Exact Solution to Time-Dependent MDPs

by Justin Boyan and Michael Littman

Originally presented at NIPS 2000
November 2000, Denver CO

Machine Learning Lunch - NIPS Review Session

Joelle Pineau *February 5 2000*

The Problem

Imagine trying to plan a route from Squirrel Hill to CMU that minimizes expected time.

Assume stochastic state transitions

$\Rightarrow$ **use Markov Decision Process (MDP)**

… now add stochastic action durations

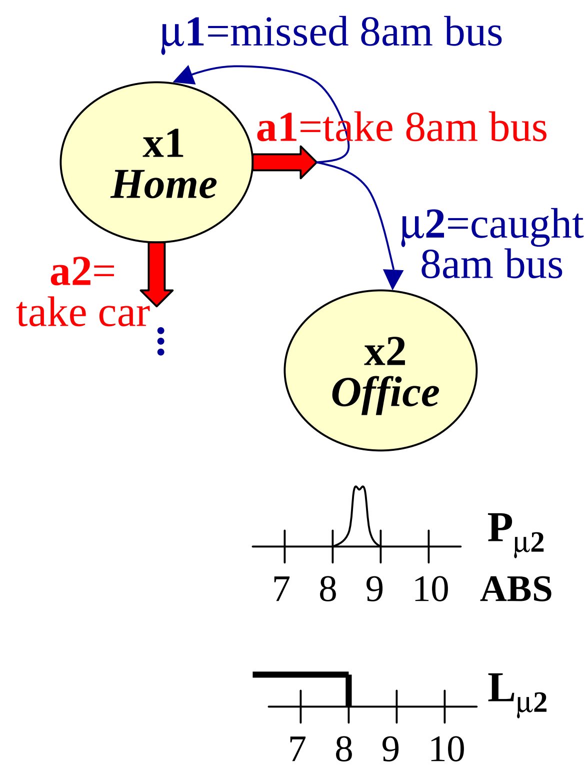
$\Rightarrow$ **use Semi-Markov Decision Process (SMDP)**

… now add time-dependent action durations

$\Rightarrow$ **use Time-Dependent MDP (TMDP)**

Defining the TMDP Model

- A TMDP is an MDP composed of states with both a discrete component and a real-valued time component: $\langle x,t \rangle \in X \times \Re$
- Formally:
  - X discrete state space
  - A discrete action space
  - M discrete set of outcomes $\mu = \langle x'_\mu, T_\mu, P_\mu \rangle$
    - $x'_\mu \in X$
    - $P_\mu(t')$ (if T_μ=ABS): pdf over arrival times
    - $P_\mu(\delta)$ (if T_μ=REL): pdf over duration
    - $T_\mu \in \langle \text{ABS,REL} \rangle$
  - L $L(\mu|x,t,a)$ the likelihood of outcome μ
  - R $R(\mu,t,\delta)$ the reward for outcome μ
  - K $K(x,t)$ the reward rate for "dawdling"



Solving a TMDP

- Find a policy via the value function V*, using Bellman equations:

$$V(x,t) = \sup_{t' \geq t} \left(\int_t^{t'} K(x,s)\,ds + \overline{V}(x,t') \right)$$ value fn (allowing dawdling)

$$\overline{V}(x,t) = \max_{a \in A} Q(x,t,a)$$ value fn (immediate action)

$$Q(x,t,a) = \sum_{\mu \in M} L(\mu | x,a,t) \cdot U(\mu,t)$$ expected Q value over outcomes

$$U(\mu,t) = \begin{cases} \int_{-\infty}^{\infty} P_\mu(t')\left[R(\mu,t,t'-t) + V(x'_\mu, t')\right] dt' & \text{(if } T_\mu\text{=ABS)} \\ \int_{-\infty}^{\infty} P_\mu(t'-t)\left[R(\mu,t,t'-t) + V(x'_\mu, t')\right] dt' & \text{(if } T_\mu\text{=REL)} \end{cases}$$

- Simply an undiscounted continuous-time MDP.

Modeling constraints for exact solution

- Assuming that:
  - The rewards, **R**, are *piecewise linear functions* of start time, arrival time, and action duration: $R(\mu,t,\delta)=R_s(\mu,t)+R_a(\mu,t+\delta)+R_d(\mu,\delta)$
  - The dawdling rewards, **K**, and the outcome probability functions, **L**, are *piecewise constant*.
  - Arrival and duration pdfs, $\mathbf{P_\mu(t')}$ and $\mathbf{P_\mu(\delta)}$, are *discrete*.
- Then:
  - All operations from the Bellman equations can be implemented exactly.
    - I.e. add, multiply, integrate, sup, max, convolution
  - Run-time of each operator is linear in # of linear components.

Experiments: Telescope Scheduling

- Consider:
  - 1 telescope with a 13-hour observation period
  - N targets with varying visibility time windows
  - State variables are target id and time.
  - Rewards based on duration of viewing a target
  - Transitions based on stochastic time to move telescope and calibrate on new target

RESULTS:	N	#states	V* pieces	Runtime(secs)
TMDP	10	11	186	0.1
MDP	10	14311	14311	1.3
TMDP	25	26	716	1.8
MDP	25	33826	33826	7.4
TMDP	50	51	1252	6.3
MDP	50	66351	66351	34.5
TMDP	100	101	2711	17.9
MDP	100	131300	131300	154.1

Conclusions

- Paper presents a stochastic model for time-dependent MDPs.
  - Demonstrated that dynamic programming with piecewise linear time-value functions can produce optimal policies efficiently.
- Two applications (city-sized route planning and telescope scheduling) were studied in simulation.
- TMDP can model many domains, using the real-valued dimension to represent a continuous resource (e.g. battery life, money, etc.)
- Future work will investigate ways of using the algorithm to generate upper and lower bounds on the optimal solution for arbitrary pdfs.