Compositional Reasoning for Markov Decision Processes
(Extended Abstract)

Yuxin Deng[1,2] ⋆ and Matthew Hennessy[3] ⋆⋆

[1] Dept. Comp. Sci. & Eng. and MOE-Microsoft Key Lab for Intell. Comp. & Syst., Shanghai Jiao Tong University, China
[2] State Key Lab of Comp. Sci., Inst. of Software, Chinese Academy of Sciences
[3] Trinity College Dublin, Ireland

Abstract. Markov decision processes (MDPs) have long been used to model qualitative aspects of systems in the presence of uncertainty. However, much of the literature on MDPs takes a monolithic approach, by modelling a system as a particular MDP; properties of the system are then inferred by analysis of that particular MDP. In this paper we develop compositional methods for reasoning about the qualitative behaviour of MDPs. We consider a class of labelled MDPs called weighted MDPs from a process algebraic point of view. For these we define a coinductive simulation-based behavioural preorder which is compositional in the sense that it is preserved by structural operators for constructing MDPs from components.

For finitary convergent processes, which are finite-state and finitely branching systems without divergence, we provide two characterisations of the behavioural preorder. The first uses a novel qualitative probabilistic logic, while the second is in terms of a novel form of testing, in which benefits are accrued during the execution of tests.

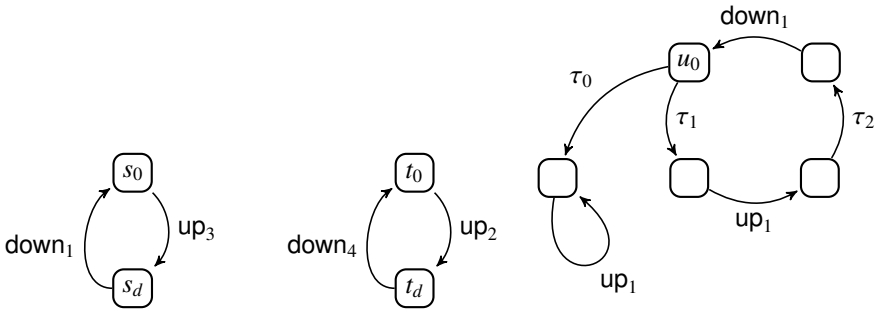
1 Introduction

Markov decision processes (MDPs) have long been used to model qualitative aspects of systems in the presence of uncertainty [12, 13, 1]. A comprehensive account of analysis techniques may be found in [12], while [13] provides a good account of *model-checking*.

However, much of the literature on MDPs takes a monolithic view of systems; essentially a system is modelled using a particular MDP, and properties of the system are then inferred by analysis of that MDP. In this paper, instead, we would like to develop compositional methods for reasoning about qualitative behaviour of Markov decision processes. This involves defining an appropriate method for comparing the behaviour MDPs which is susceptible to compositional analysis; the behaviour of a composite system should be determined by that of its components.

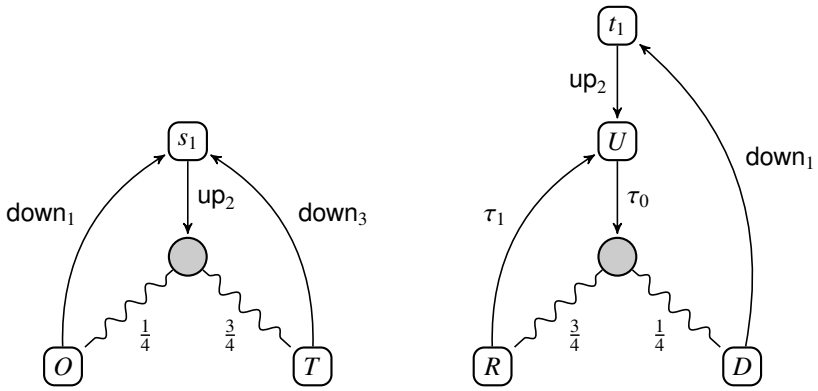
Our starting point is the idea of one system being able to *simulate* another. For example consider the following three systems:

⋆ Supported by the National Natural Science Foundation of China (61033002, 61011140074).
⋆⋆ Supported by SFI project SFI 06 IN.1 1898.



The first, a two-state machine, continually performs an up action, which accrues a benefit of 3 units, followed by a down action, which accrues a benefit of 1. The second machine performs the same actions but with benefits 2 and 4 respectively. In some sense t_0 is an improvement on s_0; intuitively t_0 can simulate the behaviour of s_0 but in so doing accrue more benefits; this is true even if one of its actions up is less beneficial than the corresponding action of s_0. The same is true for the machine u_0; it can also simulate the behaviour of s_0, with more benefit, although in this case some internal weighted actions, denoted by τ, participate in the simulation and add to the accumulation of benefits. In our terminology we will write $s_0 \sqsubseteq_{sim} t_0$, $s_0 \sqsubseteq_{sim} u_0$. However, we will have $t_0 \not\sqsubseteq_{sim} u_0$ because although u_0 can simulate the behaviour of t_0 it accumulates less benefit.

Similar informal reasoning can also be applied to probabilistic systems. Consider the following systems:

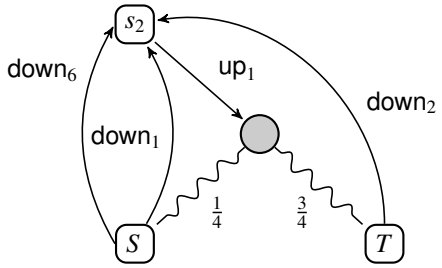


The first, from state s_1, can perform the up action with benefit 2 and a quarter of the time it ends up in a state in which down can be performed with benefit only 1. But for the remaining three-quarters it ends up in a state in which down can be performed for the larger benefit 3. The circular darkended node represents a distribution of states, with its outgoing edges describing the associated probabilities. Again intuitively we can see that s_1 is an improvement on s_0 because it can simulate s_0 and on average accrue slightly more benefits; in our theory we will have $s_0 \sqsubseteq_{sim} s_1$.

The mixture of probabilistic behaviour and internal actions introduces complications. Consider the system t_1 above which after performing an up probabilistically decides internally whether to perform a down action for benefit 1, or branch back to make

another probabilistic choice. However, each time it reverts back it accumulates a non-zero benefit via the internal weighted action τ_1, albeit with diminishing probability. Nevertheless, it will turn out that $s_0 \sqsubseteq_{sim} t_1$ and indeed $s_1 \sqsubseteq_{sim} t_1$.

Systems exhibiting both probabilistic and nondeterministic behaviour require more complicated analysis. Consider the following system:



After performing the action up it finds itself either in a state in which the action down will accrue the benefit 2, or 25% of the time there will be a nondeterministic choice between it accruing either 1 or 6. In the literature there are numerous mechanisms, such as policies, schedulers, adversaries, etc. [12, 14, 13] for resolving such choices. Here one can see if this choice systematically leads to the lower benefit 1 then s_2 will not simulate s_0 as it does not accrue sufficient benefits. This is a pessimistic outlook; an optimistic outlook means that the best choices are systematically made. If this is assumed then we will have $s_0 \sqsubseteq_{sim} s_2$; in s_2 one execution of up followed by down will yield on average the benefit $1 + (\frac{3}{4} \cdot 2 + \frac{1}{4} \cdot 6) = 4$.

The main contribution of the paper is a coinductively defined behavioural preorder $\sqsubseteq_{sim}$ between MDPs based on simulations which validate the examples discussed informally above. We confine our attention to the optimistic approach to the resolution of nondeterministic choices, although in a later paper we hope to investigate the pessimistic approach. We also show that this preorder is compositional in the sense that it is preserved by structural operators for constructing MDPs from components. The main operator is one for composing two MDPs in parallel. In $P \mid Q$ the two MPDs P and Q remain independent, execute in parallel and may communicate by synchronising on complementary actions; these internal synchronisations accrue the combined benefits of the associated complementary actions.

For finitary convergent MDPs, which are finite-state and finitely branching systems without divergence, we also provide two characterisations for the behavioural preorder $\sqsubseteq_{sim}$. The first is in terms of a *qualitative probabilistic logic* $\mathcal{L}_Q$. In addition to the standard logical connectives $\wedge$, $\vee$ and both maximal and minimal fixpoints this contains a novel qualitative *possibility* modality $\langle\alpha\rangle_w(\phi_1 \,{}_p\!\oplus\, \phi_2)$, where p is some probability between 0 and 1. Intuitively this is satisfied by an MDP which can accrue at least the benefit w by performing the action α, and subsequently satisfy the probabilistic assertion $\phi_1 \,{}_p\!\oplus\, \phi_2$. It turns out that the simulation preorder is completely determined by the logic $\mathcal{L}_Q$. Further evidence of the compatibility between the logic and the simulation relation is the fact that every system P has a *characteristic formula* $\phi(P)$ in the logic which captures its behaviour; informally system Q can simulate P if and only if it satisfies the characteristic formula $\phi(P)$.

where $s \overset{\tau}{\not\mapsto}$ *means that* s *cannot enable any* τ*-transition. We say* Φ *is an* extreme derivative *of* Δ*, with weight* w. □

Intuitively every extreme derivation $\Delta \overset{\tau}{\Longrightarrow}_w \Phi$ represents a computation from the initial distribution Δ guided by some implicit scheduler. For example, consider the hyper-derivation of an extreme derivative:

$$\begin{array}{rcl} \Delta & = & \Delta_0^{\rightarrow} + \Delta_0^{\times} \\ \Delta_0^{\rightarrow} & \overset{\tau}{\longrightarrow}_{w_0} & \Delta_1^{\rightarrow} + \Delta_1^{\times} \\ & \vdots & \\ \Delta_k^{\rightarrow} & \overset{\tau}{\longrightarrow}_{w_k} & \Delta_{k+1}^{\rightarrow} + \Delta_{k+1}^{\times} \\ & \vdots & \\ \Phi & = & \displaystyle\sum_{k=0}^{\infty} \Delta_k^{\times} \end{array} \tag{3}$$

where $w = \sum_{k\geq 0} w_k$. Initially, since $\Delta_0^{\times}$ is stable, $\Delta_0^{\rightarrow}$ contains (in its support) all states which can proceed with the computation. The implicit scheduler decides for each of these states which step to take, cumulating in the first move, $\Delta_0^{\rightarrow} \overset{\tau}{\longrightarrow}_{w_0} \Delta_1^{\rightarrow} + \Delta_1^{\times}$. At an arbitrary stage, $\Delta_k^{\rightarrow}$ contains all states which can continue; the scheduler decides which step to take for each individual state and the overall result of the schedulers decision for this stage is captured in the step $\Delta_k^{\rightarrow} \overset{\tau}{\longrightarrow}_{w_k} \Delta_{k+1}^{\rightarrow} + \Delta_{k+1}^{\times}$.

Example 5. Referring to Figure 2(a) it is easy to see that $\overline{s}$ has a unique (degenerate) extreme derivative, $\overline{s_1} \overset{\tau}{\Longrightarrow}_{9.5} (\frac{1}{4}\overline{s_l} + \frac{3}{4}\overline{s_r})$, intuitively representing the unique weighted computation from $\overline{s_1}$. However, consider the wMDP in Figure 2(b), in which there is a nondeterministic choice from state t_2; here the extreme derivatives generated from $\overline{t}$, and their associated weights, will depend on the choices made during the computation by the implicit scheduler.

First suppose that the scheduler uses the static policy which maps t_2 to $\langle\, 12, \overline{t_4} \,\rangle$. Then it is easy to see that the generated extreme derivative, which is degenerate, is $\overline{t} \overset{\tau}{\Longrightarrow}_{12} (\frac{3}{4}\overline{t_4} + \frac{1}{4}\overline{t_5})$. However using the static policy which maps t_2 to $\langle\, 4, \overline{t_1} \,\rangle$ we generate, using (3), a non-degenerate extreme derivative; after some calculations this can be seen to be $\overline{t_1} \overset{\tau}{\Longrightarrow}_{24} \overline{t_5}$.

However there are many other possible implicit schedulers, for example at different times in the computations employing either of these static policies, or even choosing nondeterministically between them. But these are the only static policies and therefore we know from Theorem 1 that if $\overline{t_1} \overset{\tau}{\Longrightarrow}_w \Delta$ then w must take the form $p \cdot 12 + (1-p) \cdot 24$ for some $0 \leq p \leq 1$. That is the set of benefits which can be generated from $\overline{t_1}$ is $\{\, 24 - 12 \cdot p \mid 0 \leq p \leq 1 \,\}$. □

Definition 9 (May testing). *In a wMDP, for any* $\Delta \in \mathcal{D}(S)$*, let*

$$\textbf{Benefits}(\Delta) = \{\, w \in \mathbb{R}_{\geq 0} \mid \Delta \overset{\tau}{\Longrightarrow}_w \Phi, \textit{ for some } \Phi \in \mathcal{D}_{sub}(S) \,\}$$

Benefit sets are compared as follows:

$$B_1 \leq^r_{\mathrm{Ho}} B_2 \text{ if for every } r_1 \in B_1 \text{ there exists some } r_2 \in B_2 \text{ such that } r_1 \leq r + r_2$$

For any two distributions Δ, Θ we write $\Delta \sqsubseteq^r_{may} \Theta$ if for every finite (testing) process T, **Benefits**$(\Delta \parallel T) \leq^r_{\mathrm{Ho}}$ **Benefits**$(\Theta \parallel T)$. *We write $\Delta \sqsubseteq_{may} \Theta$ to mean that there is some $r \in \mathbb{R}_{\geq 0}$ such that $\Delta \sqsubseteq^r_{may} \Theta$.* □

This interpretation of processes is optimistic; $\Delta \sqsubseteq^r_{\mathrm{may}} \Theta$ means that, given the investment r, every possible benefit produced by Δ can in principle be improved upon by Θ.

Note that in a bounded wMDP **Benefits**(Δ) cannot contain ∞. Moreover we can show that the parallel composition of a bounded wMDP with a finite wMDP is also bounded. This means that if we confine our attention to bounded wMDPs then benefit sets will always only contain real numbers. One way of restricting to bounded wMDPS is, by Theorem 2, to only use finitary convergent wMDPs.

Our first result shows that simulations can be used as a sound proof technique for this semantics:

Theorem 7 (Soundness). *$\Delta \mathrel{\overline{\lhd}}_r \Theta$ implies $\Delta \sqsubseteq^r_{may} \Theta$.* □

The converse is not true in general:

Example 6. Consider the two distributions $\Delta = \mathbf{0} \mathrel{{}_{\frac{1}{2}}\!\oplus} a_1.\mathbf{0}$ and $\Theta = \tau_2.\mathbf{0} \mathrel{{}_{\frac{1}{2}}\!\oplus} a_0.\mathbf{0}$. It is easy to see that $\Delta \mathrel{\overline{\not\lhd}_0} \Theta$ because there is no way to decompose Θ into $\Theta_1 \mathrel{{}_{\frac{1}{2}}\!\oplus} \Theta_2$ for some Θ_1, Θ_2 such that $a_1.\mathbf{0} \lhd_0 \Theta_2$. However, one can show that $\Delta \sqsubseteq^0_{\mathrm{may}} \Theta$. This follows from the observations below:

(i) For all weights w and tests T, **Benefits**$(\tau_w.\mathbf{0} \parallel T) = \{v + w \mid v \in$ **Benefits**$(\mathbf{0} \parallel T)\}$.
(ii) For all weights w and tests T, **Benefits**$(a_w.\mathbf{0} \parallel T) \leq^w_{\mathrm{Ho}}$ **Benefits**$(a_0.\mathbf{0} \parallel T)$.

Both assertions can be proved by structural induction on T.

Now suppose $w \in$ **Benefits**$(\Delta \parallel T)$ for an arbitrary test T. There is some stable derivative Γ such that $\Delta \parallel T \stackrel{\tau}{\Longrightarrow}_w \Gamma$. It can be shown that there are some $w_1, w_2, \Gamma_1, \Gamma_2$ with $\mathbf{0} \parallel T \stackrel{\tau}{\Longrightarrow}_{w_1} \Gamma_1$, $a_1.\mathbf{0} \parallel T \stackrel{\tau}{\Longrightarrow}_{w_2} \Gamma_2$, $w = \frac{1}{2}w_1 + \frac{1}{2}w_2$, and $\Gamma = \frac{1}{2} \cdot \Gamma_1 + \frac{1}{2} \cdot \Gamma_2$, where both Γ_1 and Γ_2 are stable. In other words, we have $w_1 \in$ **Benefits**$(\mathbf{0} \parallel T)$ and $w_2 \in$ **Benefits**$(a_1.\mathbf{0} \parallel T)$. By (i) above, $w_1 + 2 \in$ **Benefits**$(\tau_2.\mathbf{0} \parallel T)$; by (ii) above, there exists some $w_2' \in$ **Benefits**$(a_0.\mathbf{0} \parallel T)$ with $w_2 \leq w_2' + 1$. Thus, we can infer that

$$\begin{aligned} w &= \tfrac{1}{2}w_1 + \tfrac{1}{2}w_2 \\ &< \tfrac{1}{2}(w_1 + 2) + \tfrac{1}{2}(w_2 - 1) \\ &\leq \tfrac{1}{2}(w_1 + 2) + \tfrac{1}{2}w_2' \end{aligned}$$

It turns out that $\frac{1}{2}(w_1 + 2) + \frac{1}{2}w_2' \in$ **Benefits**$(\Theta \parallel T)$. Therefore, we have

$$\textbf{Benefits}(\Delta \parallel T) \leq^0_{\mathrm{Ho}} \textbf{Benefits}(\Theta \parallel T).$$

Since this reasoning is carried out for an arbitrary test T, it follows that $\Delta \sqsubseteq^0_{\mathrm{may}} \Theta$. □

Nevertheless we do have a testing characterisation for the unannotated simulation preorder:

Theorem 8 (Testing characterisation). *In a bounded wMDP,* $\overline{s} \sqsubseteq_{may} \Theta$ *if and only if* $\overline{s} \sqsubseteq_{sim} \Theta$.

Proof (Outline). One direction follows from Theorem 7. For the converse we carry out the proof in two steps: we first prove that $\overline{s} \sqsubseteq^{r}_{\text{may}} \Theta$ implies the existence of some compensation $r' \geq r$ with $\mathcal{L}_Q(0, \overline{s}) \subseteq \mathcal{L}_Q(r', \Theta)$, then appeal to Theorem 5. In the first step we proceed by constructing, for each formula ϕ, a *characteristic test* $T(\phi)$, such that if a process satisfies ϕ then it passes test $T(\phi)$ with some threshold benefit. □

An alternative approach to testing would be to use one special action ω in a test to report success and when applying such a test to a system to report the weighted average of the weight of each path leading to an occurrence of the success action; this we refer to as *expected benefits testing*. Here we will not give the formal definition of how these *expected benefits* are calculated, which is provided in [4], but simply give an informal argument to show that our simulation preorder is not sound with respect to it.

Example 7 (Simulation is unsound for expected benefits testing). Consider the following processes:

$$P = \tau_2.(\mathbf{0} \,{}_{\frac{1}{4}}\!\oplus a_0.\mathbf{0})$$
$$Q = \tau_1.(\tau_2.(\mathbf{0} \,{}_{\frac{1}{2}}\!\oplus a_0.\mathbf{0}) \,{}_{\frac{1}{2}}\!\oplus a_0.\mathbf{0})$$

It is easy to see that $P \lhd_0 Q$ as the transition $P \xrightarrow{\tau}_2 \mathbf{0} \,{}_{\frac{1}{4}}\!\oplus a_0.\mathbf{0}$ can be simulated by the hyper-derivative $Q \stackrel{\tau}{\Longrightarrow}_2 \mathbf{0} \,{}_{\frac{1}{4}}\!\oplus a_0.\mathbf{0}$. Now let T be the test $\bar{a}_0.\omega$. Both $P \parallel T$ and $Q \parallel T$ give rise to fully probabilistic wMDPs. The unique expected benefit resulted from $P \parallel T$ is $\frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 2$, i.e. $\frac{3}{2}$. On the other hand, the unique expected benefit obtained from $Q \parallel T$ is $\frac{1}{2} \cdot 1 + \frac{1}{2}(\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 3)$, i.e. $\frac{5}{4}$. As $\{\frac{3}{2}\} \not\leq^{0}_{\text{Ho}} \{\frac{5}{4}\}$, we have that P is not related to Q under expected benefits may testing; thus $\lhd$ is unsound for expected benefits testing. Note that if we consider total benefits, then $\mathbf{Benefits}(P \parallel T) = \{2\} = \mathbf{Benefits}(Q \parallel T)$. □

5 Concluding remarks

We have proposed the model of weighted Markov decision processes for compositional reasoning about the behaviour of systems with uncertainty. Amortised weighted simulation is coinductively defined to be a behavioural preorder for comparing different wMDPs. It is shown to be a precongruence relation with respect to all structural operators for constructing wMDPs from components. For finitary convergent wMDPs, we have also given logical and testing characterisations of the simulation preorder: it can be completely determined by a qualitative probabilistic logic and for each system we can find a characteristic formula to capture its behaviour; the simulation preorder also coincides with a notion of may testing preorder.

The dual of may testing is must testing. It would be interesting to investigate the must preorder given by our testing approach. We leave it as future work to provide a coinductive formulation of the preorder and study its logical characterisations.

There is a very limited literature on compositional theories of Markov decision processes particularly in the presence of weights. There is however an extensive literature on probabilistic variations of bisimulation equivalence for Markov chains; see Chapter 10 of [1] for an elementary introduction and [10] for a survey. Bisimulation equivalence has also been defined in [8] for *Interactive Markov Chains (IMCs)*, and it is shown to be compositional, in the sense of our Theorem 4: it is preserved by the operators of a process calculus interpreted as IMCs. Bisimulation and testing equivalence for Markovian process algebras are also investigated in [9, 2], but the analysis was mainly restricted to models free of nondeterminism. Recently a combination of probabilistic automata and IMCs has been studied in [7], where a notation of weak bisimulation is proposed. Since time rates are treated essentially as action names, some intuitively equivalent processes are differentiated by the weak bisimulation.

There is also an extensive literature on weighted automata [6], and probabilistic variations have also been studied [3]. However there the focus is on traditional language theoretic issues, rather than our primary concern, compositionality.

References

1. C. Baier and J.-P. Katoen. *Principles of Model Checking*. The MIT Press, 2008.
2. M. Bernardo and R. Cleaveland. A theory of testing for Markovian processes. In Proc. *CONCUR'00*, LNCS 1877, pp. 305-319. Springer, 2000.
3. K. Chatterjee, L. Doyen, and T.A. Henzinger. Probabilistic Weighted Automata. In Proc. *CONCUR'09*, LNCS 5710, pp. 244–258. Springer, 2009.
4. Y. Deng and M. Hennessy. Compositional reasoning for Markov decision processes (full version), 2010. Available at `http://basics.sjtu.edu.cn/~yuxin/temp/mdp.pdf`.
5. Y. Deng, R. van Glabbeek, M. Hennessy, and C. Morgan. Testing finitary probabilistic processes. In Proc. *CONCUR'09*, LNCS 5710, pp. 274–288. Springer, 2009.
6. M. Droste, W. Kuich, and H. Vogler (Eds.) *Handbook of Weighted Automata*. Springer, 2009.
7. E. Eisentraut, H. Hermanns, and L. Zhang. On probabilistic automata in continuous time. In Proc. *LICS'10*, pp. 342-351. IEEE Computer Society, 2010.
8. H. Hermanns. *Interactive Markov Chains: The Quest for Quantified Quality*, LNCS 2428. Springer, 2002.
9. J. Hillston. *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.
10. B. Jonsson, K.G. Larsen, and Y. Wang. Probabilistic Extensions of Process Algebras. In *Handbook of Process Algebra*, pp. 685–710. Elsevier, 2001.
11. A. Kiehn and S. Arun-Kumar. Amortised bisimulations. In Proc. *FORTE'05*, LNCS 3731, pp. 320–334. Springer, 2005.
12. M. Puterman. *Markov Decision Processes*. Wiley, 1994.
13. J. Rutten, M. Kwiatkowska, G. Norman, and D. Parker. *Mathematical Techniques for Analyzing Concurrent and Probabilistic Systems.* American Mathematical Society, 2004.
14. R. Segala. Modeling and verification of randomized distributed real-time systems. Technical Report MIT/LCS/TR-676, PhD thesis, MIT, Dept. of EECS, 1995.