Hard Constrained Semi-Markov Decision Processes

Wai-Leong Yeow[*] and **Chen-Khong Tham**[†]
National University of Singapore
21 Lower Kent Ridge Road, Singapore 119077
{waileong.yeow,eletck}@nus.edu.sg

Wai-Choong Wong
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
lwong@i2r.a-star.edu.sg

Abstract

In multiple criteria Markov Decision Processes (MDP) where multiple costs are incurred at every decision point, current methods solve them by minimising the expected primary cost criterion while constraining the expectations of other cost criteria to some critical values. However, systems are often faced with hard constraints where the cost criteria should never exceed some critical values at any time, rather than constraints based on the expected cost criteria. For example, a resource-limited sensor network no longer functions once its energy is depleted. Based on the semi-MDP (sMDP) model, we study the hard constrained (HC) problem in continuous time, state and action spaces with respect to both finite and infinite horizons, and various cost criteria. We show that the HCsMDP problem is NP-hard and that there exists an equivalent discrete-time MDP to every HCsMDP. Hence, classical methods such as reinforcement learning can solve HCsMDPs.

Introduction

Markov Decision Processes (MDP) (Puterman 1994) is a popular model of sequential decision problems which has found applications in a variety of areas: target tracking (Evans, Krishnamurthy, & Nair 2005), sensor networks (Yeow, Tham, & Wong 2005), multi-agent systems (Ghavamzadeh & Mahadevan 2004), resource management in grid computing, telecommunication networks (Yu, Wong, & Leung 2004), etc. Quite often, these applications are associated with multiple criteria (costs constraints), in which constrained MDP (CMDP) comes into play by bounding the various expected cumulative costs. However, in systems with critical resources, it may be fatal when the total resources (costs) exceed some critical point at any time. For example, energy is a vital resource in sensor networks. Once depleted, the whole network becomes dysfunctional. Overloading a computing node in a grid with jobs can cause it to fail, resulting in permanent loss of resources. Henceforth, we are motivated to introduce hard constraints into MDP. A hard constraint is a restricting condition on some cumulative cost incurred at any time. This constraint, if violated, causes the control agent to cease functioning *immediately*.

MDP with constraints have come a long way since the late 80's when Beulter et al (1986) considered average cost constraints in the semi-MDP (sMDP) domain. Feinberg, Shwartz (1996) and Altman (1999) showed many interesting results when dealing with constraints on expected total costs, including the existence of randomised optimal policies. CMDPs are also formulated as Abstract Dynamic Programming (ADP) (Gabor, Kalmar, & Szepesvari 1998). Horiguchi (2001) considered MDPs with a constraint on the termination time. However, to our best knowledge, MDPs with hard constraints have yet been studied. Attempts were made to solve MDP with soft constraints (Dolgov & Durfee 2003) but they focused on discrete-time MDP problems with positive costs and presented approximate solutions.

We study the hard constrained (HC) problem in the sMDP domain for better modelling power where the sojourn time between states is continuous and random. Hence, the states and actions can be continuous as well. Classical discrete-time MDP is simply a special case where states and actions are discrete, and sojourn time is deterministically unity. We also considered total, discounted and average cost criteria in HCsMDP. Two special properties of HCsMDP are shown in this paper: (a) HCsMDP is NP-Hard and (b) a HCsMDP problem is equivalent to some discrete-time MDP. The latter property ensures that there exists a wealth of solutions for HCsMDP such as dynamic programming (Bertsekas 2000), Q-learning (Watkins & Dayan 1992) and linear programming (Altman 1999).

Background

Consider a scenario where an agent controls a Markovian system, which is described by states. At each decision point, the agent takes an action in the current system state and the system transits to a new state. Consequently, the agent incurs $K+1$ costs. The primary cost is denoted as c_0 and other auxiliary costs form a K-dimensional cost vector $\mathbf{c}$.

Definition 1. *A discrete-time MDP is a tuple $\langle X, A, p, c\rangle$ where X and A are the set of states and actions. $p_{xa}(x')$ denotes the probability density function (pdf) of the transition to state x' from state x under action a. Each transition takes unit time and results in costs $c_0(x', x, a), \mathbf{c}(x', x, a) \in \mathbb{R}$.*[1]

[*]NUS Graduate School for Integrative Sciences & Engineering.
[†]Department of Electrical & Computer Engineering.

[1]An alternate notation $c(x, a)$, independent of x', exists in literature but it is actually $c(x, a) = \mathbb{E}_{x'}\{c(x', x, a)\}$.

The k^{th} total cost criterion for a finite horizon of N stages is

$$J^{(\pi)}_{k,N}(x_0) = \mathbb{E}_X \left[\sum_{i=0}^{N-1} c_k(x_{i+1}, x_i, a_i) | x_0 \right], \qquad (1)$$

where x_0 is the initial state, $\pi : X \rightarrow A$ is the control policy that maps the current state to an action. Vector $\mathbf{J}$ further denotes the total cost criteria for costs $k = 1 \ldots K$.

Definition 2. *A semi-MDP is a tuple $\langle X, A, f, c \rangle$ where X and A are the sets of states and actions. $f_{xa}(x', \tau)$ denotes the pdf of the transition to state x' from state x under action a in time τ. Costs associated with each transition are denoted by $c_0(x', \tau, x, a), \mathbf{c}(x', \tau, x, a) \in \mathbb{R}$. The total cost criterion is of the same form as* (1)*, with cost c_k defined as $c_k(x_{i+1}, \tau_i, x_i, a_i)$ instead.*

In classical CMDP, J_0 is minimised while keeping $\mathbf{J} \leq \mathbf{C}$. For convenience, we define the following notations:

- $\mathbb{Z}_1^K = \{1, 2, \ldots, K\}$.
- $\mathbf{J} \leq \mathbf{C} \Leftrightarrow J_k \leq C_k, \forall k \in \mathbb{Z}_1^K$.
- $\nu \subseteq \mathbb{Z}_1^K$ represents the set of constraints which are violated. Conversely, $\bar{\nu} = \mathbb{Z}_1^K - \nu$ denotes otherwise.
- $\mathbf{m}_\nu$ denotes a sub-vector of $\mathbf{m}$: $(m_k), \forall k \in \nu \subseteq \mathbb{Z}_1^K$.
- $1\{ex\}$ is a boolean indicator function that evaluates to 1 if ex is true, 0 otherwise.

Hard Constrained Semi-Markov Decision Processes (HCsMDP)

Finite Horizon HCsMDP

Finite horizon problems are usually defined with respect to the number of stages N. However, in sMDPs, it is more appropriate to define the finite horizon as total time T instead. In fact, deadline T can be seen as a hard constraint by itself where the elapsed time can never exceed T. Beyond this, the system terminates with zero costs. Furthermore, multiple costs are associated with each state transition. Each cost may represent some critical resource where the cumulative value can never exceed some critical value *at all times*. This gives rise to two variants of the same problem:

V1. The system terminates immediately with zero costs when *any* constraint on cumulative costs are violated.

V2. The system terminates *only* after deadline T.

Definition 3. *A hard constrained semi-MDP (HCsMDP) is a tuple $\langle X, A, f, T, c, \mathbf{C} \rangle$ where X, A, f, c has the same semantics as in Definition 2, $\mathbf{C}$ is a vector of critical values and T is the deadline. For the total cost criteria, the finite horizon problem is then*

$$\begin{array}{rrl} & \min_\pi J^{(\pi)}_{0,N} & s.t. \\ \textbf{V1}: & \sup_N t_{N-1} < T, & \sup_N \sum_{i=0}^{N-1} \mathbf{c}_i < \mathbf{C} \\ \textbf{V2}: & t_{N-1} = T, & \sup_N \sum_{i=0}^{N-1} \mathbf{c}_i < \mathbf{C} \end{array} \qquad (2)$$

and the infinite horizon problem is

$$\min_\pi J^{(\pi)}_{0,\infty} \qquad s.t. \quad \sup_{N \rightarrow \infty} \sum_{i=0}^{N-1} \mathbf{c}_i < \mathbf{C}. \qquad (3)$$

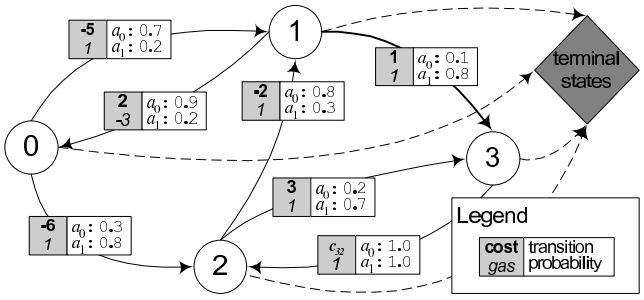

Figure 1: A semi-MDP with 4 locations. The taxi driver chooses between 2 queues at each location: a_0 and a_1. The profits (or negative costs) earned on each trip are in bold and the gas used is in italics. He faces two hard constraints: time and gas. When total elapsed time is more than T or when the taxi is out of gas, the system transits to the absorbing terminal states. The sojourn time between all locations is 1 hour except between locations 1 and 3, which is uniformly distributed between 1 to 5 hours. The cost c_{32} is varied to show the correctness of our approach.

Although the hard constraints defined above are based on cumulative (total) costs, they can be applied to discounted and average costs as well. For clarity, we first focus on the total cost criteria and discuss others later. We also illustrate an example with discrete states and actions although our theory applies to the continuous domain as well.

Consider a taxi driver's decision problem as illustrated in Fig. 1. The four states represent locations where the driver can pick up and drop off passengers (location 0 is where he starts off). At each location, he chooses between two queues to drop off his current passenger and pick up another. In one trip, the driver may get to refill his tank. This may result in more nett loss than profit but may allow the taxi to operate longer. He wishes to maximise his income (represented as negative costs), but faces two constraints: time (say, he only works for 9 hours) and gas (a critical resource). Then, **V2** enforces that the driver have to operate for the full 9 hours and to ensure that the taxi does not run out of gas during operation. Conversely, **V1** is more relaxed: the driver can stop driving before the time is up (especially when excessive losses are incurred). Although both formulations differ slightly, there is a common property.

Proposition 1. *The hard constrained semi-MDP problem for both formulations is NP-Hard.*

This result is not surprising because the general MDP problem is already NP-hard (Goldsmith, Littman, & Mundhenk 1997). Note that proofs of all propositions in this paper are stated in the Appendix.

Any violation terminates the system (V1)

To exactly solve a HCsMDP, we introduce a new set of absorbing terminal states (the rhombus in Fig. 1) where the system transits to when it terminates. Each state has a non-zero probability of being absorbed into the rhombus. This should increase with elapsed time and cumulative costs.

We construct an equivalent discrete-time MDP $\langle \tilde{X}, A, \tilde{p}, \tilde{c} \rangle$. The augmented state space $\tilde{X}$, which tracks cumulative costs $\mathbf{d} = \sum \mathbf{c}$ and elapsed time t, is

$$\tilde{X} = \{\tilde{x} = (x, t, \mathbf{d}) | \forall x \in X, t \leq T, \mathbf{d} \leq \mathbf{C}\}. \quad (4)$$

Augmented states $\tilde{x}$ with $t = T$ or $\mathbf{d} = \mathbf{C}$ represent the absorbing terminal states. We argue that the expansion of state space to track elapsed time and cumulative costs is inevitable. This is because terminating conditions depend on t and $\mathbf{d}$, and thus affects the transition pdf $\tilde{p}$. Moreover, it is also known that optimal policies for finite horizon problems are non-stationary (Puterman 1994), and both t and $\mathbf{d}$ are *sufficient* to represent all previous transitions starting from $i = 0$. There are three cases to consider for each transition from $\tilde{x}$ to $\tilde{x}'$:

C1. No violation of any constraints. Thus the system does not terminate, i.e. $t' < T$ and $\mathbf{d}' < \mathbf{C}$.

C2. Deadline T is reached and all transitions beyond T are collated into some terminal states $\{(x', T, \mathbf{d}' | \mathbf{d}' < \mathbf{C})\}$.

C3. Some constraints $\nu \subseteq \mathbb{Z}_1^K$ are violated and all transitions are collated into some terminal states represented as $\{(x', t', \mathbf{d}') | t' < T, \mathbf{d}'_\nu = \mathbf{C}_\nu\}$. We denote by $\bar{\nu}$ the set of constraints that are not violated. It is easy to see that **C1** is a special case of this, where $\bar{\nu} = \mathbb{Z}_1^K$ and $\nu = \emptyset$.

The case where both the deadline and constraints are violated does not exist since the system terminates immediately after T and *no* costs are incurred. Then, $\tilde{p}$ is related to f by

$$\tilde{p}_{\tilde{x}a}(\tilde{x}') = \tilde{p}_{xt\mathbf{d}a}(x', t', \mathbf{d}') = \begin{cases} \mathcal{U}_\nu(x', x, t' - t, \mathbf{d}' - \mathbf{d}, a) \cdot \mathcal{I}_{\bar{\nu}}(x', x, t' - t, \mathbf{d}' - \mathbf{d}, a) \cdot f_{xa}(x', t' - t) & , \mathbf{d}_\nu < \mathbf{d}'_\nu = \mathbf{C}_\nu, \mathbf{d}'_{\bar{\nu}} < \mathbf{C}_{\bar{\nu}}, t < t' < T \\ \int_{t'-t}^{\infty} f_{xa}(x', \tau) d\tau & , \mathbf{d} = \mathbf{d}' < \mathbf{C}, t < t' = T \\ 1 & , \mathbf{d} = \mathbf{d}' < \mathbf{C}, t = t' = T, x = x' \\ 1 & , \mathbf{d}_\nu = \mathbf{d}'_\nu = \mathbf{C}_\nu, \mathbf{d}_{\bar{\nu}} = \mathbf{d}'_{\bar{\nu}} < \mathbf{C}_{\bar{\nu}}, t = t' < T, x = x' \\ 0 & , \text{otherwise} \end{cases} \quad (5)$$

where $\mathcal{I}$ and $\mathcal{U}$ are indicator functions such that

$$\mathcal{I}_{\bar{\nu}}(x', x, \tau, \mathbf{b}, a) = \prod_{k \in \bar{\nu}} 1\{c_k(x', \tau, x, a) = b_k\} \text{ and}$$
$$\mathcal{U}_\nu(x', x, \tau, \mathbf{b}, a) = \prod_{k \in \nu} 1\{c_k(x', \tau, x, a) \geq b_k\}.$$

These are needed because for every transition described by (x', τ, x, a), only one unique cost $\mathbf{c}$ is incurred. It cannot be simply retrieved from $\mathbf{d}' - \mathbf{d}$ due to case **C3** above. This results in a unique pair $\{\mathbf{d}, \mathbf{d}'\}$. In fact, the first condition of (5) correspond to both **C1** and **C3** combined whereas the second condition correspond to **C2**. The third and fourth conditions correspond to the set of *absorbing* termination states.

Finally, the redefined cost $\tilde{c}$ is such that

$$\tilde{c}(\tilde{x}', \tilde{x}, a) = \begin{cases} c_0(x', t' - t, x, a) & , t < t' < T, \mathbf{d}' < \mathbf{C} \\ 0 & , \text{otherwise.} \end{cases} \quad (6)$$

Proposition 2. *$\tilde{p}$ is a proper probability density function.*

Since transitions to $\tilde{x}'$ only depends on the previous state $\tilde{x}$ and action a in (5), and $\tilde{p}$ is a proper pdf, $\langle \tilde{X}, A, \tilde{p}, \tilde{c} \rangle$ is a valid MDP. We now state:

Proposition 3. *For any non-stationary policy π in a HC-sMDP $\mathcal{M} = \langle X, A, f, T, \mathbf{c}, \mathbf{C} \rangle$, there exists a stationary policy $\tilde{\pi}$ with the same mapping of actions in a discrete-time MDP $\tilde{\mathcal{M}} = \langle \tilde{X}, A, \tilde{p}, \tilde{c} \rangle$ such that $J_{0,N}^{(\pi)} = \tilde{J}_{0,\infty}^{(\tilde{\pi})}$, where $\tilde{X}$, $\tilde{p}$ and $\tilde{c}$ are defined in* (4)*,* (5) *and* (6) *respectively.*

The optimal policy, thus, can be found through classical discrete-time MDP solvers on the equivalent problem $\tilde{\mathcal{M}}$.

Only deadline violations terminate the system (V2)

Since all cumulative costs $\mathbf{d}$ cannot exceed the critical values $\mathbf{C}$ for all $t < T$, the only solution is to keep track of $\mathbf{d}$ and prune the action space of each state such that there is no chance of violating any of the K constraints, i.e., eliminate undesirable actions $a \in \bar{A}_{\tilde{x}}$ that satisfy the following:

$$a: \quad \exists k \in \mathbb{Z}_1^K, \quad d'_k = C_k \quad \text{and} \quad \tilde{p}_{\tilde{x}a}(\tilde{x}') > 0. \quad (7)$$

Similar to the case of **V1**, classical MDP solvers can solve **V2** through the equivalent problem $\langle \tilde{X}, \tilde{A}, \tilde{p}, \tilde{c} \rangle$, where $\tilde{A}_{\tilde{x}} = A - \bar{A}_{\tilde{x}}$. However, it may be possible that $\tilde{A}_{\tilde{x}} = \emptyset$. In that case, the hard constraints are deemed as unsatisfiable.

Note that the method of pruning the action space should not be applied to **V1** as that would only result in *sub-optimal* policies. In both problems, there may exist states where the agent has a choice of either terminating with high probability or risking more costs in continuing to run the system. Then the optimal policy in these situations is obviously to cut costs and strive for early termination. Such states are very likely to exist in **V1** since there are multiple terminating conditions. In the case of the taxi driver's problem, he may want to stop driving if continuing to drive will increase losses. Hence, plainly applying action pruning will only force the agent to adopt a sub-optimal behaviour.

Discounted and Average Cost Criteria

The discounted and average cost criteria are usually discussed in infinite horizon problems since the total cost may be divergent as $N \to \infty$. We discuss both criteria in the finite horizon problem for completeness, which are trivial extensions to the total cost criterion. The same discrete-time MDP can be constructed where the cumulative costs $\mathbf{d}_n$ of the n^{th} stage is redefined as $\sum_{i=0}^{n} e^{-\gamma t_i} \mathbf{c}_i$ for γ-discounted, or $\frac{1}{t_{n-1}} \sum_{i=0}^{n} \mathbf{c}_i$ for the average cost criteria. Similarly, it is also straightforward to define $\tilde{p}$ by changing the input $\mathbf{d}' - \mathbf{d}$ to functions $\mathcal{I}$ and $\mathcal{U}$ appropriately for the new cost criteria.

Infinite Horizon: Pruning the Action Space

It can be easily seen that the infinite horizon problem is similar to that of **V2** where only the deadline T terminates the system. In fact, these two problems are equivalent if $T \to \infty$. Thus, the only feasible solution for hard constrained sMDPs in the infinite horizon case would be pruning the action space as in (7). Refer to Fig. 2 for an illustration. It is always *probable*, no matter how small the possibility is, to incur a cost of 2 at state 0 under action a_1. For

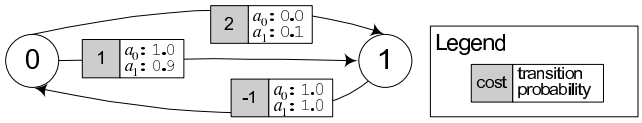

Figure 2: A discrete-time MDP that requires action pruning in order to satisfy some hard constraints.

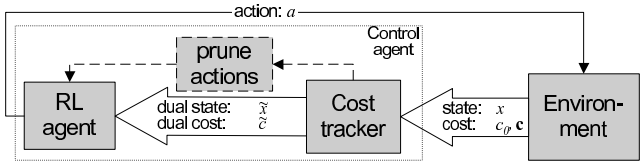

Figure 3: Structure of an RL agent. The module for pruning actions is only required for **V2** or infinite horizon problems.

a discount rate of 0.5, there is always a possibility of getting a total discounted cost[2] of 2 if action a_1 at state 0 is not pruned away. Hence, when all states communicate in a infinite horizon, the possibility of a constraint violation may not diminish to zero if some actions are not pruned (e.g., hard constraint $\mathbf{C} = 2$ in the discounted case).

However, (7) is not implementable in the infinite horizon because both t and $\mathbf{d}$ are unbounded. To overcome this, observe that in the case of discounted cost criteria,

$$\mathbf{d}_\infty = \mathbf{d}_M + \underbrace{e^{-\gamma t_M} \sum_{i=0}^{\infty} e^{\gamma(t_M - t_{M-i})} \mathbf{c}_{M+i}}_{\varepsilon} \ . \quad (8)$$

For large values of M, the rightmost term is almost negligible. Thus, an approximate solution for the infinite horizon case would be to amend the hard constraints to $\mathbf{C} - \varepsilon$, where ε is some small positive number, and prune undesirable actions for the first M steps in the equivalent discrete-time MDP. Note that this method can also be used for the average cost criterion since $\mathbf{d}$ stabilise for large values of M as well. Conversely, an average cost case can also be approximated to a γ-discounted one by applying a Tauberian approximation (Gabor, Kalmar, & Szepesvari 1998) with γ as a value sufficiently close to 1 and $\mathbf{C}_{approx} = \frac{\mathbf{C}_{actual}}{1-\gamma}$.

Solving HCsMDP

Since there exists an equivalent discrete-time MDP for a HCsMDP, there exists a wealth of solutions for MDP such as dynamic programming (Bertsekas 2000), linear programming/occupation measures (Altman 1999) and reinforcement learning methods such as Q-learning (Watkins & Dayan 1992). Unlike classical CMDP where constraints are based on expected cost criteria and optimal policies are likely to be randomised[3] (Altman 1999), optimal policies for HCsMDP problems are deterministic. After all, this is a result from classical discrete-time MDPs.

In the experiments section, we use Dynamic Programming (DP) to obtain the optimal policies for the taxi driver's problem. The basic idea of DP is to start from the terminal states, select the best action of the previous time step by considering every transition into these states, and iterating until the initial states. However, this is based on the assumption that the sojourn time between states is a discrete random variable and that the cumulative costs are discrete as well. For the case of continuous values, one possible approximated solution would be to discretise the augmented state space with the *ceiling* function, in order to capture the hard constraints. The time complexity of using DP is then $O((|X|TR^K)^2)$, where T is the number of time steps, and R is the maximum range of a cumulative cost.

The other feasible approach is to use Reinforcement Learning (RL) (Sutton & Barto 1998), with a module to keep track of augmented states (cumulative costs) as shown in Fig. 3. Since the augmented states form a discrete-time MDP, it is assured that the RL agent will converge to the optimal policy. The other advantage of RL is that computation of (5) can be avoided since it is model-free. For **V2** or infinite horizon problems, another module could be added to blacklist (and prune away) undesirable actions that violate the hard constraints.

Experiments

We solve the taxi driver's problem in this section. Transition costs and probabilities are represented as solid lines in Fig. 1. For simplicity, all transitions between locations take 1 hour (except from location 1 to 3 where the sojourn time is uniformly distributed between 1 to 5 hours). The time constraint is set at 9 hours and the gas constraint is set at 4 litres. We vary the nett profit $-c_{32}$ earned from location 3 to 2 to show differences in the optimal policies obtained from the equivalent discrete-time MDP.

Starting from location 0, the taxi driver could choose queue a_1 to increase the chances of getting a better profit. Thereafter, at location 2, the driver might want to choose a_0 to get more profit. However, depending on the profit $-c_{32}$ from location 3 to 2, a_0 might not be a good choice. If the profit is lucrative, the better policy would probably to choose route $2 \rightarrow 3 \rightarrow 2$ as opposed to routes $2 \rightarrow 1 \rightarrow 3$ or $2 \rightarrow 1 \rightarrow 0$ because the former consumes more time at the $1 \rightarrow 3$ path and the latter gives less profit. In fact, this policy has a similar structure to the optimal one that, *ignoring constraints*. Depending on the gas costs and the time constraints, the optimal policy will differ slightly.

We first examine a special (and simplified) case of the taxi driver's problem where there is only a time constraint and no gas constraint. This is a special case of HCsMDP, called deadline-sensitive semi-MDP. Subsequently, we include the gas constraints and show differences in the optimal policies between the deadline sMDP and the HCsMDP case.

Special case: deadline-sensitive semi-MDP

In the absence of other constraints, only the elapsed time is tracked at the augmented state $\tilde{x}$. Hence, the first condi-

[2]Perpetually incurring cost of 2 from state 0 to state 1 and cost of -1 on the return. The corresponding average cost is 0.5.

[3]A randomised policy refers to executing actions at some state that conforms to some probability distribution.

tion in (5) collapses to $f_{xa}(x', t'-t)$ (without the indictator functions) and the fourth condition no longer applies while the rest remains. Clearly, the resultant $\tilde{p}$ will still be a proper probability density function and Prop. 3 still applies.

We use Dynamic Programming (Bertsekas 2000) to solve the equivalent discrete-time MDP of the deadline-sensitive sMDP problem. In the case of $c_{32} = -12$, the optimal action at location 1 changes from a_1 (with $\tilde{J}^{(\pi^*)}_{0,\infty}(1,2) = -18.35$) to a_0 (with $\tilde{J}^{(\pi^*)}_{0,\infty}(1,3) = -13.96$) at elapsed time 2 and 3 hours respectively. Initially, for elapsed time 2 hours and below, the driver is more likely to travel the route $1 \to 3 \to 2$ than the route $1 \to 0 \to 2$ by taking action a_1 at location 1. This ensures more profit. However, for elapsed time $t > 2$, the best action is a_0, which avoids route $1 \to 3 \to 2$. This is because too much time will be spent in $1 \to 3$ whereas the other route gives better profit in the short term. Finally, as the deadline approaches ($t > 6$), the best action is again a_1 at location 1 for better profit. Conversely, if $c_{32} = -7$ or more, the optimal action for both locations 1 and 2 is a_0, avoiding location 3 altogether.

Taxi driver's problem

Suppose now that the total amount of gas consumed at any time cannot be more than 4 litres and every trip consumes 1 litre of gas. The only gas station available is between location 0 and 1. We focus on **V1** since it is more interesting than simply applying action pruning in the case of **V2**.

Applying Prop. 3 and Linear Programming (Altman 1999), we obtained an optimal policy for the case of $c_{32} = -12$. In the first few hours, the taxi driver attempts to fill up his tank through route $0 \to 1 \to 0$, which translates to taking action a_0 at both locations 0 and 1. Similar behaviour can be observed at location 2 in the initial hours where the optimal action is a_0 (route $2 \to 1 \to 0$). Subsequently, when the taxi is filled with gas, the policies change at both locations 1 and 2 to action a_1 instead, in order to reap more profit at location 3. However, as gas consumption nears the limit, similar behaviour as that of the deadline sMDP can be observed (when deadline draws near). The best action for location 1 becomes a_0 again in order to refill the gas tank.

Concluding remarks

We introduced HCsMDP — semi-Markov Decision Processes with hard constraints, where the cumulative costs cannot exceed some critical values at all times. In the case of problems with a finite horizon of deadline T, violating these constraints may terminate the system. We considered both cases where the former is true, and vice versa.

In the first case, we proved that the hard constrained problem is equivalent to some discrete-time MDP problem where the augmented states track the elapsed time and cumulative costs. We argue that the expansion of state space is inevitable because transition probabilities to the terminating states vary with time and cumulative costs. These new additions to the state space can also be seen as sufficient statistics that summarise the whole transition history. Optimal policies also change with elapsed time and cumulative costs as shown in the taxi driver's problem.

In the second case where cumulative costs are to be kept within the hard constraints for the whole duration T, pruning bad actions is the only way to satisfy such hard constraints. In fact, this is the same case for infinite horizon cases.

The equivalence of a hard constrained semi-Markov Decision process in the continuous time domain to a discrete-time MDP problem also indicates that a wealth of solutions exist for solving such problems. However, the equivalent problem will have continuous states when the sojourn time is a continuous variable. This is true even if the original problem has only discrete states. In that case, existing techniques can only give approximate solutions. In this paper, we use Dynamic Programming and Linear Programming to solve the equivalent discrete-time MDP for the taxi driver's problem. In particular, we showed the correctness of our approach through a study of the solutions obtained by varying some of the costs.

However, since MDP suffers from the curse of dimensionality, the expanded state space may further increase the required time complexity. In fact, HCsMDP is already NP-Hard. Methods like reinforcement learning (Sutton & Barto 1998) could be used to mediate this. The expanded MDP is largely *sparse*: the elapsed time t in the augmented state space always increases (there will not be transitions that go back in time) and the cumulative costs $\mathbf{d}$ are not massively interconnected in the transition model (since only one cost vector $\mathbf{c}$ is associated with each transition). Methods like Dynamic Bayesian Networks (Guestrin 2003), or sampling based approaches (Likhachev, Gordon, & Thrun 2004) that tackle large but sparse MDPs could be employed.

In this paper, the solutions mentioned using an equivalent discrete-time MDP strive for an exact solution for a NP-Hard problem. Thus, the future direction would be working towards heuristics or approximations.

Appendix

Proof of Prop.1. We reduce the (0,1) multi-criteria knapsack problem to a HCsMDP. Suppose there are M items and each item $a \in \mathbb{Z}_1^M$ has a primary cost v_a and auxiliary costs $[\ t_a\ \ \mathbf{w}_a\ ]'$. $\sum v_a$ is to be minimised with constraints $\sum \mathbf{w}_a < \mathbf{W}$ and $\sum t_a < T$. Then, let the state $\mathbf{x}$ be a vector of M bits. On taking action (item) a in state $\mathbf{x}$, the system deterministically transits to $\mathbf{x}'$ such that the a^{th} bit of $\mathbf{x}' = 1$, incurring costs v_a, t_a and $\mathbf{w}_a$ only if $\mathbf{x}' \neq \mathbf{x}$. Then, solving such a HCsMDP problem by minimising $\sum v_a$ while hard constrained to time T and cost $\mathbf{W}$ is equivalent to solving a (0,1) mutli-criteria knapsack problem. □

Proof of Prop. 2. We split the proof into three cases: (1) $t = T$, (2) $\mathbf{d}_\nu = \mathbf{C}_\nu$, for any $\nu \neq \emptyset$, and (3) otherwise. Also, we denote by $\mathcal{D}$ the set of all possible $\mathbf{d}$.

Case (1) Assume $\int_X \int_0^\infty \int_{\mathcal{D}} \tilde{p}_{xT\mathbf{d}a}(x', t', \mathbf{d}') d\mathbf{d}' dt' dx' \neq 1$. Then, there must be some possible transitions to $(x', T, \mathbf{d})$ other than $x = x'$ since no cost is incurred at the point $t = T$, $\mathbf{d} = \mathbf{d}' < \mathbf{C}$. But there are no such transitions besides the third condition of (5), giving a contradiction.

Case (2) Similar to Case (1), the system halts and no cost is incurred, giving $\mathbf{d}' = \mathbf{C}$ and $t = t'$. Hence, there is

no transition to other states except in the fourth condition of (5), giving $\int_X \int_0^\infty \int_{\mathcal{D}} \tilde{p}_{xt\mathbf{d}a}(x', t', \mathbf{d}') d\mathbf{d}' dt' dx' = 1$, for $\mathbf{d}_\nu = \mathbf{C}_\nu, \nu \neq \emptyset$.

Case (3) This case is equivalent to $t < t' \leq T$ and $\mathbf{d} < \mathbf{C}$. We further split this into another 2 distinct events:

$e_{\mathcal{A}}$ **:** $t' = T$ and $\mathbf{d}' < \mathbf{C}$, and

$e_{\mathcal{B}}$ **:** $t' < T$ and $\mathbf{d}'_{\bar{\nu}} < \mathbf{C}_{\bar{\nu}}$ for any $\bar{\nu}$.

Event $e_{\mathcal{A}}$ corresponds to the 2nd condition in (5), since the system halts with no cost incurred, giving $\mathbf{d} = \mathbf{d}'$. Hence,

$$\begin{aligned} \tilde{P}_{xt\mathbf{d}a}(\cdot, e_{\mathcal{A}}) &= \int_X \int_{T-t}^\infty f_{xa}(x', \tau) d\tau dx' \\ &= P_{xa}(\cdot, \tau \geq T - t). \end{aligned}$$

Event $e_{\mathcal{B}}$ corresponds to the 1st condition in (5).

$$\begin{aligned} &\tilde{P}_{xt\mathbf{d}a}(\cdot, e_{\mathcal{B}}) \\ &= \int_X \int_0^{T-t} \int_{\mathcal{D}} \begin{matrix} \mathcal{U}_\nu(x', x, \tau, \mathbf{d}' - \mathbf{d}, a) \cdot \\ \mathcal{I}_{\bar{\nu}}(x', x, \tau, \mathbf{d}' - \mathbf{d}, a) \cdot \\ f_{xa}(x', \tau) \end{matrix} d\mathbf{d}' d\tau dx' \end{aligned}$$

Since cost $\mathbf{c}$ is a function of transition (x', τ, x, a), there will only be one unique value of $\mathbf{d}'$ where $\mathcal{U}$ and $\mathcal{I}$ are non-zero. Then,

$$\begin{aligned} \tilde{P}_{xt\mathbf{d}a}(\cdot, e_{\mathcal{B}}) &= \int_X \int_0^{T-t} f_{xa}(x', \tau) d\tau dx' \\ &= P_{xa}(\cdot, \tau < T - t) \\ \text{and} \quad \tilde{P}_{\tilde{x}a}(\cdot, \cdot) &= \tilde{P}_{xt\mathbf{d}a}(\cdot, e_{\mathcal{A}} \cup e_{\mathcal{B}}) = 1. \end{aligned}$$

□

Proof of Prop. 3. For the first M decision points, the primary cost criterion of the equivalent MDP

$$\begin{aligned} \tilde{J}^{(\tilde{\pi})}_{0,M}(\tilde{x}_0) &= \mathbb{E}_{\tilde{X}} \left[\sum_{i=0}^{M-1} \tilde{c}_0(\tilde{x}_{i+1}, \tilde{x}_i, a_i) | \tilde{x}_0 \right] \\ &= \sum_{i=0}^{M-1} \int_X \int_0^\infty \int_{\mathcal{D}} \begin{matrix} \tilde{c}_0(x, \tau, \mathbf{d}, x_i, t_i, \mathbf{d}_i, a_i) \cdot \\ \tilde{p}_{x_i t_i \mathbf{d}_i a_i}(x, \tau, \mathbf{d}) \end{matrix} d\mathbf{d} d\tau dx, \end{aligned}$$

where $\mathcal{D}$ is the set of all possible $\mathbf{d}$. Since $\tilde{c}(x, \tau, \mathbf{d}, x_i, t_i, \mathbf{d}_i, a_i)$ is only non-zero when $t_i < \tau < T$, we only need to evaluate $\tilde{p}$ using the first condition in (5) with $\nu = \emptyset$ and $\bar{\nu} = \mathbb{Z}_1^K$. This gives the cost criterion to be

$$\sum_{i=0}^{M-1} \int_X \int_{t_i}^T \int_{\mathcal{D}} \begin{matrix} \mathcal{I}_{\mathbb{Z}_1^K}(x, x_i, \tau - t_i, \mathbf{d} - \mathbf{d}_i, a_i) \cdot \\ f_{x_i a_i}(x, \tau - t_i) c_0(x, \tau - t_i, x_i, a_i) \end{matrix} d\mathbf{d} d\tau dx.$$

Moreover, for any transition (x, τ, x_i, a_i), there is only one such $\mathbf{d}$ where $\mathcal{I}_{\mathbb{Z}_1^K}$ is non-zero as $\mathbf{c}$ is a function. Then,

$$\tilde{J}^{(\tilde{\pi})}_{0,M}(\tilde{x}_0) = \sum_{i=0}^{M-1} \int_X \int_0^{T-t_i} f_{x_i a_i}(x, \tau') c_0(x, \tau', x_i, a_i) d\tau' dx.$$

Since the range of τ' is kept within 0 and $T - t_i$, t_{i+1} never exceeds T for all $i < M$. The same applies for $\mathbf{d}_i < \mathbf{C}$. Thus, $\sup_M t_{M-1} < T$, $\sup_M \mathbf{d}_{M-1} < \mathbf{C}$ and

$$\begin{aligned} \tilde{J}^{(\tilde{\pi})}_{0,\infty}(\tilde{x}_0) &= \lim_{M \to \infty} \mathbb{E}_{X,\tau} \left[\sum_{i=0}^{M-1} c_0(x_{i+1}, \tau_i, x_i, a_i) | x_0 \right] \\ &= J^{(\pi)}_{0,N}(x_0) \end{aligned}$$

□

References

Altman, E. 1999. *Constrained Markov Decision Processes*. Chapman & Hall/CRC.

Bertsekas, D. P. 2000. *Dynamic Programming and Optimal Control: 2nd Edition*. Athena Scientific.

Beutler, F. J., and Ross, K. W. 1986. Time-Average Optimal Constrained Semi-Markov Decision Processes. *Advances in Applied Probability* 18(2):341–359.

Dolgov, D., and Durfee, E. 2003. Approximating Optimal Policies for Agents with Limited Execution Resources. In *Proc. of IJCAI'03*.

Evans, R.; Krishnamurthy, V.; and Nair, G. 2005. Networked Sensor Management and Data Rate Control for Tracking Maneuvering Targets. *IEEE Trans. Signal Processing* 53(6):1979–1991.

Feinberg, E. A., and Shwartz, A. 1996. Constrained Discounted Dynamic Programming. *Mathematics of Operations Research* 21:922–945.

Gabor, Z.; Kalmar, Z.; and Szepesvari, C. 1998. Multi-criteria Reinforcement Learning. In *Proc. of ICML'98*.

Ghavamzadeh, M., and Mahadevan, S. 2004. Learning to Communicate and Act Using Hierarchical Reinforcement Learning. In *Proc. of AAMAS'04*.

Goldsmith, J.; Littman, M. L.; and Mundhenk, M. 1997. The complexity of plan existence and evaluation in probablistic domains. In *Proc. of UAI'97*.

Guestrin, C. 2003. *Planning Under Uncertainty in Complex Structured Environments*. Ph.D. Dissertation, Stanford University, Stanford, California.

Horiguchi, M. 2001. Markov decision processes with a stopping time constraint. *Mathematical Methods of Operations Research* 53(2):279–295.

Likhachev, M.; Gordon, G.; and Thrun, S. 2004. Planning for Markov Decision Processes with Sparse Stochasticity. In *Proc. of NIPS'2004*.

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

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Watkins, C. J., and Dayan, P. 1992. Q-learning. *Machine Learning* 8(3):279–292.

Yeow, W.-L.; Tham, C.-K.; and Wong, W.-C. 2005. Energy Efficient Multiple Target Tracking in Sensor Networks. In *Proc. of IEEE GLOBECOM'05*.

Yu, F.; Wong, V. W. S.; and Leung, V. C. M. 2004. A New QoS Provisioning Method for Adaptive Multimedia in Cellular Wireless Networks. In *Proc. of INFOCOM'04*.