Extending the Bellman equation to continuous actions and continuous time in the discounted case

Emmanuel Rachelson [1]

Frédérick Garcia [2]
Patrick Fabiani [1]

[1]ONERA-DCSD — Toulouse, France

[2]INRA-BIA —Toulouse, France

ISAIM, January 2-4th, 2007

ONERA THE FRENCH AEROSPACE LAB INRA

What happens when you try to consider random continuous observable decision epochs in an MDP framework ?

Emmanuel Rachelson [1]

Frédérick Garcia [2]
Patrick Fabiani [1]

[1]ONERA-DCSD — Toulouse, France

[2]INRA-BIA —Toulouse, France

ISAIM, January 2-4th, 2007

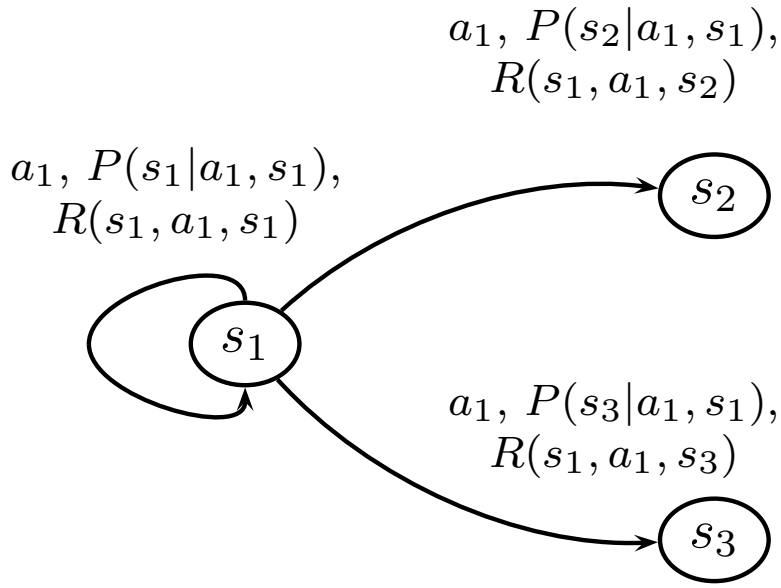
ONERA THE FRENCH AEROSPACE LAB INRA

Plan

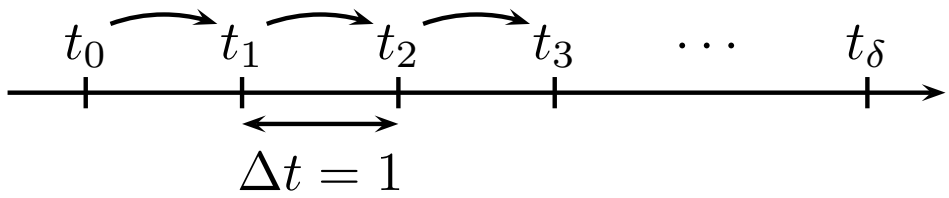
Markov Decision Processes

MDP: 5-tuple $< S, A, P, r, T >$

Policy: $\pi : \begin{cases} S & \rightarrow & A \\ s & \mapsto & a \end{cases}$



Discounted criterion: $V_\gamma^\pi(s) = E(\sum_{\delta=0}^{\infty} \gamma^\delta r(s_\delta, \pi(s_\delta)) | s_0 = s)$



ONERA THE FRENCH AEROSPACE LAB | INRA

Optimality

Optimal policy:

$$\pi^* = \arg\max_{\pi \in \mathscr{D}} V_\gamma^\pi$$

$$\pi^* = \arg\max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s')$$

Bellman equation:

$$V^*(s) = LV^*(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right\}$$

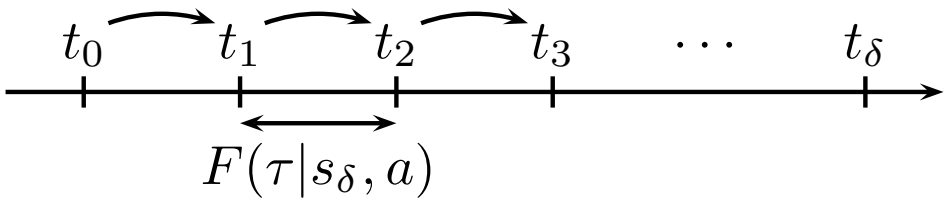
SMDP and CTMDP

random decision epochs

sojourn time: $F(\tau|s,a)$
$\longrightarrow$ independent of s'
$\longrightarrow$ independent of δ

Similar to the MDP case

CTMDP: exponential distribution on τ



Time-dependency

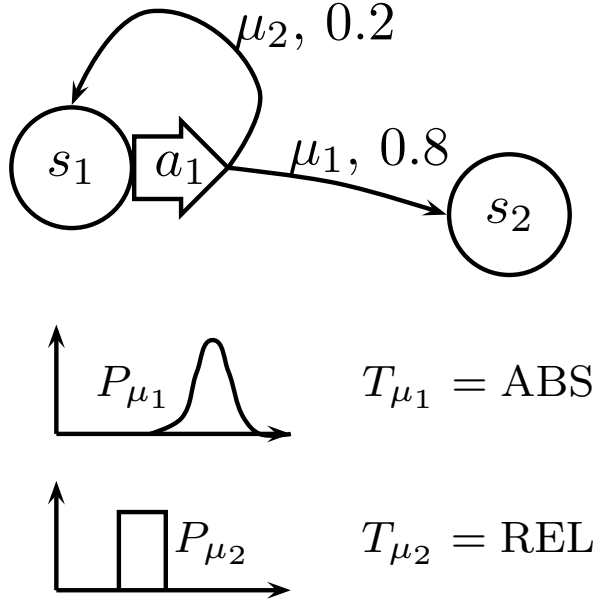
Continuous τ but ...

non-observable (no time-dependent dynamics)
independence between τ and s'

TMDP

TMDP (Boyan and Littman, 01):

- discrete state space S
- discrete action space A
- Outcome space M. Outcome μ:
  - an outcome state s'_μ
  - a flag T_μ
  - a pdf on time P_μ
- likelihood function $L(\mu|s,t,a)$
- reward function $R(\mu,t,t')$
- dawdling cost function $K(s,t)$



Policies

TMDP policy: $\pi(s,t) = (t',a)$

Optimality equations:

$$V(s,t) = \sup_{t' \geq t} \left(\int_t^{t'} K(s,\theta) d\theta + \overline{V}(s,t') \right) \quad (1)$$

$$\overline{V}(s,t) = \max_{a \in A} Q(s,t,a) \quad (2)$$

$$Q(s,t,a) = \sum_{\mu \in M} L(\mu|s,t,a) \cdot U(\mu,t) \quad (3)$$

$$U(\mu,t) = \begin{cases} \int_{-\infty}^{\infty} P_\mu(t')[R(\mu,t,t') + V(s'_\mu,t')]dt' \\ \int_{-\infty}^{\infty} P_\mu(t'-t)[R(\mu,t,t') + V(s'_\mu,t')]dt' \end{cases} \quad (4)$$

Questions

- Is t a state variable ?
- Continuous state variables = bounded ?
- Acyclic dynamics ?
- Bounded vs infinite horizon ?
- Causality principle ?

- Is there a *wait* action?
- Is there a more generic way of representing "*wait*" actions ?
- Continuous actions ? Parametric actions ?

Questions

- Is t a state variable ?
- Continuous state variables = bounded ?
- Acyclic dynamics ?
- Bounded vs infinite horizon ?
- Causality principle ?

- Is there a *wait* action?
- Is there a more generic way of representing "*wait*" actions ?
- Continuous actions ? Parametric actions ?

Questions

- Is t a state variable ?
- Continuous state variables = bounded ?
- Acyclic dynamics ?
- Bounded vs infinite horizon ?
- Causality principle ?

- Is there a *wait* action?
- Is there a more generic way of representing "*wait*" actions ?
- Continuous actions ? Parametric actions ?

Questions

- Is t a state variable ?
- Continuous state variables = bounded ?
- Acyclic dynamics ?
- Bounded vs infinite horizon ?
- Causality principle ?

- Is there a *wait* action?
- Is there a more generic way of representing "*wait*" actions ?
- Continuous actions ? Parametric actions ?

Questions

- Is t a state variable ?
- Continuous state variables = bounded ?
- Acyclic dynamics ?
- Bounded vs infinite horizon ?
- Causality principle ?

- Is there a *wait* action?
- Is there a more generic way of representing "*wait*" actions ?
- Continuous actions ? Parametric actions ?

Questions

- Is t a state variable ?
- Continuous state variables = bounded ?
- Acyclic dynamics ?
- Bounded vs infinite horizon ?
- Causality principle ?

- Is there a *wait* action?
- Is there a more generic way of representing "*wait*" actions ?
- Continuous actions ? Parametric actions ?

Questions

- Is t a state variable ?
- Continuous state variables = bounded ?
- Acyclic dynamics ?
- Bounded vs infinite horizon ?
- Causality principle ?

- Is there a *wait* action?
- Is there a more generic way of representing "*wait*" actions ?
- Continuous actions ? Parametric actions ?

Questions

- Is t a state variable ?
- Continuous state variables = bounded ?
- Acyclic dynamics ?
- Bounded vs infinite horizon ?
- Causality principle ?

- Is there a *wait* action?
- Is there a more generic way of representing "*wait*" actions ?
- Continuous actions ? Parametric actions ?

Framework

We need a formal framework,
close enough to MDP,
in order to take into account continuous time and the associated parametric actions.
We wish to establish criteria
and to derive optimality equations from it.

Framework

We need a formal framework,
close enough to MDP,
in order to take into account continuous time and the associated parametric actions.
We wish to establish criteria
and to derive optimality equations from it.

Framework

We need a formal framework,
close enough to MDP,
in order to take into account continuous time and the associated parametric actions.
We wish to establish criteria
and to derive optimality equations from it.

Framework

We need a formal framework,
close enough to MDP,
in order to take into account continuous time and the associated parametric actions.
We wish to establish criteria
and to derive optimality equations from it.

Planning horizon vs. temporal horizon

planning horizon: number of steps allowed before termination

temporal horizon: upper bound on t

Planning horizon vs. temporal horizon

planning horizon: number of steps allowed before termination

temporal horizon: upper bound on t

Our approach: both infinite horizons with discounted criterion

Planning horizon vs. temporal horizon

planning horizon: number of steps allowed before termination
temporal horizon: upper bound on t

Our approach: both infinite horizons with discounted criterion

Other cases:

- finite planning horizon: same as standard MDP
- finite temporal horizon: bounded observable resource
- total reward criterion: reduction of the discounted case

Model

XMDP:

- S hybrid state space
  $\longrightarrow$ emphasis on t: (s,t)
- $A(X)$ parametric action space
  $\longrightarrow$ $a_i(x)$: action a_i with parameter x.
- p transition pdf
  $\longrightarrow$ $p(s'|s,a(x)) \equiv p(s',t'|s,t,a(x))$
- r reward function
  $\longrightarrow$ $r(s,t,a(x))$
- T decision epoch indexes

Hypothesis

- Action durations have a strictly positive lower bound

$$\rightarrow t_{\delta+1} - t_{\delta} \geq \alpha > 0$$

- Upper semi-continuity of r

$$\rightarrow \limsup_{x \rightarrow x_0} r(s,t,a(x)) \leq r(s,t,a(x_0))$$

Hypothesis

- Action durations have a strictly positive lower bound

  $$\longrightarrow t_{\delta+1} - t_{\delta} \geq \alpha > 0$$

- Upper semi-continuity of r

  $$\longrightarrow \limsup_{x \longrightarrow x_0} r(s,t,a(x)) \leq r(s,t,a(x_0))$$

Bellman equation for MDP

1. $V = LV \Rightarrow V = V^*$
   - 1.1 $V \geq LV \Rightarrow V \geq V^*$
     $\forall \pi \in \mathscr{D}$, $V - V^\pi \geq (L^\pi)^{(n)} V - V^\pi \geq s_n$
     with $\lim_{n \to \infty} s_n = 0$
   - 1.2 $V \leq LV \Rightarrow V \leq V^*$
2. Existence of such a V

   $a_1 = \arg\max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) U(s')$

   $\forall U, V \; LU(s) - LV(s) \leq \gamma \sum_{s' \in S} P(s'|s,a_1)[U(s') - v(s')]$

   therefore $||LU - LV|| \leq \gamma ||U - V||$, L is a contraction mapping.

Bellman equation for MDP

1. $V = LV \Rightarrow V = V^*$
   - 1.1 $V \geq LV \Rightarrow V \geq V^*$

     $\forall \pi \in \mathscr{D},\ V - V^\pi \geq \left(L^\pi\right)^{(n)} V - V^\pi \geq s_n$

     with $\lim_{n \to \infty} s_n = 0$
   - 1.2 $V \leq LV \Rightarrow V \leq V^*$
2. Existence of such a V

   $a_1 = \arg\max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) U(s')$

   $\forall U, V\ LU(s) - LV(s) \leq \gamma \sum_{s' \in S} P(s'|s,a_1)[U(s') - v(s')]$

   therefore $||LU - LV|| \leq \gamma ||U - V||$, L is a contraction mapping.

Bellman equation for MDP

1. $V = LV \Rightarrow V = V^*$
   - 1.1 $V \geq LV \Rightarrow V \geq V^*$

     $\forall \pi \in \mathscr{D},\ V - V^\pi \geq \left(L^\pi\right)^{(n)} V - V^\pi \geq s_n$

     with $\lim_{n \to \infty} s_n = 0$
   - 1.2 $V \leq LV \Rightarrow V \leq V^*$
2. Existence of such a V

   $a_1 = \arg\max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) U(s')$

   $\forall U, V\ LU(s) - LV(s) \leq \gamma \sum_{s' \in S} P(s'|s,a_1)[U(s') - v(s')]$

   therefore $||LU - LV|| \leq \gamma ||U - V||$, L is a contraction mapping.

Bellman equation for MDP

1. $\underline{V = LV \Rightarrow V = V^*}$
   - 1.1 $\underline{V \geq LV \Rightarrow V \geq V^*}$
     $\forall \pi \in \mathscr{D},\ V - V^\pi \geq (L^\pi)^{(n)} V - V^\pi \geq s_n$
     with $\lim_{n \to \infty} s_n = 0$
   - 1.2 $\underline{V \leq LV \Rightarrow V \leq V^*}$
2. Existence of such a V

$$a_1 = \arg\max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) U(s')$$

$$\forall U, V\ LU(s) - LV(s) \leq \gamma \sum_{s' \in S} P(s'|s,a_1)[U(s') - v(s')]$$

therefore $||LU - LV|| \leq \gamma ||U - V||$, L is a contraction mapping.

L^π and L

$$L^\pi V(s,t) = r(s,t,\pi(s,t)) + \int_{\substack{t' \in \mathbb{R} \\ s' \in S}} \gamma^{t'-t} p(s',t'|s,t,\pi(s,t)) V(s',t') ds' dt'$$

$$LV(s,t) = \sup_{\pi \in \mathscr{D}} \left\{ r_\pi(s,t) + \int_{\substack{t' \in \mathbb{R} \\ s' \in S}} \gamma^{t'-t} p_\pi(s',t'|s,t) V(s',t') ds' dt' \right\}$$

Bellman equation for XMDP

1. $V = LV \Rightarrow V = V^*$
   - 1.1 $V \geq LV \Rightarrow V \geq V^*$
     $\forall \pi \in \mathscr{D}$, $V - V^\pi \geq (L^\pi)^{(n)} V - V^\pi \geq s_n$
     with $\lim_{n \to \infty} s_n = 0$.
     key features: r bounded, $t_{\delta+1} - t_\delta \geq \alpha$.
   - 1.2 $V \leq LV \Rightarrow V \leq V^*$
2. Existence of such a V
   key features: upper semi-continuity of r, $t_{\delta+1} - t_\delta \geq \alpha$.
   $||LU - LV|| \leq \gamma^\alpha ||U - V||$, L is a contraction mapping.

Bellman equation for XMDP

1. $V = LV \Rightarrow V = V^*$

   1.1 $V \geq LV \Rightarrow V \geq V^*$

   $\forall \pi \in \mathscr{D}$, $V - V^\pi \geq (L^\pi)^{(n)} V - V^\pi \geq s_n$

   with $\lim_{n \to \infty} s_n = 0$.

   key features: r bounded, $t_{\delta+1} - t_\delta \geq \alpha$.

   1.2 $V \leq LV \Rightarrow V \leq V^*$

2. Existence of such a V

   key features: upper semi-continuity of r, $t_{\delta+1} - t_\delta \geq \alpha$.

   $||LU - LV|| \leq \gamma^\alpha ||U - V||$, L is a contraction mapping.

Bellman equation for XMDP

1. $V = LV \Rightarrow V = V^*$

   1.1 $V \geq LV \Rightarrow V \geq V^*$

   $\forall \pi \in \mathscr{D}$, $V - V^\pi \geq (L^\pi)^{(n)} V - V^\pi \geq s_n$

   with $\lim_{n \to \infty} s_n = 0$.

   key features: r bounded, $t_{\delta+1} - t_\delta \geq \alpha$.

   1.2 $V \leq LV \Rightarrow V \leq V^*$

2. Existence of such a V

   key features: upper semi-continuity of r, $t_{\delta+1} - t_\delta \geq \alpha$.

   $||LU - LV|| \leq \gamma^\alpha ||U - V||$, L is a contraction mapping.

Bellman equation for XMDP

1. $V = LV \Rightarrow V = V^*$
   - 1.1 $V \geq LV \Rightarrow V \geq V^*$
     $\forall \pi \in \mathscr{D}$, $V - V^\pi \geq (L^\pi)^{(n)} V - V^\pi \geq s_n$
     with $\lim_{n \to \infty} s_n = 0$.
     key features: r bounded, $t_{\delta+1} - t_\delta \geq \alpha$.
   - 1.2 $V \leq LV \Rightarrow V \leq V^*$
2. Existence of such a V

   key features: upper semi-continuity of r, $t_{\delta+1} - t_\delta \geq \alpha$.
   $||LU - LV|| \leq \gamma^\alpha ||U - V||$, L is a contraction mapping.

Reformulation

XMDP: generalization of MDP

Existence of a similar optimality equation: $V^* = LV^*$

$$LV(s,t) = \max_{a \in A} \sup_{x \in X} \left\{ r(s,t,a(x)) + \int_{\substack{t' \in \mathbb{R} \\ s' \in S}} \gamma^{t'-t} p(s',t'|s,t,a(x)) V(s',t') ds' dt' \right\}$$

Back to TMDP

wait:

- deterministic
- no effects on s

$$\Rightarrow \mathit{wait}(0) \text{ is a } \mathit{no-op}.$$

$\Rightarrow$ possibility to insert $\mathit{wait}(0)$ anywhere in any policy

$\Rightarrow$ a policy "(t', a)" summarizes any policy on TMDP

Equivalence ?

TMDP = XMDP with:

- discrete s + continuous t
- one single parametric action: *wait*
- other actions with no parameter
- total reward criterion

Equivalence ?

Transition model

$$p(s',t'|s,t,a) = \sum_{\mu \in M_{s'}} L(\mu|s,t,a) P_\mu(t'-t)$$

$$p(s',t'|s,t,wait(\tau)) = \delta_{s,t+\tau}(s',t')$$

Reward model

$$r(s,t,a) = \sum_{\mu \in M} \int_{t' \in \mathbb{R}^+} L(\mu|s,t,a) P_\mu(t'-t) r(\mu,t,t') dt'$$

$$r(s,t,wait(\tau)) = \int_t^{t+\tau} K(s,\theta) d\theta$$

Equivalence ?

$$V(s,t) = \sup_{t' \geq t} \left(\int_t^{t'} K(s,\theta) d\theta + \overline{V}(s,t') \right)$$

$$\overline{V}(s,t) = \max_{a \in A} Q(s,t,a)$$

$$Q(s,t,a) = \sum_{\mu \in M} L(\mu | s,t,a) \cdot U(\mu,t)$$

$$U(\mu,t) = \begin{cases} \int_{-\infty}^{\infty} P_\mu(t')[R(\mu,t,t') + V(s'_\mu,t')]dt' \\ \int_{-\infty}^{\infty} P_\mu(t'-t)[R(\mu,t,t') + V(s'_\mu,t')]dt' \end{cases}$$

Equivalence ?

. . . with the previous remarks . . .

Equivalence ?

$$V^*(s,t) = \sup_{\tau \in \mathbb{R}^+} \left(r(s,t,wait(\tau)) + \max_{a \in A \setminus \{wait\}} \left\{ \sum_{s' \in S} L(\mu_{s'}|s,a,t) \cdot \int_{t' \in \mathbb{R}^+} P_{\mu_{s'}}(t'-t) \left[r(\mu_{s'},t,t') + V^*(s',t') \right] dt' \right\} \right)$$

Equivalence

What allows for equation separation in the TMDP case ?

- *wait* is deterministic
- *wait* doesn't change the state

Policy optimization using polynomial representations

Piecewise polynomial interpolation of the model:

$$\begin{cases} d^\circ(P_\mu) = a \\ d^\circ(r) = d^\circ(V_n) = b \quad \Rightarrow d^\circ(V_{n+1}) = a+b+c+1. \\ d^\circ(L) = c \end{cases}$$

Policy representation using polynomial representations

Exact optimization possible if $a+b+c+1=b$

$$\rightarrow \text{with} \begin{cases} L \text{ piecewise constant} \\ P_\mu \text{ discrete pdf} \\ d^\circ(r) < 5 \end{cases}, d^\circ(V_{n+1}) = d^\circ(V_n),$$

(Boyan and Littman, 01) : $a=-1, b=1, c=0$

Approximate optimization by degree reduction of V_{n+1}

$\rightarrow$ $TMDP_{poly}$ method using spline interpolation

Extensions

- Several continuous actions ?
  $turn(\theta), go(x, V), \ldots$
- Methods based on other representations for continuous dynamics?
  $\longrightarrow$ standard pdf expressions
  $\longrightarrow$ representation-free algorithms (see perspectives)

Summary

Problem addressed: dealing with continuous observable time in MDP.

Goal: providing a sound framework for continuous or discrete time-dependent MDP and proving general optimality equations.

Contribution: The 'XMDP' extension to MDP and adaptation of Bellman equation.

What XMDP is not:

- not a completely new formalism $\longrightarrow$ extension of the MDP model
- not an algorithm $\longrightarrow$ algorithms might depend on the dynamics' representation

Summary

Problem addressed: dealing with continuous observable time in MDP.
Goal: providing a sound framework for continuous or discrete time-dependent MDP and proving general optimality equations.
Contribution: The 'XMDP' extension to MDP and adaptation of Bellman equation.

What XMDP is not:

- not a completely new formalism $\longrightarrow$ extension of the MDP model
- not an algorithm $\longrightarrow$ algorithms might depend on the dynamics' representation

Summary

Problem addressed: dealing with continuous observable time in MDP.
Goal: providing a sound framework for continuous or discrete time-dependent MDP and proving general optimality equations.
Contribution: The 'XMDP' extension to MDP and adaptation of Bellman equation.

What XMDP is not:

- not a completely new formalism $\longrightarrow$ extension of the MDP model
- not an algorithm $\longrightarrow$ algorithms might depend on the dynamics' representation

Summary

Problem addressed: dealing with continuous observable time in MDP.
Goal: providing a sound framework for continuous or discrete time-dependent MDP and proving general optimality equations.
Contribution: The 'XMDP' extension to MDP and adaptation of Bellman equation.

What XMDP is not:

- not a completely new formalism $\longrightarrow$ extension of the MDP model
- not an algorithm $\longrightarrow$ algorithms might depend on the dynamics' representation

Summary

Problem addressed: dealing with continuous observable time in MDP.
Goal: providing a sound framework for continuous or discrete time-dependent MDP and proving general optimality equations.
Contribution: The 'XMDP' extension to MDP and adaptation of Bellman equation.

What XMDP is not:

- not a completely new formalism $\longrightarrow$ extension of the MDP model
- not an algorithm $\longrightarrow$ algorithms might depend on the dynamics' representation

Perspectives

Difficulty in temporal planning: concurrency.

Concurrency and MDP: $\begin{cases} \text{CoMDP, Dec-MDP (concurrent actions)} \\ \text{GSMDP (concurrent events)} \end{cases}$

Our current focus: time-dependent GSMDP.

$\rightarrow$ can be translated to XMDP.

Main concern: “execution path” space too big.

Our approach: simulation-based approximate policy iteration.