRM $\mathcal{M}^\pi = (S, Q^\pi, i^\pi, r^\pi)$, where

$$Q^\pi(s,s') := Q(s,\pi(s),s'), \quad i^\pi(s,s') := i(s,\pi(s),s') \quad \text{and} \quad r^\pi(s) := r(s,\pi(s)).$$

A reward measure $\mathcal{R}$ for the CTMDP $\mathcal{M}$ induces for each policy π a value V^π for $\mathcal{M}^\pi$.

Definition 5.2. *Let $\mathcal{M}$ be a CTMDP with reward measure $\mathcal{R}$ and for each $\pi \in \Pi$ let V^π be the value of π with respect to $\mathcal{R}$. The value V^* of $\mathcal{M}$ is defined as*

$$V^*(s) := \sup_{\pi \in \Pi} V^\pi(s).$$

A policy $\pi^* \in \Pi$ is called optimal if

$$\forall s \in S \; \forall \pi \in \Pi : \; V^{\pi^*}(s) \geq V^\pi(s).$$

In order to optimize the CTMDP we can transform $\mathcal{M}$ by embedding or uniformization into an MDP and by continuization into another CTMDP. The transformations follow the Big Picture as presented in Sect. 4.7 (Fig. 4.7) with the difference that all action-dependent quantities (i.e. Q, i and r) are transformed in an action-wise manner. The following theorem states that these transformations preserve both the optimal value and the optimal policies.

Theorem 5.1. *Let $\mathcal{M}$ be a CTMDP with policy space Π, optimal value V^* and a set of optimal policies $\Pi^* \subseteq \Pi$. Further let $\widehat{\mathcal{M}}$ be a transformed model (MDP or CTMDP) as in Fig. 4.7 with policy space $\widehat{\Pi}$, value $\widehat{V}^*$ and optimal policies $\widehat{\Pi}^* \subseteq \widehat{\Pi}$. Then*

$$V^* = \widehat{V}^* \quad \textit{and} \quad \Pi^* = \widehat{\Pi}^*.$$

Proof. Note that V^* and $\widehat{V}^*$ are defined over policies, i.e.

$$V^* = \sup_{\pi \in \Pi} V^\pi \quad \text{and} \quad \widehat{V}^* = \sup_{\pi \in \widehat{\Pi}} \widehat{V}^\pi.$$

All the transformations in Fig. 4.7 do not transform S, *Act* and e, thus $\Pi = \widehat{\Pi}$. Furthermore, for each $\pi \in \Pi$ the transformations preserve the value V^π, i.e. $V^\pi = \widehat{V}^\pi$ and thus $V^* = \sup_{\pi \in \Pi} V^\pi = \sup_{\pi \in \widehat{\Pi}} \widehat{V}^\pi = \widehat{V}^*$. In order to show that $\Pi^* = \widehat{\Pi}^*$ let $\pi^* \in \Pi^*$. Then for all s and for all $\pi \in \Pi$ by definition of π^* it holds that

$$V^{\pi^*}(s) \geq V^\pi(s) = \widehat{V}^\pi(s) \quad \text{and} \quad V^{\pi^*}(s) = \widehat{V}^{\pi^*}(s).$$

and therefore π^* is optimal for $\widehat{M}$, i.e. $\pi^* \in \widehat{\Pi}^*$. In complete analogy it follows that $\widehat{\Pi}^* \subseteq \Pi^*$ and the equality for the sets of optimal policies follows. □

5.2 Average Reward Measure

All the necessary work has already been done for analyzing CTMDPs by transformation to MDPs. It remains to provide optimality equations for the average reward and algorithms which can be used directly on CTMDPs. Consider a CTMDP (S, Act, e, Q, i, r) with average reward measure and let

$$\overline{r}(s,a) = \sum_{s' \neq s} i(s,a,s') Q(s,a,s') + r(s,a)$$

denote the continuized rate reward. Define the Bellman operators $\mathcal{B}_{\text{av}} \colon \mathbb{R}^S \to \mathbb{R}^S$ and $\mathcal{B}^g_{\text{bias}} \colon \mathbb{R}^S \to \mathbb{R}^S$ (parametrized by $g \in \mathbb{R}^S$) as follows:

$$(\mathcal{B}_{\text{av}} g)(s) := \max_{a \in e(s)} \left\{ \sum_{s' \in S} Q(s,a,s') g(s') \right\}$$

$$(\mathcal{B}^g_{\text{bias}} h)(s) := \max_{a \in e^g(s)} \left\{ \overline{r}(s,a) + \sum_{s' \in S} Q(s,a,s') h(s') \right\} - g(s)$$

$$\text{where } e^g(s) := \left\{ a \in e(s) \mid \sum_{s' \in S} Q(s,a,s') g(s') = 0 \right\}$$

These operators look similar to the Bellman operators (3.10) and (3.11) in the discrete-time case. The difference is that instead of searching for fixed-points we need to search for zeros of $\mathcal{B}_{\text{av}}$ and $\mathcal{B}^g_{\text{bias}}$ (see (4.19)). This gives the first and the second Bellman optimality equations

$$\max_{a \in e(s)} \left\{ \sum_{s' \in S} Q(s,a,s') g(s') \right\} = 0 \tag{5.1}$$

$$\max_{a \in e^g(s)} \left\{ \overline{r}(s,a) + \sum_{s' \in S} Q(s,a,s') h(s') \right\} - g(s) = 0. \tag{5.2}$$

The following existence theorem is the analogue version of Theorem 5.2 for discrete-time MDPs.

Theorem 5.2 (Existence Theorem).

(i) The average optimal value function g^* is a solution to (5.1)*, i.e* $\mathcal{B}_{\text{av}} g^* = 0$*. For $g = g^*$ there exists a solution h to* (5.2)*, i.e.* $\mathcal{B}^{g^*}_{\text{bias}} h = 0$*. If g and h are solutions to* (5.1) *and* (5.2) *then $g = g^*$.*

(ii) There exists an optimal policy π^* and it holds that $g^{\pi^*} = g^*$.

(iii) For any solution h to (5.2) *with $g = g^*$ an optimal policy π^* can be derived from*

$$\pi^*(s) \in \operatorname*{argmax}_{a \in e^{g^*}(s)} \left\{ \overline{r}(s,a) + \sum_{s' \in S} Q(s,a,s') h(s') \right\}.$$

For a direct proof we refer to [17]. We propose here another proof sketch based on uniformization and its value-preserving property.

Proof. Without loss of generality we assume that $E(s,a) \leq 1$ and set the uniformization rate $\mu := 1$ such that the uniformization is value-preserving. The μ-uniformized MDP is given by $\mathcal{M}^\mu = (S, Act, e, P^\mu, R^\mu)$ where

$$P^\mu(s,a,s') = \delta_{s,s'} + Q(s,a,s') \quad \text{and} \quad R^\mu(s,a) = \overline{r}(s,a).$$

If $(g^\mu)^*$ denotes the optimal average reward for $\mathcal{M}^\mu$ then by Theorem 5.1 it holds that $g^* = (g^\mu)^*$. Since finding a fixed point of some operator $\mathcal{T}$ is equivalent to

finding a zero of the operator $\mathcal{B} = \mathcal{T} - id$, where id is the identity operator, part (i) follows. Furthermore, Theorem 3.7 guarantees the existence of an optimal policy for $\mathcal{M}^\mu$ and by Theorem 5.1 also for $\mathcal{M}$ such that parts (ii) and (iii) follow. □

We restate the policy iteration algorithm from [17] since our CTMDP model as introduced in Definition 5.1 allows also impulse rewards.

Theorem 5.3 (Policy Iteration). *Let $\mathcal{M} = (S, Act, e, Q, i, r)$ be a CTMDP and $\overline{r}(s,a)$ the continuized rate reward. For an initial policy $\pi_0 \in \Pi$ define the following iteration scheme:*

1. ***Policy evaluation**: Compute a solution $(g^{\pi_n}, h^{\pi_n}, w)^T$ to*

$$\begin{pmatrix} -Q^{\pi_n} & 0 & 0 \\ I & -Q^{\pi_n} & 0 \\ 0 & I & -Q^{\pi_n} \end{pmatrix} \begin{pmatrix} g^{\pi_n} \\ h^{\pi_n} \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ \overline{r}^{\pi_n} \\ 0 \end{pmatrix}$$

2. ***Policy improvement**: Define for each state s the set of improving actions*

$$B_{n+1}(s) := \left\{ a \in e(s) \mid \begin{array}{r} \sum_{s'} Q(s,a,s') g^{\pi_n}(s') > 0 \ \vee \\ (\sum_{s'} Q(s,a,s') g^{\pi_n}(s') = 0 \\ \Rightarrow \overline{r}(s,a) + \sum_{s'} Q(s,a,s') h^{\pi_n}(s') > g^{\pi_n}(s)) \end{array} \right\}$$

and choose an improving policy π_{n+1} such that

$$\pi_{n+1}(s) \in B_{n+1}(s) \text{ if } B_{n+1}(s) \neq \emptyset \quad \text{or} \quad \pi_{n+1}(s) := \pi_n(s) \text{ if } B_{n+1}(s) = \emptyset.$$

Termination: If $\pi_{n+1} = \pi_n$ then π_n is an optimal policy. Otherwise go to the policy evaluation phase with π_{n+1}.

The values g^{π_n} are non-decreasing and policy iteration terminates in a finite number of iterations with an optimal policy π_n and optimal average reward g^{π_n}.

The policy evaluation phase in this algorithm can be derived from the evaluation phase of the policy iteration algorithm in Theorem 3.8 for the uniformized model. However, the main difference between these algorithms is the policy improvement phase. Here $B_{n+1}(s)$ provides all actions which lead to at least some improvement in the policy π_n whereas in Theorem 3.8 a greedy maximal improving policy is chosen: $G_{n+1}(s)$ respectively $H_{n+1}(s)$. Note that $G_{n+1}(s) \cup H_{n+1}(s) \subseteq B_{n+1}(s)$. Of course, the choice of π_{n+1} in Theorem 5.3 can also be established by the greedy improving policy.

Example 5.1 (Bridge circuit). Consider a brige circuit as outlined in the reliability block diagram in Fig. 5.1.

The system is up, if there is at least one path of working components from s to t and it is down if on every path there is at least one failed component. Each working component $C \in \{L_1, L_2, B, R_1, R_2\}$ can fail after an exponentially distributed time with rate λ_C and there is a single repair unit, which can fix a failed component C after an exponentially distributed time with rate μ_C.

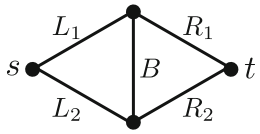

Fig. 5.1. The reliability block diagram of the bridge circuit system. An edge represents a component, which can be working or failed.

We assume that the components L_1 and L_2 (respectively R_1 and R_2) are identical and the parameter values for all components are

$$\begin{array}{lll} \lambda_{L_i} = 1.0 & \lambda_B = 0.1 & \lambda_{R_i} = 2.0 \\ \mu_{L_i} = 10.0 & \mu_B = 100.0 & \mu_{R_i} = 10.0. \end{array}$$

The action model allows the repair unit to be assigned to a failed component or to decide not to repair. We further assume that repair is preemptive, i.e. if during repair of a failed component another component fails, then the repair unit can decide again which component to repair. Note that due to the memoryless property of the exponential repair distribution, the remaining repair time in order to complete the repair does not depend on the elapsed time for repair. We want to find optimal repair policies, in order to pursue the following two goals:

(G1): maximize the MTTF (mean time to failure)
(G2): maximize the availability (i.e. the fraction of uptime in the total time).

Figure 5.2 shows an excerpt of the state space (with 32 states), which we apply to both goals (G1) and (G2). Note that for both measures (MTTF and availability) we define the reward structure which consists only of the rate reward r, which is 1 on up states and 0 on down states. The difference between both goals affects the state space as follows: For (G1) the 16 down states are absorbing (for every policy), while for (G2) a repair of failed components is also allowed in down system states.

We optimize (G1) by transforming the CTMDP by embedding into a discrete-time SSP (S, P, R) (cf. Definition 3.4 and Fig. 4.7) and hereby aggregate all absorbing down states to the goal state for the SSP. By embedding transformation, the reward $R(s, a)$ is the expected sojourn time in state s under action a in the CTMDP model, i.e. for all $a \in e(s)$

$$R(s,a) = \begin{cases} \frac{1}{E(s,a)}, & \text{for } s \neq \mathit{goal} \\ 0, & \text{for } s = \mathit{goal} \end{cases},$$

where $E(s, a)$ is the exit rate. Table 5.1 shows the resulting optimal policy and its corresponding maximal MTTF value function.

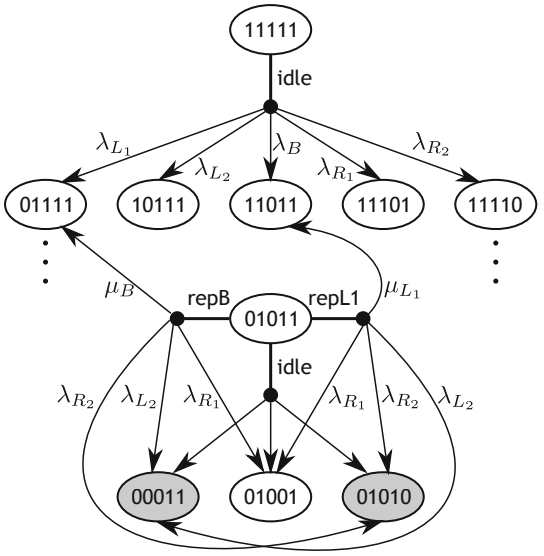

Fig. 5.2. State space of the bridge circuit CTMDP model. State encoding 01011 represents (from left to right) that L_1 has failed, L_2 is working, B has failed, R_1 is working and R_2 is working. From every state where at least some component has failed, there are repair actions and from each state there is also the idle action indicating that the repair unit can choose not to repair. Shadowed states represent down system states.

Table 5.1. Optimal policy and maximal MTTF value function of the bridge ciruit system (Example 5.1)

11111	11110	11101	11011	11010	11001	10111	10110	10101
idle	repR2	repR1	repB	repB	repB	repL2	repR2	repR1
1.449	1.262	1.262	1.448	1.234	1.234	1.324	1.095	1.087

10011	10010	01111	01110	01101	01011	01001	goal
repB	repB	repL1	repR2	repR1	repB	repB	idle
1.291	1.073	1.324	1.087	1.095	1.291	1.073	0.000

Problem (G2) is optimized by applying the CTMDP policy iteration algorithm for the average reward as outlined in Theorem 5.3. Beginning with the initial policy constantly *idle*, policy iteration converged in 6 iteration steps to the optimal policy given in Table 5.2.

Note that for (G2) the model can be shown to be unichain. Thus, the average reward g^π is constant for all policies π such that $Q^\pi g^\pi = 0$. For this reason, the policy evaluation and improvement phases in the policy iteration algorithm can be simplified. □

Table 5.2. Optimal policy for the availability of the bridge ciruit system (Example 5.1). The maximal availability is 0.917757 independent of the initial state.

11111	11110	11101	11100	11011	11010	11001	11000	10111	10110	10101
idle	repR2	repR1	repR1	repB	repR2	repR1	repR1	repL2	repR2	repR1
10100	10011	10010	10001	10000	01111	01110	01101	01100	01011	01010
repR1	repB	repR2	repB	repR1	repL1	repR2	repR1	repR2	repB	repB
01001	01000	00111	00110	00101	00100	00011	00010	00001	00000	
repR1	repR2	repL1	repL1	repL2	repL1	repL1	repL1	repL2	repL2	

6 Conclusion and Outlook

In this tutorial, we presented an integrated picture of MRMs and MDPs over finite state spaces for both discrete and continuous time. The theory and application area for this kind of models is very popular and broad. For this reason, we just focussed on the fundamentals of the theory. We reviewed the most important basic facts from literature, which are inevitable for a deeper understanding of Markovian models. Furthermore, we set up the theory step by step from discrete-time MRMs up to continuous-time MDPs and pointed out important links between these theories. We also connected these models by a number of model transformations and highlighted their properties. In order to show the applicability of these models, we introduced small prototypical examples, coming from the domain of performance and dependability evaluation and optimization. Of course, many current applications in the optimization of dependable systems suffer from the curse of dimensionality. However, there are established techniques which can be used in order to overcome this curse and make evaluation and optimization for large practical models accessible, e.g. approximate solutions (e.g. approximate dynamic programming), simulative approaches (reinforcement learning) or the use of structured models. There are also several important extensions to the Markov model types we could not address in this introductory tutorial, such as partially observable MDPs (used especially in the area of AI), denumerable state and action spaces (e.g. for queueing systems) or even continuous state and action spaces (leading directly to control theory).

A Appendix

A.1 Lemmata and Remarks

Lemma A.1.

(i) Let N be a non-negative discrete random variable with expected value and $(a_n)_{n\in\mathbb{N}}$ a bounded sequence. Then

$$\mathbb{E}\left[\sum_{n=0}^{N} a_n\right] = \sum_{n=0}^{\infty} a_n P(N \geq n).$$

(ii) Let T be a non-negative continuous random variable with expected value and $a : [0, \infty) \to \mathbb{R}$ an integrable and bounded function. Then

$$\mathbb{E}\left[\int_0^T a(t)\,dt\right] = \int_0^\infty a(t) P(T \geq t)\,dt.$$

Proof. We show only (i) – the proof for (ii) is analogous when summation is replaced by integration. Since a_n is bounded and $\mathbb{E}[N] = \sum_{n=0}^\infty P(N \geq n) < \infty$ it follows that $\sum_{n=0}^\infty a_n P(N \geq n)$ converges absolutely. From

$$\sum_{n=0}^\infty a_n P(N \geq n) = \sum_{n=0}^\infty \sum_{k=n}^\infty a_n P(N = k) = \sum_{n=0}^\infty \sum_{k=0}^\infty a_n P(N = k) \mathbb{1}_{\{k \geq n\}}(n,k)$$

we can interchange both infinite summations by the Fubini theorem. It follows

$$\begin{aligned}\sum_{n=0}^\infty a_n P(N \geq n) &= \sum_{k=0}^\infty \sum_{n=0}^\infty a_n P(N = k) \mathbb{1}_{\{n \leq k\}}(n,k) \\ &= \sum_{k=0}^\infty \sum_{n=0}^k a_n P(N = k) = \mathbb{E}\left[\sum_{n=0}^N a_n\right]. \qquad \square\end{aligned}$$

Remark A.1. As presented in Sect. 4.2.2 a CTMC $\mathcal{M} = (S, Q)$ induces the stochastic processes X_n, τ_n, T_n, N_t and Z_t. If we fix a uniformization rate $\mu > 0$ then $\mathcal{M}$ also induces the ***uniformized stochastic processes*** $\widetilde{X}_n$, $\widetilde{\tau}_n$, $\widetilde{T}_n$, $\widetilde{N}_t$ and $\widetilde{Z}_t$ over $\Omega = (S \times (0, \infty])^{\mathbb{N}}$. Here $\widetilde{X}_n$ is the n-th visited state in the uniformized DTMC and $\widetilde{\tau}_n$ is the time up to transition in state $\widetilde{X}_n$, i.e. all $\widetilde{\tau}_n$ are independent and exponentially distributed with rate μ. Moreover, the total elapsed time $\widetilde{T}_n := \sum_{k=0}^{n-1} \widetilde{\tau}_k$ for the first n transitions is Erlang distributed with n phases and rate μ and the number $\widetilde{N}_t := \max\{n \mid \widetilde{T}_n(\omega) \leq t\}$ of (uniformized) transitions up to time $t \geq 0$ is Poisson distributed with parameter μt. Note also that $\widetilde{Z}_t := \widetilde{X}_{\widetilde{N}_t} = X_{N_t} = Z_t$ for all $t \geq 0$. Thus, when uniformization is considered as adding exponentially distributed self-loop transitions to states of the CTMC $\mathcal{M}$, then the continuous-time state process Z_t is not modified at all. Therefore, the probability measures P_s of the CTMC for all $s \in S$ are left invariant under uniformization and especially the transient probability matrix $P(t)$ as defined in (4.1). $\square$

Lemma A.2. *Consider a CTMC $\mathcal{M} = (S, Q)$ with discrete-time and continuous-time state processes X_n and Z_t. Let*

$$n_T(s, s') := \mathbb{E}_{s_0}\left[\sum_{k=1}^{N_T} \mathbb{1}_{\{X_{k-1}=s, X_k=s'\}}\right]$$

be the expected number of transitions from state s to state $s' \neq s$ within the time interval $[0, T]$ from a fixed initial state $X_0 = s_0$. Then

$$n_T(s, s') = Q(s, s') \int_0^T P_{s_0}(Z_t = s)\,dt. \tag{A.1}$$

Proof. If s is absorbing then clearly (A.1) holds and we can assume in the following that $E(s) > 0$. We abbreviate in the following the notation by $\mathbb{E} := \mathbb{E}_{s_0}$ and $P := P_{s_0}$. Define

$$n_T(s) := \sum_{s' \neq s} n_T(s,s') = \mathbb{E}\left[\sum_{k=1}^{N_T} \mathbb{1}_{\{X_{k-1}=s\}}\right] = \sum_{k=1}^{\infty} P(X_{k-1}=s) P(N_T \geq k)$$

as the number of complete visits to state s, such that s has been left before time T. From $P(X_{k-1}=s, X_k = s') = P(X_k = s' | X_{k-1}=s) P(X_{k-1}=s)$ it follows that

$$n_T(s,s') = P(s,s') \sum_{k=1}^{\infty} P(X_{k-1}=s) P(N_T \geq k) = P(s,s') n_T(s),$$

where $P(s,s') = \delta_{s,s'} + \frac{Q(s,s')}{E(s)}$ is the embedded transition probability. We use uniformization as means for proof with a uniformization rate $\mu \geq \max_{s \in S} E(s)$. Let $\widetilde{X}_k$, $\widetilde{\tau}_k$, $\widetilde{T}_k$ and $\widetilde{N}_t$ be the uniformized stochastic processes as defined in Remark A.1. Then $\widetilde{T}_k$ is Erlang distributed with density $f_{\widetilde{T}_k}(t) = e^{-\mu t} \frac{\mu^k t^{k-1}}{(k-1)!}$ for $t \geq 0$ and $\widetilde{N}_t$ has the Poisson probabilities $P(\widetilde{N}_t = k) = e^{-\mu t} \frac{(\mu t)^k}{k!}$. The total accumulated time for complete visits in s up to time T fulfills

$$\frac{1}{E(s)} \mathbb{E}\left[\sum_{k=1}^{N_T} \mathbb{1}_{\{X_{k-1}=s\}}\right] = \frac{1}{\mu} \mathbb{E}\left[\sum_{k=1}^{\widetilde{N}_T} \mathbb{1}_{\{\widetilde{X}_{k-1}=s\}}\right]$$

and therefore

$$n_T(s) = \frac{E(s)}{\mu} \mathbb{E}\left[\sum_{k=1}^{\widetilde{N}_T} \mathbb{1}_{\{\widetilde{X}_{k-1}=s\}}\right]$$

is a fraction of the number of uniformized transitions up to time T from state s to some arbitrary state s'. It follows that

$$\begin{aligned}
n_T(s) &= \frac{E(s)}{\mu} \mathbb{E}\left[\sum_{k=1}^{\widetilde{N}_T} \mathbb{1}_{\{\widetilde{X}_{k-1}=s\}}\right] = \frac{E(s)}{\mu} \sum_{k=1}^{\infty} P(\widetilde{X}_{k-1}=s) P(\widetilde{N}_T \geq k) \\
&= \frac{E(s)}{\mu} \sum_{k=0}^{\infty} P(\widetilde{X}_k = s) P(\widetilde{T}_{k+1} \leq T) \\
&= \frac{E(s)}{\mu} \sum_{k=0}^{\infty} P(\widetilde{X}_k = s) \int_0^T e^{-\mu t} \frac{\mu^{k+1} t^k}{k!}\, dt \\
&= E(s) \int_0^T e^{-\mu t} \sum_{k=0}^{\infty} P(\widetilde{X}_k = s) \frac{(\mu t)^k}{k!}\, dt = E(s) \int_0^T P(Z_t = s)\, dt
\end{aligned}$$

since

$$P(Z_t = s) = \sum_{k=0}^{\infty} P(Z_t = s \mid \widetilde{N}_t = k) P(\widetilde{N}_t = k) = \sum_{k=0}^{\infty} P(\widetilde{X}_k = s) e^{-\mu t} \frac{(\mu t)^k}{k!}.$$

Thus, (A.1) follows from $P(s,s') = \frac{Q(s,s')}{E(s)}$ for $s' \neq s$ and $n_T(s,s') = P(s,s')n_T(s)$. □

Remark A.2. In the proof of Lemma A.2, we have applied uniformization as a detour in order to show that

$$n_T(s,s') = Q(s,s') \int_0^T P_{s_0}(Z_t = s)\, dt.$$

There is also a more direct way to show this equation by an argument that is used in the proof of the PASTA property ("Poisson Arrivals See Time Averages") [41]. The PASTA property is a tool that is frequently used in the theory of queueing systems. Consider a system that is represented by the Markov chain $\mathcal{M} = (S, Q)$ with state process Z_t for $t \geq 0$ and fix two states s and s' with $Q(s,s') > 0$. Let $\tau_{s,s'}$ be the exponentially distributed time with rate $Q(s,s')$ that governs the transition from s to s' as shown in Sect. 4.1. Further define an independent sequence of such random variables $\tau_{s,s'}^{(n)}$, $n \in \mathbb{N}$ with same distribution as $\tau_{s,s'}$. Then the process $A_t := \max\left\{n \mid \sum_{k=0}^{n-1} \tau^{(k)} \leq t\right\}$ is a Poisson process with rate $Q(s,s')$ and is regarded as a stream of arriving jobs to the system. Since $\tau_{s,s'}$ is memoryless it holds that when the system is in state s and an arrival occurs then the system performs a transition to s'. Therefore, the counting process $Y_t := \sum_{k=1}^{N_t} \mathbb{1}_{\{X_{k-1}=s, X_k=s'\}}$ is precisely the number of arrivals of A_t to the system (up to time t) that find the system in state s. Further let $U_t := \mathbb{1}_{\{Z_t=s\}}$ be the process that indicates whether the system is in state s at time t. It holds that Y_t can be represented as a stochastic Riemann-Stiltjes integral of the process U_t with respect to the arrival process A_t, i.e. for all $T \geq 0$ it holds that $Y_T = \int_0^T U_t\, dA_t$ with probability 1. Note that for each $t \geq 0$ the set of future increments $\{A_{t+s} - A_s \mid s \geq 0\}$ and the history of the indicator process $\{U_s \mid 0 \leq s \leq t\}$ are independent. Thus the "lack of anticipation assumption" as needed for [41] is satisfied and it follows that

$$n_T(s,s') = \mathbb{E}_{s_0}[Y_T] = Q(s,s') \cdot \mathbb{E}_{s_0}\left[\int_0^T U_t\, dt\right] = Q(s,s') \int_0^T P_{s_0}(Z_t = s)\, dt.$$

□

Lemma A.3. *Let T be a non-negative continuous random horizon length for a CTMRM $\mathcal{M} = (S, Q, i, r)$ and independent of the state process Z_t of $\mathcal{M}$. Further let $N_T = \max\{n \mid T_n \leq T\}$ be the random number of transitions up to time T. If the k-th moment of T exists, then it also exists for N_T.*

In order to prove this theorem we need the following definition.

Definition A.1. *For two random variables X and Y with distributions F_X and F_Y we say that X is **stochastically smaller** then Y (denoted by $X \preceq Y$) if $F_X(x) \geq F_Y(x)$ for all $x \in \mathbb{R}$.*

It follows that if $X \preceq Y$ then $\mathbb{E}[g(X)] \leq \mathbb{E}[g(Y)]$ for a monotonically increasing function g.

Proof. Let X_n, τ_n, T_n and N_t be the stochastic processes as defined in Sect. 4.2.2. Further choose $\mu := \max\{E(s) \mid s \in S\}$ as a uniformization rate and $\widetilde{X}_n, \widetilde{\tau}_n, \widetilde{T}_n$ and $\widetilde{N}_t$ the uniformized processes as in Remark A.1. First we show that $N_T \preceq \widetilde{N}_T$: From $\mu \geq E(s)$ for all s it follows that $\widetilde{\tau}_n \preceq \tau_n$ for all n and thus $\widetilde{T}_n \preceq T_n$. Therefore

$$N_T = \max\{n \mid T_n \leq T\} \preceq \max\{n \mid \widetilde{T}_n \leq T\} = \widetilde{N}_T$$

and thus $\mathbb{E}[N_T^k] \leq \mathbb{E}[\widetilde{N}_T^k]$. In order to show that $\mathbb{E}[N_T^k]$ is finite we show $\mathbb{E}[\widetilde{N}_T^k] < \infty$. It holds that $P(\widetilde{N}_T = n) = P(\widetilde{T}_n \leq T < \widetilde{T}_{n+1})$ and therefore

$$\mathbb{E}[\widetilde{N}_T^k] = \sum_{n=0}^{\infty} n^k P(\widetilde{T}_n \leq T < \widetilde{T}_{n+1}).$$

We show that the sequence $P(\widetilde{T}_n \leq T < \widetilde{T}_{n+1})$ is decreasing fast enough.

$$P(\widetilde{T}_n \leq T < \widetilde{T}_{n+1}) = \int_{z=0}^{\infty} f_T(z) \int_{v=0}^{z} f_{\widetilde{T}_n}(v) \int_{u=z-v}^{\infty} f_{\widetilde{\tau}_n}(u)\, du\, dv\, dz =$$
$$\int_{z=0}^{\infty} f_T(z) \int_{v=0}^{z} f_{\widetilde{T}_n}(v) e^{-\mu(z-v)}\, dv\, dz = \int_{z=0}^{\infty} f_T(z) e^{-\mu z} \int_{v=0}^{z} \frac{\mu^n v^{n-1}}{(n-1)!}\, dv\, dz =$$
$$\int_{z=0}^{\infty} f_T(z) \frac{e^{-\mu z}(\mu z)^n}{n!}\, dz.$$

Therefore

$$\mathbb{E}[\widetilde{N}_T^k] = \sum_{n=0}^{\infty} n^k \int_{z=0}^{\infty} f_T(z) \frac{e^{-\mu z}(\mu z)^n}{n!}\, dz = \int_{z=0}^{\infty} f_T(z) \mathbb{E}[\widetilde{N}_z^k]\, dz,$$

where $\widetilde{N}_z$ is the number of uniformized transitions up to time z which is Poisson distributed with parameter μz. Now the k-th moment $g(z) := \mathbb{E}[\widetilde{N}_z^k]$ is a polynomial in z of degree k and therefore

$$\mathbb{E}[\widetilde{N}_T^k] = \int_{z=0}^{\infty} g(z) f_T(z)\, dz < \infty,$$

as a polynomial of degree k in the moments of T. □

Remark A.3. In the case $k = 1$ it holds that $\mathbb{E}[\widetilde{N}_T] = \mu \mathbb{E}[T]$ represents exactly Little's law: If jobs enter a queue at rate μ and if their mean residence time in the queue is $\mathbb{E}[T]$, then there are on average $\mu \mathbb{E}[T]$ jobs in the queue. For $k \geq 2$ the theorem generalizes Little's law and allows to compute $\mathbb{E}[\widetilde{N}_T]$ analytically since the coefficients of g can be computed analytically.

A.2 Laurent Series Expansion for Continuous Time Models

Proposition A.1. *Let $Q \in \mathbb{R}^{n \times n}$ be the generator matrix of a CTMC over a finite state space and $P(t) = e^{Qt}$ the transient probability matrix. Then the*

CTMC is exponentially ergodic, i.e. there exists $\delta > 0$ and $L > 0$, such that

$$||P(t) - P^*|| \leq Le^{-\delta t}$$

for all $t \geq 0$, where $||A|| := \max_i \sum_j |A_{i,j}|$ is the matrix maximum norm.

In [18] an equivalent statement is described for finite state DTMCs. We transfer and modify the proof to the continuous-time case.

Proof. Since $P_{i,j}(t) \to P^*_{i,j}$ for all i, j, it follows that for an arbitrary fixed $\varepsilon > 0$ there exists $T > 0$, such that for all i

$$\sum_j |P_{i,j}(T) - P^*_{i,j}| \leq e^{-\varepsilon} < 1$$

and therefore $||P(T) - P^*|| \leq e^{-\varepsilon}$. Now split $t = n_t T + s_t$ with $n_t \in \mathbb{N}$ and $s_t \in [0, T)$.

$$\begin{aligned}
||P(t) - P^*|| &= ||P(T)^{n_t} P(s_t) - P^*|| && P(s+t) = P(s)P(t) \\
&= ||\left(P(T)^{n_t} - P^*\right)\left(P(s_t) - P^*\right)|| && P(t)P^* = P^*,\ P^*P^* = P^* \\
&\leq ||P(T)^{n_t} - P^*|| \cdot ||P(s_t) - P^*|| && \text{subadditivity of norm} \\
&= ||\left(P(T) - P^*\right)^{n_t}|| \cdot ||P(s_t) - P^*|| \\
&\leq ||P(T) - P^*||^{n_t} \cdot ||P(s_t) - P^*|| \\
&\leq e^{-\varepsilon n_t} ||P(s_t) - P^*|| \\
&= e^{-\varepsilon t/T} e^{\varepsilon s_t/T} ||P(s_t) - P^*||.
\end{aligned}$$

Defining

$$\delta := \frac{\varepsilon}{T} \quad \text{and} \quad L := \sup_{s \in [0,T)} \left(e^{\varepsilon s/T} ||P(s) - P^*|| \right) < \infty$$

gives $||P(t) - P^*|| \leq Le^{-\delta t}$. □

Define the transient deviation matrix $\Delta(t) := P(t) - P^*$ and the total deviation matrix $H := \int_0^\infty \Delta(t)\, dt$ (componentwise integration). From Proposition A.1 it follows that the integral defining H converges since $||\Delta(t)|| \leq Le^{-\delta t}$.

Theorem A.1. *For $\alpha > 0$ let $W(\alpha) := \int_0^\infty e^{-\alpha t} P(t)\, dt$ be the Laplace transform of $P(t)$. Then there exists $\delta > 0$, such that for all $0 < \alpha < \delta$ the Laurent series of $W(\alpha)$ is given by*

$$W(\alpha) = \alpha^{-1} P^* + \sum_{n=0}^{\infty} (-\alpha)^n H^{n+1}.$$

Proof. Since $P(t)P^* = P^*P(t) = P^*P^* = P^*$ it follows that $\Delta(t+s) = \Delta(t)\Delta(s)$ for all $s, t \geq 0$. Now

$$\begin{aligned} W(\alpha) &= \int_0^\infty e^{-\alpha t}(P(t) - P^* + P^*)\,dt \\ &= \alpha^{-1}P^* + \int_0^\infty e^{-\alpha t}\Delta(t)\,dt \\ &= \alpha^{-1}P^* + \int_0^\infty \left(\sum_{n=0}^\infty \frac{(-\alpha t)^n}{n!}\right)\Delta(t)\,dt \\ &= \alpha^{-1}P^* + \sum_{n=0}^\infty (-\alpha)^n \int_0^\infty \frac{t^n}{n!}\Delta(t)\,dt, \end{aligned}$$

where the last equality follows from Lebesgue's dominated convergence theorem, since for all i, j the sequence $\sum_{n=0}^N \frac{(-\alpha t)^n}{n!}\Delta_{i,j}(t)$ can be dominated by the integrable function $Ce^{(\alpha-\delta)t}$ (for $\delta > 0$ from Proposition A.1 and some $C > 0$) for all $0 < \alpha < \delta$. We show by induction that

$$\int_0^\infty \frac{t^n}{n!}\Delta(t)\,dt = H^{n+1}. \tag{A.2}$$

For $n = 0$ this is true by definition of H. Let (A.2) be true for an arbitrary $n \in \mathbb{N}$. Then

$$\begin{aligned} \int_0^\infty \frac{t^{n+1}}{(n+1)!}\Delta(t)\,dt &= \int_0^\infty \left(\int_0^t \frac{s^n}{n!}\,ds\right)\Delta(t)\,dt = \int_0^\infty \frac{s^n}{n!}\int_s^\infty \Delta(t)\,dt\,ds \\ &= \int_0^\infty \frac{s^n}{n!}\int_0^\infty \Delta(s+t)\,dt\,ds = \int_0^\infty \frac{s^n}{n!}\Delta(s)\int_0^\infty \Delta(t)\,dt\,ds \\ &= \left(\int_0^\infty \frac{s^n}{n!}\Delta(s)\,ds\right)H = H^{n+1} \end{aligned}$$

and the Laurent series follows. □

A.3 Collection of Proofs

Proof (of Theorem 2.1). (i) For an arbitary $s_0 \in S$ it holds

$$\begin{aligned} V_N(s_0) = \mathbb{E}_{s_0}\left[\sum_{i=1}^N R(X_{i-1}, X_i)\right] &= \sum_{s_1,\ldots,s_N}\left(\left(\sum_{i=1}^N R(s_{i-1}, s_i)\right)\prod_{i=1}^N P(s_{i-1}, s_i)\right) \\ = \sum_{s_1} P(s_0, s_1)\Bigg(R(s_0, s_1)\sum_{s_2,\ldots,s_N}\prod_{i=2}^N P(s_{i-1}, s_i) + \\ \sum_{s_2,\ldots,s_N}\sum_{i=2}^N R(s_{i-1}, s_i)\prod_{i=2}^N P(s_{i-1}, s_i)\Bigg). \end{aligned}$$

Now since $\sum_{s_1} R(s_0, s_1)P(s_0, s_1) = R(s_0)$ and for each $s_1 \in S$ it holds that

$$\sum_{s_2,\ldots,s_N} \prod_{i=2}^{N} P(s_{i-1}, s_i) = 1 \quad \text{and}$$

$$\sum_{s_2,\ldots,s_N} \left(\sum_{i=2}^{N} R(s_{i-1}, s_i) \right) \prod_{i=2}^{N} P(s_{i-1}, s_i) = \mathbb{E}_{s_1} \left[\sum_{i=2}^{N} R(X_{i-1}, X_i) \right] = V_{N-1}(s_1)$$

it follows that

$$V_N(s_0) = R(s_0) + \sum_{s_1} P(s_0, s_1) V_{N-1}(s_1).$$

In case V_∞ exists then $V_\infty(s) = \lim_{N\to\infty} V_N(s)$ for all $s \in S$ and statement (ii) follows from (i) by taking the limit on both sides. □

Proof (of Proposition 2.1). We are going to sketch a proof for this fact in case the Markov chain is aperiodic. Since V_∞ exists for the model (S, P, R) if and only if it exists for $(S, P, |R|)$ we can assume without loss of generality that $R(s, s') \geq 0$ for all $s, s' \in S$. Note that here $|R|$ has to be interpreted as the transition-based reward function with $|R|(s, s') := |R(s, s')|$. The reason is that the state-based view on the absolute reward values $\sum_{s'\in S} P(s, s')|R(s, s')|$ in general differs from $|\sum_{s'\in S} P(s, s')R(s, s')|$ which is the absolute value of the state-based view on the reward values!

"⇒": Assume that V_∞ exists and $R(\widetilde{s}, \widetilde{s}') > 0$ for some states $\widetilde{s}, \widetilde{s}' \in S_i^r$ and thus $R(\widetilde{s}) = \sum_{s'\in S} P(\widetilde{s}, s')R(\widetilde{s}, s') > 0$. For all $k \in \mathbb{N}$ it holds that $\mathbb{E}_s [R(X_{k-1}, X_k)]$ is the reward gained for the k-th transition when starting in s. Therefore

$$\begin{aligned} \mathbb{E}_{\widetilde{s}} [R(X_{k-1}, X_k)] &= \sum_{s'\in S} P^{k-1}(\widetilde{s}, s') \sum_{s''\in S} P(s', s'')R(s', s'') \\ &= \sum_{s'\in S} P^{k-1}(\widetilde{s}, s')R(s') \geq P^{k-1}(\widetilde{s}, \widetilde{s})R(\widetilde{s}). \end{aligned}$$

Since P is aperiodic and $\widetilde{s}$ is recurrent it follows that $P^{k-1}(\widetilde{s}, \widetilde{s})$ converges to $\rho_{\widetilde{s}}(\widetilde{s}) > 0$, where $\rho_{\widetilde{s}}$ is the limiting distribution from $\widetilde{s}$ (see Sect. 2.1.2). Therefore the sequence $\mathbb{E}_{\widetilde{s}} \left[\sum_{k=1}^{N} |R(X_{k-1}, X_k)| \right] \geq \sum_{k=1}^{N} P^{k-1}(\widetilde{s}, \widetilde{s})R(\widetilde{s})$ is unbounded, which is a contradiction to the existence of V_∞.

"⇐": Assume that $R(s, s') = 0$ for all $s, s' \in S_i^r$ and all $i = 1, \ldots, k$. We anticipate a result from Proposition 2.2 in Sect. 2.4, which states that the limiting matrix $P^\infty := \lim_{n\to\infty} P^n$ exists since P is aperiodic. In [18] it is shown that P is geometric ergodic, i.e. there exists $n_0 \in \mathbb{N}$, $c > 0$ and $\beta < 1$ such that

$$||P^n - P^\infty|| \leq c\beta^n$$

for all $n \geq n_0$, where $||.||$ is the maximum norm. (This result as stated holds for unichain models, but it can also be directly extended to the multichain case). First of all, we want to show that $P^\infty R = 0$, i.e. for all $s \in S$ it holds that

$$(P^\infty R)(s) = \sum_{s'\in S} P^\infty(s, s') \sum_{s''\in S} P(s', s'')R(s', s'') = 0.$$

If $s \in S_i^r$ is recurrent then we only have to consider those terms in the summation for which s' and s'' are in the same closed recurrent class S_i^r. But since both $s', s'' \in S_i^r$ it follows that $R(s', s'') = 0$ and thus $(P^\infty R)(s) = 0$. On the other hand if $s \in S^t$ is transient then $P^\infty(s, s') = 0$ for all $s' \in S^t$ and otherwise if s' is recurrent then again $P(s', s'') = 0$ or $R(s', s'') = 0$ dependent on whether s' and s'' are in the same closed recurrent class. (Compare this also to the representation of $P^\infty = P^*$ in (2.14).) Combining together it follows for all $s \in S$ and $k \geq n_0+1$ that

$$\mathbb{E}_s\left[R(X_{k-1}, X_k)\right] = \sum_{s' \in S} P^{k-1}(s, s') \sum_{s'' \in S} P(s', s'')R(s', s'') = \sum_{s' \in S} P^{k-1}(s, s')R(s')$$

$$\leq \max_{s \in S} \left\{ \sum_{s' \in S} P^{k-1}(s, s')R(s') \right\} = ||P^{k-1}R|| = ||P^{k-1}R - P^\infty R|| \leq c\beta^{k-1}||R||.$$

Therefore

$$\mathbb{E}_s\left[\sum_{k=1}^{N} R(X_{k-1}, X_k)\right] \leq \sum_{k=1}^{N} c\beta^{k-1}||R||$$

converges as $N \to \infty$ since $\beta < 1$. □

Proof (of Theorem 3.2). The convergence of V_n to $(V^\gamma)^*$ has been already remarked in Remark 3.2. It further holds

$$||V^{\pi_\varepsilon} - (V^\gamma)^*|| \leq ||V^{\pi_\varepsilon} - V_{n+1}|| + ||V_{n+1} - (V^\gamma)^*||.$$

From (2.11) it follows that for every policy π the linear operator T^π defined by $T^\pi V := R^\pi + \gamma P^\pi V$ is also a contraction with the same Lipschitz constant $q := \gamma < 1$ as for $\mathcal{T}$. Let $V^{\pi_\varepsilon} = T^{\pi_\varepsilon} V^{\pi_\varepsilon}$ be the fixed point of T^{π_ε}. By definition of π_ε in (3.7) (i.e. $\pi_\varepsilon(s)$ is a maximizing action) it follows that $T^{\pi_\varepsilon} V_{n+1} = \mathcal{T} V_{n+1}$. Thus, for the first term it holds

$$\begin{aligned} ||V^{\pi_\varepsilon} - V_{n+1}|| &\leq ||V^{\pi_\varepsilon} - \mathcal{T}V_{n+1}|| + ||\mathcal{T}V_{n+1} - V_{n+1}|| \\ &= ||T^{\pi_\varepsilon} V^{\pi_\varepsilon} - T^{\pi_\varepsilon} V_{n+1}|| + ||\mathcal{T}V_{n+1} - \mathcal{T}V_n|| \\ &\leq q||V^{\pi_\varepsilon} - V_{n+1}|| + q||V_{n+1} - V_n||. \end{aligned}$$

Therefore

$$||V^{\pi_\varepsilon} - V_{n+1}|| \leq \frac{q}{1-q}||V_{n+1} - V_n||.$$

In analogy it follows for the second term

$$||V_{n+1} - (V^\gamma)^*|| \leq q||V_n - (V^\gamma)^*|| \leq q\left(||V_n - V_{n+1}|| + ||V_{n+1} - (V^\gamma)^*||\right)$$

and thus

$$||V_{n+1} - (V^\gamma)^*|| \leq \frac{q}{1-q}||V_{n+1} - V_n||.$$

By combining the inequalities together it follows that

$$||V^{\pi_\varepsilon} - (V^\gamma)^*|| \leq \frac{2q}{1-q}||V_{n+1} - V_n||.$$

Hence the conclusion follows from $||V_{n+1} - V_n|| < \frac{1-\gamma}{2\gamma}\varepsilon$ for $q = \gamma$. □

Proof (of Proposition 4.1). The proof is analogous to the proof of Proposition 2.1 in the discrete-time setting. By Definition 4.2 the value function V_∞ is defined if and only if it holds for all $s \in S$ that both terms $\mathbb{E}_s\left[\sum_{k=1}^{N_T} |i(X_{k-1}, X_k)|\right]$ and $\mathbb{E}_s\left[\int_0^T |r(Z_t)|\, dt\right]$ converge as $T \to \infty$. For simplicity, we only sketch the proof for the rate reward. Without loss of generality we assume that $r(s) \geq 0$ for all $s \in S$.

"$\Rightarrow$": V_∞ is defined if and only if $\int_0^T \mathbb{E}_s[r(Z_t)]\, dt$ converges with $T \to \infty$ and thus $\mathbb{E}_s[r(Z_t)] \to 0$ as $t \to \infty$. But if s is recurrent then $\lim_{t\to\infty} P(t)(s,s) = P^*(s,s) > 0$ and from $\mathbb{E}_s[r(Z_t)] = \sum_{s'\in S} P(t)(s,s')r(s') \geq P(t)(s,s)r(s)$ it follows that $r(s) = 0$.

"$\Leftarrow$": Let $r(s) = 0$ for all recurrent states s. As in the discrete-time case, one can show that the transient probability matrix $P(t)$ of the finite-state CTMC (S, Q) is exponentially ergodic, i.e. there exists $L > 0$ and $\delta > 0$ such that $||P(t) - P^*|| \leq Le^{-\delta t}$ for all $t \geq 0$ where $||.||$ is the maximum norm (see Proposition A.1). We first show that

$$(P^* r)(s) = \sum_{s'\in S} P^*(s,s')r(s') = 0$$

for all $s \in S$ (see also the representation of P^* in (4.3)). If $s \in S_i^r$ is recurrent then $P^*(s,s') = 0$ if $s' \in S \setminus S_i^r$ and $r(s') = 0$ if $s' \in S_i^r$. Otherwise, if $s \in S^t$ is transient then $P^*(s,s') = 0$ for all transient states $s' \in S^t$ and $r(s') = 0$ for all recurrent states $s' \in S \setminus S^t$. It follows for all $s \in S$ that

$$\begin{aligned}\mathbb{E}_s\left[\int_0^T r(Z_t)\, dt\right] &= \int_0^T \sum_{s'\in S} P(t)(s,s')r(s')\, dt \leq \int_0^T ||P(t)r||\, dt \\ &= \int_0^T ||P(t)r - P^* r||\, dt \leq \int_0^T Le^{-\delta t}||r||\, dt\end{aligned}$$

converges as $T \to \infty$. □

Proof (of Lemma 4.1). We show the first equality by regarding the representation of $V(s)$ in (4.13) as a total expectation. We can interchange both expectations in the middle term by the Fubini theorem (or law of total expectation), i.e.

$$V(s) = \mathbb{E}\left[\mathbb{E}_s\left[\int_0^T \overline{r}(Z_t)\, dt \mid T\right]\right] = \mathbb{E}_s\left[\mathbb{E}\left[\int_0^T \overline{r}(Z_t)\, dt \mid Z_t\right]\right].$$

Here $\mathbb{E}\left[\int_0^T \overline{r}(Z_t)\, dt \mid Z_t\right]$ is a conditional expectation given knowledge of all the Z_t for $t \geq 0$, i.e. it is a random variable over Ω that takes the values $\mathbb{E}\left[\int_0^T \overline{r}(Z_t(\omega))\, dt\right]$ for $\omega \in \Omega$. Since the state space is finite it holds that the map $t \mapsto \overline{r}(Z_t(\omega))$ is bounded for all $\omega \in \Omega$ and it follows by Lemma A.1 that

$$V(s) = \mathbb{E}_s\left[\int_0^\infty \overline{r}(Z_t) P_T(T \geq t)\, dt\right].$$

For the second equality of $V(s)$ in Lemma 4.1 note that $Z_t(\omega) = X_{N_t(\omega)}(\omega)$ and $N_t(\omega)$ piecewise constant in t for all $\omega \in \Omega$. Therefore $r(Z_t(\omega)) = r(X_n(\omega))$ for all $t \in [T_n(\omega), T_{n+1}(\omega))$ and it follows that

$$\mathbb{E}_s\left[\int_0^\infty \overline{r}(Z_t) P_T(T \geq t)\, dt\right] = \mathbb{E}_s\left[\sum_{n=0}^{\infty} \overline{r}(X_n) \int_{T_n}^{T_{n+1}} P_T(T \geq t)\, dt\right]. \qquad \square$$

Proof (of Theorem 4.3). Equation (4.16) can be established by multipliying (4.15) with $\alpha + E(s)$ and using $E(s) = -Q(s,s)$ when rearranging terms. Thus we only have to show (4.15). If s is absorbing then $Q(s,s') = 0$ for all s', $E(s) = 0$ and $P(t)(s,s') = \delta_{s,s'}$. The conclusion follows from (4.14) since $V^\alpha(s) = \int_0^\infty \overline{r}(s) e^{-\alpha t}\, dt = \frac{\overline{r}(s)}{\alpha}$. Assume in the following that s is non-absorbing and thus $E(s) > 0$. From Lemma 4.1 it holds that

$$V^\alpha(s) = \mathbb{E}_s\left[\sum_{n=0}^{\infty} \overline{r}(X_n) \int_{T_n}^{T_{n+1}} e^{-\alpha t}\, dt\right] = \mathbb{E}_s\left[\sum_{n=0}^{\infty} e^{-\alpha T_n} \overline{r}(X_n) \int_0^{\tau_n} e^{-\alpha t}\, dt\right],$$

since $T_{n+1} = T_n + \tau_n$. Define $R(X_n, \tau_n) := \overline{r}(X_n) \int_0^{\tau_n} e^{-\alpha t}\, dt$. Because τ_0 given $X_0 = s$ is exponentially distributed with rate $E(s) > 0$ it follows by Lemma A.1 that

$$\mathbb{E}_s\left[R(X_0, \tau_0)\right] = \frac{\overline{r}(s)}{\alpha + E(s)}$$

and thus

$$\begin{aligned}
V^\alpha(s) &= \mathbb{E}_s\left[\sum_{n=0}^{\infty} e^{-\alpha \sum_{k=0}^{n-1} \tau_k} R(X_n, \tau_n)\right] \\
&= \mathbb{E}_s\left[R(X_0, \tau_0)\right] + \mathbb{E}_s\left[e^{-\alpha\tau_0} \sum_{n=1}^{\infty} e^{-\alpha \sum_{k=1}^{n-1} \tau_k} R(X_n, \tau_n)\right] \\
&= \frac{\overline{r}(s)}{\alpha + E(s)} + \mathbb{E}_s\left[e^{-\alpha\tau_0} \sum_{n=0}^{\infty} e^{-\alpha \sum_{k=0}^{n-1} \tau_{k+1}} R(X_{n+1}, \tau_{n+1})\right] \\
&= \frac{\overline{r}(s)}{\alpha + E(s)} + \mathbb{E}\left[e^{-\alpha\tau_0} V^\alpha(X_1) \mid X_0 = s\right],
\end{aligned}$$

where $V^\alpha(X_1)$ is the random variable representing the discounted value when the process starts in X_1. Now since $V^\alpha(X_1)$ is independent of τ_0 (given $X_0 = s$) it follows that

$$\begin{aligned}
&\mathbb{E}\left[e^{-\alpha\tau_0} V^\alpha(X_1) \mid X_0 = s\right] = \mathbb{E}\left[e^{-\alpha\tau_0} \mid X_0 = s\right] \mathbb{E}\left[V^\alpha(X_1) \mid X_0 = s\right] \\
&= \int_0^\infty e^{-\alpha t} \cdot E(s) e^{-E(s)t}\, dt \cdot \sum_{s' \neq s} V^\alpha(s') P(s,s') = \frac{E(s)}{\alpha + E(s)} \sum_{s' \neq s} V^\alpha(s') P(s,s') \\
&= \sum_{s' \neq s} \frac{Q(s,s')}{\alpha + E(s)} V^\alpha(s'),
\end{aligned}$$

where the last equation follows from $Q(s,s') = P(s,s')E(s)$. $\square$

References

1. Altman, E.: Constrained Markov Decision Processes. Chapman & Hall (1999)
2. Altman, E.: Applications of Markov Decision Processes in Communication Networks. In: Feinberg, E.A., Shwartz, A. (eds.) Handbook of Markov Decision Processes. International Series in Operations Research & Management Science, vol. 40, pp. 489–536. Springer, US (2002)
3. Baier, C., Haverkort, B., Hermanns, H., Katoen, J.-P.: Model-Checking Algorithms for Continuous-Time Markov Chains. IEEE Transactions on Software Engineering 29(6), 524–541 (2003)
4. Bäuerle, N., Rieder, U.: Markov Decision Processes with Applications to Finance. Springer, Heidelberg (2011)
5. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
6. Benini, L., Bogliolo, A., Paleologo, G.A., De Micheli, G.: Policy Optimization for Dynamic Power Management. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 18, 813–833 (1998)
7. Bertsekas, D.: Dynamic Programming and Optimal Control, 3rd edn., vol. I. Athena Scientific (1995) (revised in 2005)
8. Bertsekas, D.: Dynamic Programming and Optimal Control, 4th edn., vol. II. Athena Scientific (1995) (revised in 2012)
9. Bertsekas, D., Tsitsiklis, J.: An analysis of stochastic shortest path problems. Mathematics of Operations Research 16(3), 580–595 (1991)
10. Bertsekas, D., Tsitsiklis, J.: Neuro-Dynamic Programming, 1st edn. Athena Scientific (1996)
11. Beynier, A., Mouaddib, A.I.: Decentralized Markov decision processes for handling temporal and resource constraints in a multiple robot system. In: Proceedings of the 7th International Symposium on Distributed Autonomous Robotic System, DARS (2004)
12. Bolch, G., Greiner, S., de Meer, H., Trivedi, K.S.: Queueing Networks and Markov Chains - Modelling and Performance Evaluation with Computer Science Applications, 2nd edn. Wiley (2006)
13. Cassandra, A.R.: A survey of POMDP applications. In: Working Notes of AAAI 1998 Fall Symposium on Planning with Partially Observable Markov Decision Processes, pp. 17–24 (1998)
14. Diz, F.J., Palacios, M.A., Arias, M.: MDPs in medicine: opportunities and challenges. In: Decision Making in Partially Observable, Uncertain Worlds: Exploring Insights from Multiple Communities, IJCAI Workshop (2011)
15. Fox, B.L., Landi, D.M.: An algorithm for identifying the ergodic subchains and transient states of a stochastic matrix. Communications of the ACM 11(9), 619–621 (1968)
16. Gouberman, A., Siegle, M.: On Lifetime Optimization of Boolean Parallel Systems with Erlang Repair Distributions. In: Operations Research Proceedings 2010 - Selected Papers of the Annual International Conference of the German Operations Research Society, pp. 187–192. Springer (January 2011)
17. Guo, X., Hernandez-Lerma, O.: Continuous-Time Markov Decision Processes - Theory and Applications. Springer (2009)
18. Heidergott, B., Hordijk, A., Van Uitert, M.: Series Expansions For Finite-State Markov Chains. Probability in the Engineering and Informational Sciences 21(3), 381–400 (2007)

19. Hou, Z., Filar, J.A., Chen, A. (eds.): Markov Processes and Controlled Markov Chains. Springer (2002)
20. Howard, R.A.: Dynamic Programming and Markov Processes. John Wiley & Sons, New York (1960)
21. Hu, Q., Yue, W.: Markov Decision Processes with their Applications. Springer (2008)
22. Janssen, J., Manca, R.: Markov and Semi-Markov Reward Processes. In: Applied Semi-Markov Processes, pp. 247–293. Springer, US (2006)
23. Janssen, J., Manca, R.: Semi-Markov Risk Models for Finance, Insurance and Reliability. Springer (2007)
24. Jensen, A.: Markoff chains as an aid in the study of Markoff processes. Skandinavisk Aktuarietidskrift 36, 87–91 (1953)
25. Stidham Jr., S., Weber, R.: A survey of Markov decision models for control of networks of queues. Queueing Systems 13(1-3), 291–314 (1993)
26. Mahadevan, S.: Learning Representation and Control in Markov Decision Processes: New Frontiers. Foundations and Trends in Machine Learning 1(4), 403–565 (2009)
27. Mahadevan, S., Maggioni, M.: Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes. Journal of Machine Learning Research 8, 2169–2231 (2007)
28. Mausam, Kolobov, A.: Planning with Markov Decision Processes: An AI Perspective. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers (2012)
29. Momtazi, S., Kafi, S., Beigy, H.: Solving Stochastic Path Problem: Particle Swarm Optimization Approach. In: Nguyen, N.T., Borzemski, L., Grzech, A., Ali, M. (eds.) IEA/AIE 2008. LNCS (LNAI), vol. 5027, pp. 590–600. Springer, Heidelberg (2008)
30. Obal, W.D., Sanders, W.H.: State-space support for path-based reward variables. In: Proceedings of the Third IEEE International Performance and Dependability Symposium on International Performance and Dependability Symposium, IPDS 1998, pp. 233–251. Elsevier Science Publishers B. V. (1999)
31. Ott, J.T.: A Markov Decision Model for a Surveillance Application and Risk-Sensitive Markov Decision Processes. PhD thesis, Karlsruhe Institute of Technology (2010)
32. Powell, W.B.: Approximate Dynamic Programming - Solving the Curses of Dimensionality. Wiley (2007)
33. Puterman, M.L.: Markov Decision Processes - Discrete Stochastic Dynamic Programming. John Wiley & Sons INC. (1994)
34. Qiu, Q., Pedram, M.: Dynamic power management based on continuous-time Markov decision processes. In: Proceedings of the 36th Annual ACM/IEEE Design Automation Conference, DAC 1999, pp. 555–561. ACM (1999)
35. Sanders, W.H., Meyer, J.F.: A Unified Approach for Specifying Measures of Performance, Dependability, and Performability. Dependable Computing for Critical Applications 4, 215–238 (1991)
36. Schaefer, A.J., Bailey, M.D., Shechter, S.M., Roberts, M.S.: Modeling medical treatment using Markov decision processes. In: Brandeau, M.L., Sainfort, F., Pierskalla, W.P. (eds.) Operations Research and Health Care. International Series in Operations Research & Management Science, vol. 70, pp. 593–612. Kluwer Academic Publishers (2005)
37. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. A Bradford Book. MIT Press (March 1998)

38. Trivedi, K.S., Malhotra, M.: Reliability and Performability Techniques and Tools: A Survey. In: Messung, Modellierung und Bewertung von Rechen- und Kommunikationssystemen. Informatik aktuell, pp. 27–48. Springer, Heidelberg (1993)
39. Tsitsiklis, J.N.: NP-Hardness of checking the unichain condition in average cost MDPs. Operations Research Letters 35(3), 319–323 (2007)
40. White, D.J.: A Survey of Applications of Markov Decision Processes. The Journal of the Operational Research Society 44(11), 1073–1096 (1993)
41. Wolff, R.W.: Poisson Arrivals See Time Averages. Operations Research 30(2), 223–231 (1982)