A Formalism for Stochastic Decision Processes with Asynchronous Events

Håkan L. S. Younes and **Reid G. Simmons**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{lorens, reids}@cs.cmu.edu

Abstract

We present the generalized semi-Markov decision process (GSMDP) as a natural model for stochastic decision processes with asynchronous events in hope to spur interest in asynchronous models, often overlooked in AI literature.

Introduction

Stochastic processes with *asynchronous* events and actions have received little attention in the AI literature despite there being an abundance of asynchronous systems in the real world. The canonical example of an asynchronous process is a simple queuing system with a single service station, for example modeling your local post office. Customers arrive at the post office, wait in line until the service station is vacant, spend some time being serviced by the clerk, and finally leave. We can think of the arrival and departure (due to service completion) of a customer as two separate *events*. There is no synchronization between the arrival and departure of customers, i.e. the two events just introduced are *asynchronous*, so this is clearly an example of an asynchronous system.

Some attention has recently been given to planning with *concurrent* actions. Guestrin, Koller, & Parr (2002) and Mausam & Weld (2004) use discrete-time *Markov decision processes* (MDPs) to model and solve planning problems with concurrent actions, but the approach is restricted to instantaneous actions executed in synchrony. Rohanimanesh & Mahadevan (2001) consider planning problems with temporally extended actions that can be executed in parallel. By restricting the temporally extended actions to *Markov options*, the resulting planning problems can be modeled as discrete-time *semi*-Markov decision processes (SMDPs).

All three of the approaches cited above model time as a discrete quantity. This is a natural model of time for synchronous systems driven by a global clock. Asynchronous systems, on the other hand, are best represented using a dense (continuous) model of time (Alur, Courcoubetis, & Dill 1993). Continuous-time MDPs (Howard 1960) can be used to model asynchronous systems, but are restricted to events and actions with exponential trigger time distributions. Continuous-time SMDPs (Howard 1971) lift the restriction on trigger time distributions, but cannot model asynchrony.

We therefore propose the *generalized* semi-Markov decision process (GSMDP), based on the GSMP model of discrete event systems (Glynn 1989), as a model for asynchronous stochastic decision processes. A GSMDP, unlike an SMDP, remembers if an event enabled in the current state has been continuously enabled in previous states without triggering. This is key in modeling asynchronous processes, which typically involve events that race to trigger first in a state, but the event that triggers first does not necessarily disable the competing events. For example, if a customer is currently being serviced at the post office, the fact that another customer arrives does not mean that the service of the first customer has to start over from scratch. Rohanimanesh & Mahadevan (2001) note that if they were to allow semi-Markov options in their models, the resulting global models would no longer be SMDPs for the same reason. It should therefore be clear to the reader that the GSMDP formalism is a true generalization of the SMDP formalism. Intuitively, a GSMDP can be viewed as the composition of asynchronous SMDPs.

Generalized Semi-Markov Decision Processes

The generalized semi-Markov process (GSMP), first introduced by Matthes (1962), is an established formalism in queuing theory for modeling continuous-time stochastic discrete event systems (Glynn 1989). We add a decision dimension to the formalism by distinguishing a subset of the events as controllable and adding rewards, thereby obtaining the generalized semi-Markov *decision* process (GSMDP).

A Model of Stochastic Discrete Event Systems

A GSMP consists of a countable set of states S and a finite set of events E. At any point in time, the process occupies some state $s \in S$ in which a subset E_s of the events are enabled. With each event $e \in E$ is associated a positive distribution G_e governing the time e must remain enabled before it triggers, and a next-state probability distribution $p_e(s'|s)$. The enabled events in a state race to trigger first, and the event that triggers causes a transition to a state $s' \in S$ according to the next-state probability distribution for the triggering event. The time we spend in a specific state s before an event occurs is a random variable T_s. A GSMP is

a semi-Markov process only if the distribution of T_s, for all $s \in S$, is independent of history.

As an example of a GSMP, consider the post office "system" mentioned in the introduction. The state of this simple queuing system is the number of customers currently in the post office. There are two events representing customer arrival and customer departure, respectively. The arrival event is always enabled and the distribution associated with this event represents the inter-arrival time for customers. The departure event is only enabled when there are customers in the post office. The distribution associated with the departure event represents the time it takes to service a single customer. When the departure event triggers, the state changes from n to $n-1$. The arrival event causes a transition from state n to $n+1$, unless the post office is full in which case the state does not change.

Formal Semantics. To formally define the semantics of a GSMP model, we associate a real-valued clock t_e with each event that indicates the time remaining until e is scheduled to trigger in the current state. The process starts in some initial state s with events E_s enabled. For each enabled event $e \in E_s$, we sample a trigger time according to the distribution G_e and set t_e to the sampled value. For disabled events, we set $t_e = \infty$. Let e^* be the event in E_s with the smallest clock value, i.e. $e^* = \arg\min_{e \in E_s} t_e$. The event e^* becomes the triggering event in s. Provided that all trigger time distributions are continuous, the probability of two events triggering at exactly the same time is zero so e^* is uniquely defined. When e^* triggers after t_{e^*} time units in s, we sample a next state s' according to $p_{e^*}(s'|s)$ and update each clock t_e as follows:

1. if $e \in E_{s'} \cap \big(\{e^*\} \cup \big(E \setminus E_s\big)\big)$, then t'_e is sampled from G_e;
2. if $e \in E_{s'} \cap \big(E_s \setminus \{e^*\}\big)$, then $t'_e = t_e - t_{e^*}$;
3. otherwise, if $e \notin E_{s'}$ then $t'_e = \infty$.

The first rule covers events that are enabled in s' and either triggered or were not enabled in s. All such events are rescheduled. Events that remain enabled across state transitions without triggering are not rescheduled (rule 2). It is this rule that introduces history dependence and therefore breaks the semi-Markov property, thus a GSMP is not necessarily a semi-Markov process. The third and final rule states that events disabled in s' are scheduled not to trigger. Given a new state s' and new clock values t'_e for each $e \in E$, we repeat the procedure just specified with $s = s'$ and $t_e = t'_e$ so long as $E_s \neq \emptyset$.

By adding the clocks to the description of states we obtain an extended state-space $X \subset S \times \mathbb{R}^{|E|}_{\geq 0}$. Given an extended state $x \in X$, the next-state distribution over X is well-defined, which means that we can define a Markov chain with state-space X that corresponds to a GSMP with state-space S. This will be a *general state-space* Markov chain (GSSMC; Shedler 1993) because the state-space has both discrete and continuous components. Let $f_e(t)$ be the probability density function for the distribution G_e associated with event e. The next-state distribution for the GSSMC is defined as $f(x'|x) = p_{e^*}(s'|s) \prod_{e \in E} \tilde{f}_e(t'_e|s', x)$, where $\tilde{f}_e(t'_e|s', x)$ is

1. $f_e(t'_e)$, if $e \in E_{s'} \cap \big(\{e^*\} \cup \big(E \setminus E_s\big)\big)$;
2. $\delta(t'_e - (t_e - t_{e^*}))$, if $e \in E_{s'} \cap \big(E_s \setminus \{e^*\}\big)$;
3. $\delta(t'_e - \infty)$, if $e \notin E_{s'}$.

Here, $\delta(t - t_0)$ is the Dirac delta function (Dirac 1927, p. 625) with the property that $\int_{-\infty}^{x} \delta(t - t_0)\,dt$ is 0 for $x < t_0$ and 1 for $x \geq t_0$. In particular, $\int_{-\infty}^{x} \delta(t - \infty)\,dt$ is 0 for any finite x and 1 for $x = \infty$.

Observation Model. In general, the future trigger times of enabled events are not known to an observer of the process. Only the discrete part of the state-space is fully observable. However, the time that an event has been enabled is know to an observer, and this information is sufficient to provide the observer with a probability distribution over extended states.

Let $O \subset S \times \mathbb{R}^{|E|}_{\geq 0}$ be the set of observations. An observation $o = \langle s, \vec{u} \rangle \in O$ consists of s, the observed discrete part of the current extended state, and a vector $\vec{u}$ with elements u_e for each $e \in E$ being the time that event e has been enabled ($u_e = 0$ if $e \notin E_s$). Given an observation $o = \langle s, \vec{u} \rangle$, a probability density function $f(x|o)$ over X is defined as $f(x|o) = \prod_{e \in E} \tilde{f}_e(t_e | t_e > u_e, s)$ if $x = \langle s, \vec{t} \rangle$ and $f(x|o) = 0$ otherwise, where $\tilde{f}_e(t_e | t_e > u_e, s)$ is $f_e(t_e | t_e > u_e)$ if $e \in E_s$ and $\delta(t_e - \infty)$ otherwise.

Clearly, u_e is only significant for $e \in E_s$. Furthermore, if the distribution G_e associated with e is memoryless, i.e. $f_e(t | t > t_0) = f_e(t)$ as is the case for the exponential distribution, we do not need to know for how long e has been enabled. Thus, an observation only needs to consist of s and u_e for all $e \in E_s$ such that G_e is not a memoryless distribution. A GSMP with all events associated with an exponential distribution is simply a continuous-time Markov chain (Glynn 1989).

We define a function $obs : X \times O \times S \to O$ that given an extended state x, an observation of x, and the observable part s' of a successor x' of x, provides the observation of x'. We have $obs(x, o, s') = \langle s', \vec{u}' \rangle$, where $\vec{u}'$ consists of elements u'_e for each $e \in E$, with u'_e being

1. $u_e + t_{e^*}$, if $e \in E_{s'} \cap \big(E_s \setminus \{e^*\}\big)$;
2. 0 otherwise.

The first case covers events that remain enabled across state transitions without triggering. The time that e has remained enabled is simply the time it had remained enabled when entering state s (u_e) plus the time spent in s (t_{e^*}). The second case covers events that were not previously enabled or just triggered. Clearly, these events have not been enabled without triggering so the observation is 0 in this case. Note that the continuous component of x' is irrelevant to the observation of x'.

We could of course record the time an event has been enabled in the extended state rather than the trigger time of the event. An extended state would in that case be fully observable, but the result would be a general state-space *semi*-Markov process instead of a Markov chain.

$$
\begin{aligned}
v_\alpha^\pi(o) &= \int_X f(x|o)\left(\int_0^{t_{e^*}} e^{-\alpha t} c(s,\pi(o))\,dt + e^{-\alpha t_{e^*}} \int_X f(x'|x,o)\big(k(s,e^*,s') + v_\alpha^\pi(obs(x,o,s'))\big)\,dx'\right)dx \\
&= \int_X f(x|o)\left(\frac{1}{\alpha}\big(1-e^{-\alpha t_{e^*}}\big)c(s,\pi(o)) + e^{-\alpha t_{e^*}}\left(\hat{k}(s,e^*) + \sum_{s'\in S} p_{e^*}(s'|s) v_\alpha^\pi(obs(x,o,s'))\right)\right)dx
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
v_\alpha^\pi(s) &= \int_0^\infty \lambda_s^\pi e^{-\lambda_s^\pi t} \sum_{e\in E_s^\pi} \frac{\lambda_e}{\lambda_s^\pi}\left(\frac{1}{\alpha}\big(1-e^{-\alpha t}\big)c(s,\pi(s)) + e^{-\alpha t}\left(\hat{k}(s,e) + \sum_{s'\in S} p_e(s'|s) v_\alpha^\pi(obs(s'))\right)\right)dt \\
&= \frac{1}{\lambda_s^\pi + \alpha}\left(c(s,\pi(s)) + \sum_{e\in E_s^\pi} \lambda_e\left(\hat{k}(s,e) + \sum_{s'\in S} p_e(s'|s) v_\alpha^\pi(obs(s'))\right)\right)
\end{aligned} \tag{2}
$$

Actions, Policies, and Rewards

Given a GSMP with event set E, we identify a set $A \subset E$ of controllable events, or *actions*. The remaining events are called *exogenous events*. Actions differ from exogenous events in that they can be disabled at will in a state, while an exogenous event e always remains enabled in a state s if $e \in E_s$. A control policy π determines which actions should be enabled at a given time in a state. We allow the action choice to depend on the entire execution history of the process, which can be captured in an observation $o \in O$ as described above. Thus, a policy is a mapping from observations to sets of actions: $\pi : O \to 2^A$. A GSMDP controlled by a policy π is a GSSMC with E_s replaced by $E^\pi(o) = \pi(o) \cup \big(E_s \setminus A\big)$ in the definition of e^*, $f(x|o)$, and $obs(x,o,s')$. The next-state distribution is redefined as $f(x'|x,o) = p_{e^*}(s'|s)\prod_{e\in E} \tilde{f}_e(t'_e|s',x,o)$, where $\tilde{f}_e(t'_e|s',x,o)$ is defined as $\tilde{f}_e(t'_e|s',x)$ with $E^\pi(o)$ replacing E_s and $E^\pi_{obs(x,o,s')}$ replacing $E_{s'}$.

For the post office example, we could make the departure event into an action. This would signify that we can open and close the service station at will. If we close the service station (i.e. disable the departure event) while a customer is being serviced, the time we have spent with the customer is forgotten and the customer must be serviced from scratch if we reopen the service station (i.e. enable the departure event).

In addition to actions, we specify a reward structure to obtain a GSMDP. We assume a traditional reward structure with a lump sum reward $k(s,e,s')$ associated with the transition from state s to s' caused by the triggering of event e, and a continuous reward rate $c(s,A')$ associated with set of actions $A' \subset A$ being enabled in s (cf. Puterman 1994). The expected lump sum reward if event e triggers in state s is $\hat{k}(s,e) = \sum_{s'\in S} p_e(s'|s)k(s,e,s')$.

The expected infinite-horizon discounted value of an observation o for a policy π is given by (1). The parameter α is the *discount rate*, which can be interpreted as the rate of a termination event with exponential trigger time distribution (Howard 1960). It means that a unit reward earned t time units into the future counts as a $e^{-\alpha t}$ reward at present time. Note that if rewards were allowed to depend on the extended state x of the process, and not only on the real state s, we would not be able to get rid of the nested integrations in (1).

Now, let s be a state such that each event $e \in E_s^\pi$ is associated with an exponential distribution having rate λ_e: $G_e = Exp(\lambda_e)$. Let $\lambda_s^\pi = \sum_{e\in E_s^\pi} \lambda_e$. The time spent in s before an event triggers is then a random variable with distribution $Exp(\lambda_s^\pi)$ and the probability that a specific event e triggers first is λ_e/λ_s^π. These nice properties of the exponential distribution allows us to write the expected infinite-horizon discounted value of s as (2). We write $obs(s')$ for the observation of the next state because it is independent of the current state. If all events are associated with memoryless distributions, then $obs(s')$ can be replaced with s' in (2), which then represents an alternative formulation for continuous-time Markov decision processes.

Discussion

Unless we make limiting assumptions regarding a GSMDP model, for example that all distributions are memoryless, then we most likely have to resort to approximation schemes in order to solve the GSMDP. A straightforward approach would be to discretize time, however, we will suffer greatly from the curse of dimensionality if we do so naively.

Younes & Simmons (2004) present a technique for approximating a GSMDP with a continuous-time MDP by approximating each distribution G_e with a continuous *phase-type distribution* (Neuts 1981). The continuous-time MDP can then be solved using standard techniques. The approximation essentially amounts to a discretization into random-length intervals of the observation for how long an event has been enabled. The length of an interval is a random variable with an exponential distribution.

Alternatively, we could use *discrete* phase-type distributions (Neuts 1975) to obtain a discrete-time MDP that approximates our GSMDP. Bobbio *et al.* (2003) describe an algorithm for approximating an arbitrary positive distribution with a discrete phase-type distribution, which could be used for the purpose of approximating a GSMDP with a discrete-time MDP. One clear advantage with this approach over using continuous phase-type distributions is that determinis-

tic distributions can be represented exactly. A disadvantage with approximating a GSMDP with a discrete-time model is that we would have to take into account the possibility of two events triggering at the same time, and it may not be immediately obvious what to do in such a case.

A third possibility is of course to use a function approximator, for example k-nearest neighbor, to represent the value function of a GSMDP. The GSMDP could then be solved using fitted value iteration (Gordon 1995).

Even if we cannot hope to find optimal policies for GSMDPs, it may still be worthwhile trying to determine characteristics of optimal policies. We have defined a policy as a mapping from observations to sets of actions, where an observation includes a clock value for each enabled event. This means that actions can be enabled and disabled at any point in time while in a specific state and not only at the triggering of an event or action. If all trigger time distributions are exponential, i.e. if the GSMDP is a continuous-time MDP, then we do not need to take into consideration the time that events have been enabled in order to maximize the expected infinite-horizon discounted reward. The SMDP case, when the action choice can change between transition, has been analyzed by Chitgopekar (1969), Stone (1973), and Cantaluppi (1984) under various assumptions. An analysis of this sort would be valuable for the general case with asynchronous events as well, and would ideally provide us with conditions for the trigger time distributions under which the optimal policy has a certain structure (for example piecewise constant).

References

Alur, R.; Courcoubetis, C.; and Dill, D. 1993. Model-checking in dense real-time. *Information and Computation* 104(1):2–34.

Bobbio, A.; Horváth, A.; Scrapa, M.; and Telek, M. 2003. Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. *Performance Evaluation* 54(1):1–32.

Cantaluppi, L. 1984. Optimality of piecewise-constant policies in semi-Markov decision chains. *SIAM Journal on Control and Optimization* 22(5):723–739.

Chitgopekar, S. S. 1969. Continuous time Markovian sequential control processes. *SIAM Journal on Control* 7(3):367–389.

Dirac, P. A. M. 1927. The physical interpretation of the quantum dynamics. *Proceedings of the Royal Society of London. Series A* 113(765):621–641.

Glynn, P. W. 1989. A GSMP formalism for discrete event systems. *Proceedings of the IEEE* 77(1):14–23.

Gordon, G. J. 1995. Stable function approximation in dynamic programming. In *Proc. Twelfth International Conference on Machine Learning*, 261–268. Morgan Kaufmann Publishers.

Guestrin, C.; Koller, D.; and Parr, R. 2002. Multiagent planning with factored MDPs. In *Advances in Neural Information Processing Systems 14: Proc. 2001 Conference*. Cambridge, MA: The MIT Press. 1523–1530.

Howard, R. A. 1960. *Dynamic Programming and Markov Processes*. New York, NY: John Wiley & Sons.

Howard, R. A. 1971. *Dynamic Probabilistic Systems*, volume II. New York, NY: John Wiley & Sons.

Matthes, K. 1962. Zur Theorie der Bedienungsprozesse. In *Trans. Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, 513–528. Publishing House of the Czechoslovak Academy of Sciences.

Mausam, and Weld, D. S. 2004. Solving concurrent Markov decision processes. In *Proc. Nineteenth National Conference on Artificial Intelligence*. AAAI Press.

Neuts, M. F. 1975. Probability distributions of phase type. In *Liber Amicorum Professor emeritus dr. H. Florin*. Leuven, Belgium: Katholieke Universiteit Leuven. 173–206.

Neuts, M. F. 1981. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore, MD: Johns Hopkins University Press.

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

Rohanimanesh, K., and Mahadevan, S. 2001. Decision-theoretic planning with concurrent temporally extended actions. In *Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence*, 472–479. Morgan Kaufmann Publishers.

Shedler, G. S. 1993. *Regenerative Stochastic Simulation*. Boston, MA: Academic Press.

Stone, L. D. 1973. Necessary and sufficient conditions for optimal control of semi-Markov jump processes. *SIAM Journal on Control* 11(2):187–201.

Younes, H. L. S., and Simmons, R. G. 2004. Solving generalized semi-markov decision processes using continuous phase-type distributions. In *Proc. Nineteenth National Conference on Artificial Intelligence*. AAAI Press.