TIME-DEPENDENT UTILITY DECISION MAKING USING THE TIMDP MODEL

Guilherme S. Bastos*, Fabio T. Ramos†, Luiz E. de Souza*, Carlos H. C. Ribeiro‡

*Federal University of Itajubá
Information Technology and Systems Engineering Institute
Itajubá, MG, Brazil

†University of Sydney
Australian Centre for Field Robotics
Sydney, NSW, Australia

‡Aeronautical Institute of Technology
Computation Science Division Department
São José dos Campos, SP, Brazil

Emails: sousa@unifei.edu.br, f.ramos@acfr.usyd.edu.au, edival@unifei.edu.br, carlos@ita.br

Abstract— Time-dependent Markov Decision Process (TiMDP) is a recently and important development made for time-dependent sequential decision making. This model consists of stochastic state transitions and stochastic time-dependent action durations. However, time-dependent utility theory can be used in TiMDP to model the rewards over time. This paper presents the modeling of the rewards as time-dependent utility, solves a TiMDP basic example step by step by a proper approximation, and finally shows and solves a more complex example using normal distribution for representation of action duration.

Keywords— Decision Making Theory, Utility Theory, MDP, Planning, Time-dependency.

1 Introduction

Consider the following problem: an accident have occurred, three people are injured, there is only one doctor (agent) that can give medical care for each one at a time, their lives are dependent on medical care, What does the doctor do? In this problem, the injury level is different for each person and there are uncertainties on life maintenance after medical care. Analyzing these parameters, it is almost obvious that the right decision on the attendance sequence could maximize the probability of life saving. Decision theory is often claimed to be able to make the most rational choice (Parsons and Wooldridge, 2002) being the basic theory to solve this practical and common sequential decision problem.

Sequential decision problems have been tackled very intensively in the last few years, and it is well known that the theoretical framework based on Markov Decision Process (MDP) (Puterman, 1990) is the best way to model and solve them, returning optimal results in many cases (Boutilier et al., 1999). However, real problems have an additional and specific parameter, which is time dependency. MDP theory only considers fixed time between epochs that can be easily understood and modeled as iteration steps. To avoid this limitation, Semi-MDPs (SMDPs) (Sutton et al., 1999), and, more recently, Time-Dependent MDPs (TiMDPs) have been proposed by Boyan and Littman (2000). In those models, the transition between states takes a specific time t (durative action). In a TiMDP, time is observable, thus the agent can wait the best moment to make the decision (or execute the action in the present state). Considering the SMDP, there is a time duration probability for the durative action, that is, the agent cannot wait for the best moment to execute the action. The TiMDP model also has likelihood time-dependent functions that activate the action outcome for the present time.

In the accident scenario, the person lifetime, which can be defined as an utility function (Russell and Norvig, 2009), decreases over time and can be formally understood as a time-dependent utility (Horvitz and Rutledge, 1991). This is only one application that can be modeled as a TiMDP problem. Other instances like vehicle routing and scheduling problems with time-window constraints (Solomon, 1987; Ichoua et al., 2003; Ji, 2005; Bastos, 2010) can be also be modeled as TiMDPs. Due to paper size limitation, only sequential decision problems with time-dependent utilities will be considered herein.

The rest of this paper is organized as follows. In Section 2 we discuss the concept of time-dependent utility theory. Section 3 presents the TiMDP general model with a step by step solution of a simple example. Section 4 shows a sequential decision problem with time-dependent utilities and durative actions. The conclusion of this work is presented in Section 5,with a brief discussion of related problems to be tackled and suggestions for future work.

2 Time-dependent utility

An agent needs a measurement value to select the best option (or make a decision) among others. This measurement is the value of the utility function (Li and Soh, 2004). This value is also defined, in decision theoretic planning, as value function (cumulated rewards in sequential decision making) (Boutilier et al., 1999). An expected utility (EU) can be calculated for problems with nondeterministic actions (Russell and Norvig, 2009):

$$\begin{aligned} EU(A|E) &= \sum_i P(Result_i|E, Do(A)) \\ &\cdot U(Result_i(A)) \end{aligned} \quad (1)$$

where $Result_i(A)$ are the possible outcome states for a nondeterministic action A, E summarizes the agent's available evidence about the world, and $Do(A)$ is a proposition informing that action A is executed in the current state.

This common utility representation may not be used in complex real problems, in which actions to be executed are durative and have priorities. Often, it is necessary to solve more urgent tasks and to leave others in wait (Bastos et al., 2008). For solving this question, time-dependent utility theory (Horvitz and Rutledge, 1991) can be used. In this theory, utility is a function of time, greater than zero, which can be increasing or decreasing.

2.1 Decreasing Time-Dependent Utility Function

Decreasing functions can be used to represent a task lifespan and give some idea of priorities to the decision maker. For example, there are two injured people that must receive medical care by the only doctor present in a scenario. They have different injury levels and will die if do not receive medical care as soon as possible. Thus, the doctor needs to take a right decision in the attempt to save both lives, choosing which person to attend first. This decision could be made easily, for this simple example, if the doctor knows the time-dependent utility function representing the importance of a person life (that is, the death risk) in the current time. This function must map important information like age, life decreasing rate, injury level and so forth, to a utility value (this mapping is not the focus of this work, and it is assumed known by the decision maker). Therefore, the right doctor decision is the one that executes the right attendance sequence, considering durative actions, without the utility function reaching the zero value (death).

The decreasing time-dependent utility function can be represented by any decreasing function, but for functionality and simplicity we use exponential or linear functions for its representation:

$$\begin{aligned} U(A,t) &= U(A,t_o) \cdot e^{-k_1 \cdot t} \quad (2) \\ U(A,t) &= U(A,t_o) - k_2 \cdot t, U(A,t) \geq 0 \end{aligned}$$

where U(A,t) is the utility for choosing action A at time t, t_o is the initial time, and k_1 and k_2 are parameters for adjusting the exponential and linear functions, respectively, for the problem requirements.

2.2 Increasing Time-Dependent Utility Function

Increasing functions can be used for instance to represent profits along time. For example, sometimes it is interesting to choose the task execution sequence based on greater rewards, as is the case for the vehicle refueling problem, in which the utility of the refueling state increases over time. Thus, as the fuel level decreases, the utility of refueling increases, and after a certain time and depending on the current position of the vehicle (distance from the fueling station), the refueling decision will be taken. Unlike the decreasing utility function that has a minimum value (zero in the most of the cases), in this case it is reasonable to assume a maximum value. For vehicle refueling in particular, it is important to agree upon a maximum utility value that will refer to an empty tank. The utility model is defined:

$$\begin{aligned} U(A,t) &= U_{max} \cdot (1 - e^{-k_3 \cdot t}) \quad (3) \\ U(A,t) &= U(A,t_o) + k_4 \cdot t, U(A,t) \leq 0 \end{aligned}$$

where $U(A,t)$ is the utility for choosing action A at time t, t_o is the initial time, U_{max} is the maximum utility, and k_3 and k_4 are parameter constants for adjusting the exponential and linear functions, respectively, for the problem requirements.

3 Time-dependent MDP Model

Time-dependent MDPs (TiMDPs) were first proposed by Boyan and Littman (2000) to model and solve sequential decision problems with the following attributes: (1) Stochastic state transitions; and (2) Stochastic time-dependent action durations.

It is not cited by Boyan and Littman (2000), but rewards can be also modeled by a time-dependent utility. Therefore, we add the follow attribute for the TiMDP model: (3) Time-dependent rewards.

Formally, a TiMDP consists of the following components:

S	Discrete space state
A	Discrete action space
M	Discrete set of outcomes, each of the form $\mu = \langle s'_\mu, T_\mu, P_\mu \rangle$: $s'_\mu \in S$: the resulting space $T_\mu \in \{\text{ABS,REL}\}$: specifies the type of the resulting time distribution (absolute or relative) $P_\mu(t')$(if $T_\mu =$ ABS): pdf (probability density function) over absolute arrival times of μ $P_\mu(\delta)$(if $T_\mu =$ REL): pdf over durations of μ
L	$L(\mu\|s,t,a)$ is the likelihood of outcome μ given state s, time t AND action a
R	$R(\mu,t,\delta)$ is the reward for the outcome μ at time t with duration δ

The TiMDP model is represented by the following Bellman equations:

$$\begin{aligned} V(s,t) &= \max_{a \in A} Q(s,t,a) \\ Q(s,t,a) &= \sum_{\mu \in M} L(\mu|s,a,t).U(\mu,t) \\ U(\mu,t) &= \int_{-\infty}^{\infty} P_\mu(t')[R(\mu,t,t'-t) + \\ & \quad V(s'_\mu,t')]dt'(\text{ if } T_\mu = \text{ABS}) \\ U(\mu,t) &= \int_{-\infty}^{\infty} P_\mu(t'-t)[R(\mu,t,t'-t) + \\ & \quad V(s'_\mu,t')]dt'(\text{ if } T_\mu = \text{REL}) \end{aligned} \tag{4}$$

where $U(\mu,t)$ is the utility of outcome μ in time t, $V(s,t)$ is the time-value function for the immediate action, and $Q(s,t,a)$ is the expected Q time-value over outcomes.

We can note that the calculations of $U(\mu,t)$ are convolutions of the result-time pdf P_μ with the lookahead value $R+V$. The likelihood function L represents the probability of an outcome occurrence for action a in time t, and can be also used to model problems with time-windows (Bresina et al., 2002).

This model is used to solve time-dependent problems with finite time horizon and represents an undiscounted continuous-time MDP.

3.1 *Discrete Solution for Relative Time Distributions by Backwards Convolution*

In the general TiMDP model (Boyan and Littman, 2000), the time-value functions for each state can be arbitrarily complex and therefore impossible to represent exactly. The TiMDP problem is solved by representing R and V as a piecewise linear (PWL) function, L as a piecewise constant (PWC) function, and P_μ is discretized. This representation ensures closure under the convolutions and avoids an increased number of iterations. This solution is fast and exact (for the approximated functions), but there are the following drawbacks: (1) loss of information caused by the initial approximations; (2) insertion of new breakpoints in the piecewise functions over iterations; and (3) need for an analytic solution of the convolution integral.

Li and Littman (2005) explored the practical solution of value iteration considering that P_μ is a PWC function. This way, the degree of convoluted functions would grow up during the iterations, making impossible its solution in a reasonable time. To prevent this behavior, Li and Littman (2005) introduced the Lazy Approximation Algorithm, in which the resultant PWL convoluted function is approximated to a PWC function on each iteration. Hence, the imprecisions and state space augmentation introduced by discretization of P_μ is avoided in this solution method.

In a recent work performed by Rachelson et al. (2009), the related functions of the TiMDP model are represented by piecewise polynomial (PWP) functions. In order to limit the degree growing of the iteration results, the introduced algorithm executes when needed, a decreasing step, reducing the degree of the results in the current iteration by PWP interpolation.

In order to simplify the solution algorithm, we propose the discretization of all involved functions in the model and solution of the convolutions by a discrete numerical method. This approximation does not provide a solution as fast as the original one, but it is an easier and direct way to solve problems with few states. The only problem here is that the convolution present in the TiMDP model is not solved as conventional convolution integral. A conventional convolution integral can be represented by:

$$h(t) = \int_{-\infty}^{\infty} g(t')k(t-t')dt' \tag{5}$$

The discrete formulation of a convolution is:

$$h(j) = k(j) * g(j) = \sum_i g(i)k(j-i) \tag{6}$$

This convolution involves a delay represented by the k function over the g function. However, in the TiMDP there is a negative delay, and the convolution integral is now:

$$h(t) = \int_{-\infty}^{\infty} g(t')k(t'-t)dt' \tag{7}$$

We characterize it as a *backwards convolution*, and its discrete solution is:

$$h(j) = k(j) \bullet g(j) = \sum_i g(i)k(j+i) \tag{8}$$

So, using our solution method, the time-value function V for relative P_μ is:

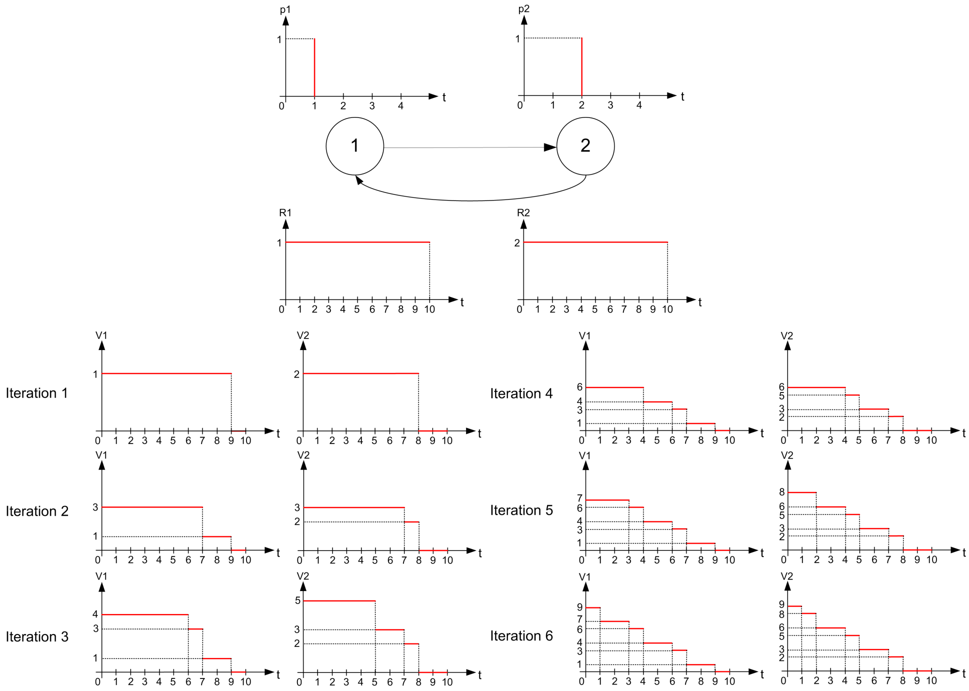

Figure 1: TiMDP Example solved step by step using value iteration algorithm

$$V(s,t) \;=\; \max_{a\in A} \sum_{\mu\in M} L(\mu|s,a,t)\cdot P_\mu(t) \bullet \quad (9)$$
$$[R(\mu,t) + V(s'_\mu,t)]$$

For discretized problems with absolute time distributions, the integral of Eq. 4 can be solved by numerical methods such as the *Newton-Cotes Rule* (Thisted, 1988).

3.2 A General TiMDP example

The example presented in Fig. 1 is a good start point for understanding the value iteration in TiMDPs. The problem is composed by two states, one action per state, constant rewards (R) over time t, and unitary probability function (L) over all time horizon. In this case, if the agent is at *State 1*, it will receive a reward equals one after one time period (the action is durative and takes deterministically one time period) going to *State 2*. In *State 2*, the agent will receive a reward equals two after two time period. The rewards can be cumulated until the end of time horizon.

The system starts with time-value functions V equals to zero. Then, the problem is solved by value iteration using Bellman equations (Eq. 4) with our approximations presented in Section 3.1. The value iteration converges at the sixth iteration, and the solution of V gives important information for the agent decision making. For example, the agent is at *State 2* at time 2, it knows that can receive an accumulated reward of 6 units following the given policies. In this case, the agent can wait until time 3 and receive the same cumulated reward. Thus, for TiMDPs, policies are dependent of current time and state.

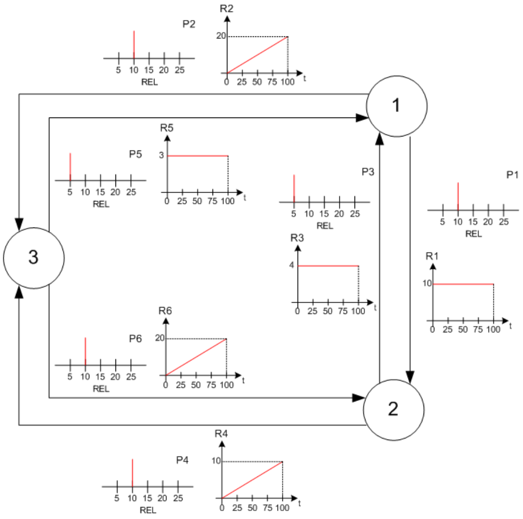

Figure 2: Sequential decision making problem using time-dependent utility

4 A time-dependent utility application

In this section, we present sequential a decision making example using time-dependent utilities (or rewards varying over time) modeled and solved by a TiMDP. The example is presented in Fig. 2; it has three states, two selectable actions per state, a finite horizon with limit of *100 time periods*, an unitary likelihood function over all time horizon, and deterministic action durations.

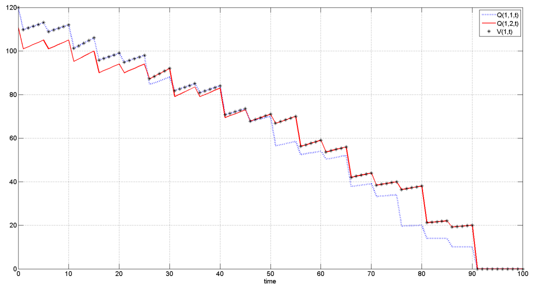

Figure 3: Value function - $V(1,t)$

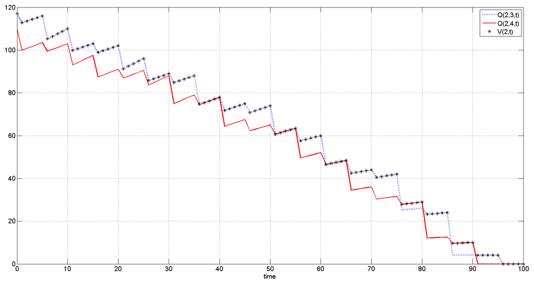

Figure 4: Value function - $V(2,t)$

The problem was solved by value iteration using Bellman equations with our approximations (Eq. 9). The results for the time-value functions V and Q are presented in Figs. 3, 4 and 5.

In the graphics, we have the time-value function $V(State, t)$, which is the maximum between the $Q(State, Action, t)$ time functions.

Figure 6 shows the policies depending on the time. Such policies define the actions that the agent must choose based on the maximum Q value for a state, in time t. For example, if the agent is at *State 3* and the current time is *77*, it must choose *Action 6*, therefore moving to *State 2*.

4.1 Stochastic Action Duration

In the TiMDP model, actions can be durative and uncertain (represented by pdfs). We used a *Normal Distribution* to represent *P1* in our example, with mean 10 and variance 3. Normal distribution are very convenient for this kind of problem, in which action is durative with different durations over executions. For real situations with a reliable database of past action durations, a *Normal Distribution* is a good approximation for the action duration pdf, because it tends to cluster around a single mean value with the proper variance. The solution for *State 1* is shown in Fig. 7.

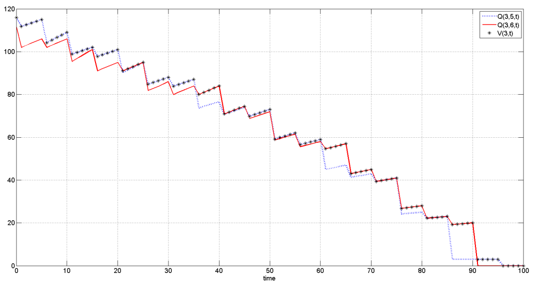

Figure 5: Value function - $V(3,t)$

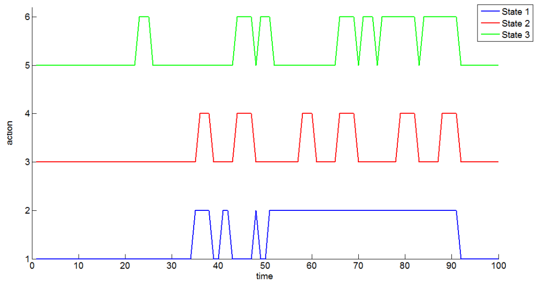

Figure 6: Policies over time

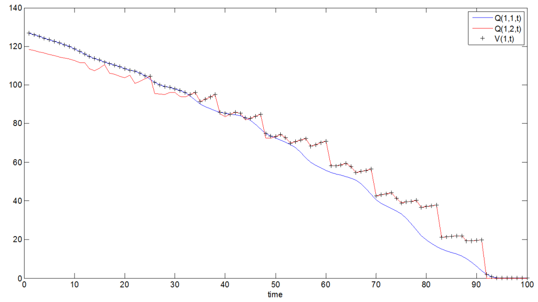

Figure 7: Value function $V(1,t)$, $P1 = N(10,3)$

Comparing this result with the original problem (Fig. 2), it is clear that the function $Q(1,1,t)$ becomes smoother. There is also a change in aspect for function $Q(1,2,t)$. In fact, these changes in the function may change the overall policies due to the uncertainty in action durations that is related to inherent variances.

The policies are shown in the Fig. 8. Comparing to Fig. 6 we note a difference between the policies, which is caused by the uncertainty added in the duration of *Action 1*. For example, now the policy in *State 1* at time 25 is *Action 2*, against *Action 1* in the original problem. The uncertainty added to the action duration that belongs to *State 1* has also caused a changing in the policy for *State 3*. Therefore, it is very important to model correctly the pdfs in order to avoid wrong decisions.

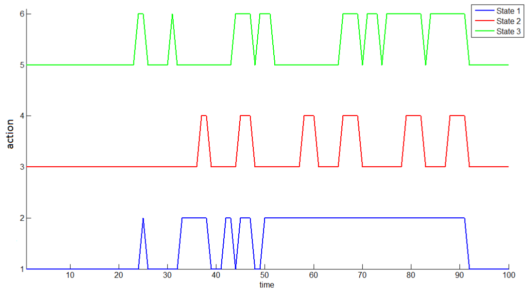

Figure 8: Policies over time - $P1 = N(10,3)$

5 Conclusion and future work

This paper has presented an application of TiMDP using time-dependent utility. This concept can be easily applied to real time-dependent sequential decision making problems with optimal results. The use of pdfs – specifically normal distributions in our case – to represent actions durations often make representative difference in policy results over time, therefore a precise database for pdfs representation has a crucial importance in the quality of results. More work in this sense is important to evaluate the impact of pdfs parameters over the policies.

This work is just an initial part of a bigger project, where we are trying to model and make good approximations of real problems to TiMDP representation. Another kind of problems like vehicle dispatching, traveler salesman problem and vehicle routing problem, with time-dependent utilities and time-windows representations is been developed for single and multi-agent systems. For a near future, we will apply these concepts to real problems, like truck dispatching in open-pit mining, and use Reinforcement Learning to adapt the systems parameters, searching for better policies.

Acknowledgments

Guilherme S. Bastos is grateful to CAPES and FAPEMIG for the financial support. Carlos H. C. Ribeiro is grateful to CNPq and FAPESP for the partial financial support.

References

Bastos, G. (2010). *Methods for Truck Dispatching in Open-pit Mining*, PhD thesis, Aeronautical Institute of Technology, Brazil.

Bastos, G., Ribeiro, C. and Souza, L. (2008). Variable utility in multi-robot task allocation systems, *Robotic Symposium, IEEE Latin American* pp. 179–183.

Boutilier, C., Dean, T. and Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage, *Journal of Artificial Intelligence Research* **11**(1): 94.

Boyan, J. and Littman, M. (2000). Exact solutions to time-dependent MDPs, *Advances in Neural Information Processing System* **2**(12): 10.

Bresina, J., Dearden, R., Meuleau, N., Smith, D. and Washington, R. (2002). Planning under continuous time and resource uncertainty: A challenge for AI, *AIPS Workshop on Planning for Temporal Domains*, Citeseer, pp. 91–97.

Horvitz, E. and Rutledge, G. (1991). Time-dependent utility and action under uncertainty, *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence, Los Angeles, CA*, Citeseer, pp. 151–158.

Ichoua, S., Gendreau, M. and Potvin, J. (2003). Vehicle dispatching with time-dependent travel times, *European journal of operational research* **144**(2): 379–396.

Ji, X. (2005). Models and algorithm for stochastic shortest path problem, *Applied Mathematics and Computation* **170**(1): 503–514.

Li, L. and Littman, M. (2005). Lazy approximation for solving continuous finite-horizon MDPs, *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, Vol. 20, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 1175.

Li, X. and Soh, L. (2004). Applications of decision and utility theory in multi-agent systems, *CSE Technical reports* p. 85.

Parsons, S. and Wooldridge, M. (2002). *An Introduction to Game Theory and Decision Theory*, Kluwer Academic Publishers.

Puterman, M. (1990). Markov decision processes, *Handbooks in Operations Research and Management Science* **2**: 331–434.

Rachelson, E., Fabiani, P. and Garcia, F. (2009). TiMDPpoly: An Improved Method for Solving Time-Dependent MDPs, *Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on*, IEEE, pp. 796–799.

Russell, S. and Norvig, P. (2009). *Artificial intelligence: a modern approach*, Prentice hall.

Solomon, M. (1987). Algorithms for the vehicle routing and scheduling problems with time window constraints, *Operations research* **35**(2): 254–265.

Sutton, R., Precup, D. and Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning, *Artificial intelligence* **112**(1): 181–211.

Thisted, R. (1988). *Elements of statistical computing: numerical computation*, Chapman & Hall/CRC.