Speeding Up Planning in Markov Decision Processes via Automatically Constructed Abstractions

Alejandro Isaza, Csaba Szepesvári, Vadim Bulitko and Russell Greiner

Department of Computing Science, University of Alberta

Motivation

TASK: *Stochastic shortest path*
= reaching some goal state when the effects of actions are stochastic

- special case of planning
- subclass of Markov Decision Problems
- medium size: fully enumerated state space

Here: *Multiple SSPs with the same domain*
Goal: *Speed up using Abstractions*

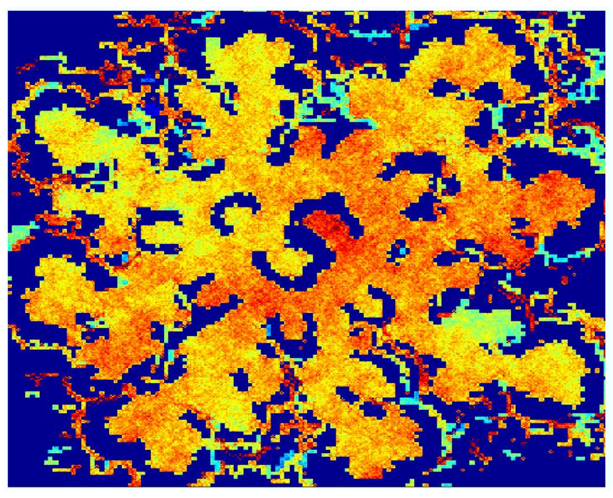
- construct a *multi-level hierarchy* of progressively simpler abstractions
- find a policy for the most abstract level, then recursively refine into a solution to the original problem.

Results:

- fully automated
- near-optimal solutions
- speed-up of ∼100 over a state-of-the-art MDP solver

Features

- Options based abstraction
- Multiple levels
- Deterministic abstractions

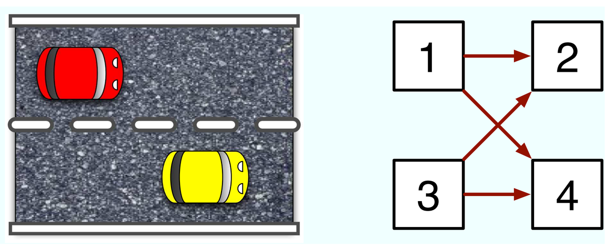


Path planning for agents in commercial video games (uncertainty of transitions ≈ map congestion)

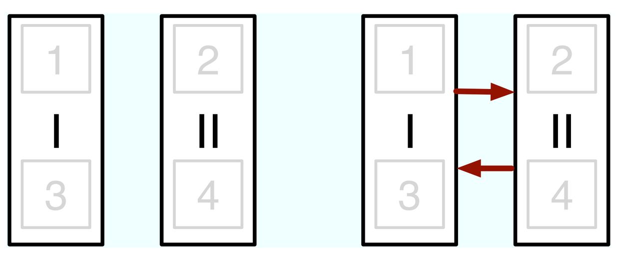
Control of multi-link robotic manipulators (uncertainty ≈ unmodeled dynamics, sensor and actuator noise)

Algorithm

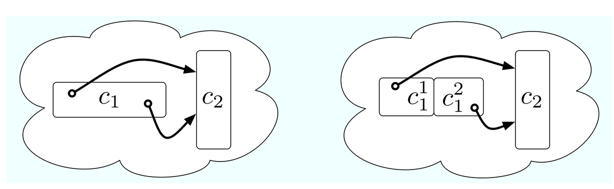
Building abstractions



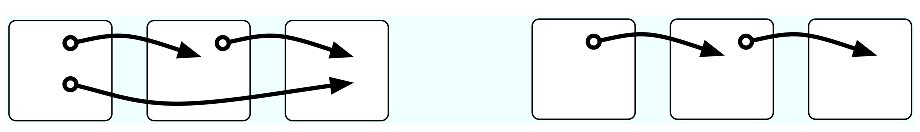
Highway domain



Clustering



Cluster splitting



Pruning

Planning

- Build a region around the goal, solve at the ground level
- Plan in the abstract graph
- Follow the ground options to execute the plan
- Follow the ground solution when entering the goal region

Related work

	This work	D97	KD03	Ha98	AsHu04/7	Moo99	LaKae01/2
State aggregation	+	+	+	+	+	+	+
Automatically built abstractions	+	-	-	-	+	+	+
Options	+	-	-	+	+	+	+
Option discovery	+	-	-	-	+	+	-
Lifted policy perf. bound	+	-	-	-	-	-	-
Deterministic abstractions	+	-	-	-	-	+	+
Experiments with $\lvert X\rvert \geq 10^3$	+	-	-	-	+	+	+

Theory

Results

Processes:
Lifting a policy $\tilde{\pi}$ of $\tilde{M}$ to M: $\tilde{\pi} \longmapsto H(\tilde{\pi})$
Projecting a policy π of M with ρ to $\tilde{X}$: $\pi \longmapsto L_\rho(\pi) \equiv (L_{\rho,\pi}(c), L_{\rho,\pi}(p))$.

Theorem 1:
Interpretation: The expected error of the lifted value function of the abstract policy $\tilde{\pi}$, relative to the base level policy π, is small if *(i)* aggregation does not loose detail of v_π and *(ii)* the projected costs and transitions underlying π are matched by the costs (resp., transitions) associated with $\tilde{\pi}$.

π: proper policy of M	$\tilde{\pi}$: proper policy of $\tilde{M}$
w: $w = w_\pi$	$\tilde{w}$: $\tilde{w} = w_{L_\rho(\pi)}$
γ: discount factor corresponding to w	$\tilde{\gamma}$: discount factor corresponding to $\tilde{w}$
λ: $\lambda = \max_x \frac{\tilde{w}(\tilde{x}(x))}{w(x)}$	

$$\|v_\pi - Ev_{\tilde{\pi}}\|_{w,\infty} \leq \frac{\|Av_\pi - v_\pi\|_{w,\infty}}{1-\gamma} + \lambda \frac{\|L_{\rho,\pi}(c) - \tilde{c}_{\tilde{\pi}}\|_{\tilde{w},\infty} + c_{\max}\|L_{\rho,\pi}(p) - \tilde{p}_{\tilde{\pi}}\|_{\tilde{w},1/\infty}}{1-\tilde{\gamma}}. \tag{1}$$

Theorem 2 (Simulation):
Interpretation: We can accurately simulate π with some policy $\tilde{\pi}$ of the abstract MDP provided all of the terms are small.

Call the right-hand side of (**??**) $B(\pi, \tilde{\pi})$. Let $w' = w_{H(\tilde{\pi})}$ and let γ' be the corresponding discount factor. Let $\hat{w} : X \to \mathbb{R}^+$ be arbitrary. Then

$$\|v_\pi - v_{H(\tilde{\pi})}\|_{\hat{w},\infty} \leq \left(\max_x \frac{w(x)}{\hat{w}(x)}\right) B(\pi,\tilde{\pi}) + \left(\max_x \frac{w'(x)}{\hat{w}(x)}\right) B(H(\pi),\tilde{\pi}).$$

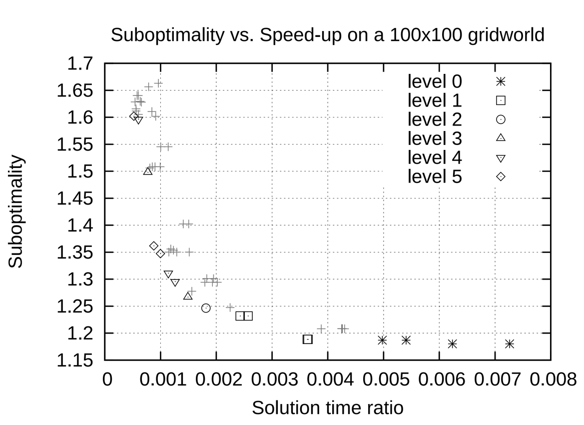
Notation

X:	state space
g:	$g \in X$; goal state
$M_{p,c}$:	$M_{p,c} = (X,p,c)$; Markov cost process; transitions: $p(y\vert x)$, costs: $c(x,y)$, $c \geq 0$
$v_{p,c}$:	$v_{p,c} : X \to \mathbb{R}$; cost-to-go function; $v_{p,c}(x) \triangleq \mathbb{E}[\sum_{t=0}^\infty c(x_t, x_{t+1})]$
w_p:	$w_p \triangleq v_{p,1}$; expected number of steps until the goal is reached
γ_p:	$(1-\gamma_p)^{-1} = \max_x w_p(x)$; discount factor underlying p
M:	$M = (X,A,p,c,g)$; SSP with action set A; transitions: $p(y\vert x,a)$, costs: $c(x,a,y)$
π:	$\pi : X \to A$; policy
p_π, c_π:	transitions and costs under π
v_π:	$v_\pi \triangleq v_{p_\pi,c_\pi}$; cost-to-go underlying π
w_π:	$w_\pi \triangleq w_{p_\pi}$; expected number of steps until the goal is reached
$\|p-q\|_{w,1/\infty}$:	$\|p-q\|_{w,1/\infty} \triangleq \max_x \sum_y \vert p(y\vert x) - q(y\vert x)\vert w(y)/w(x)$
$\|v\|_{w,\infty}$:	$\|v\|_{w,\infty} \triangleq \max_x \vert v(x)\vert / w(x)$

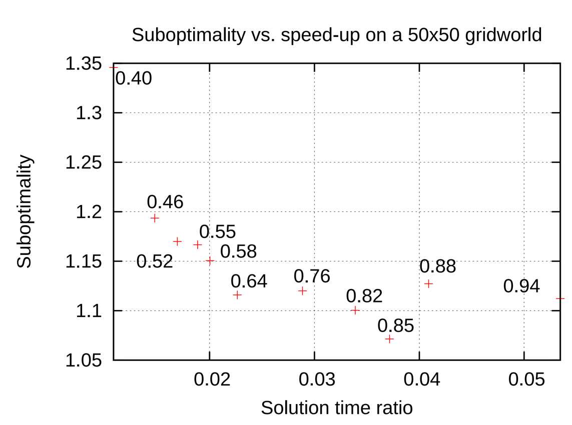
Abstractions

M:	$M = (X,A,p,c,g)$; original MDP
$c_{\max}$:	maximum cost in M
$\tilde{M}$:	$\tilde{M} = (\tilde{X}, \tilde{A}, \tilde{p}, \tilde{c}, \{g\})$; abstract MDP
$\tilde{x}(x)$:	abstract state of x
$S(x)$:	$S(x) = \{x' \in X : \tilde{x}(x) = \tilde{x}(x')\}$; peers of x
ρ:	state-randomization measure; $\rho : X \to [0,1]$, $\sum_{z\in S(x)} \rho(z) = 1\ \forall x \in X$
A_ρ:	$A_\rho : \mathbb{R}^X \to \mathbb{R}^{\tilde{X}}$, $(A_\rho v)(x) = \sum_{z \in S(x)} \rho(z) v(z)$; value aggregator
E:	$E : \mathbb{R}^{\tilde{X}} \to \mathbb{R}^X$, $(Ev)(x) = v(\tilde{x}(x))$; value extension

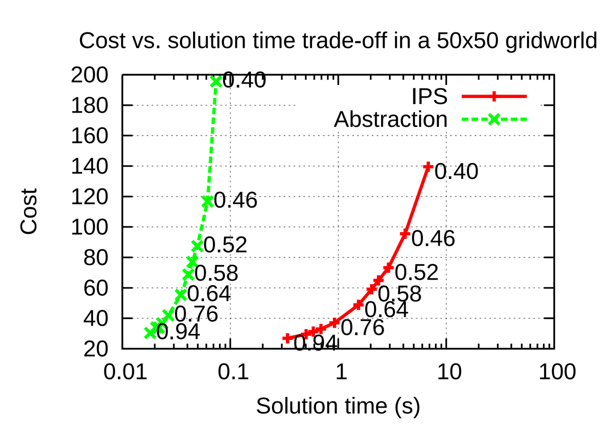
Results



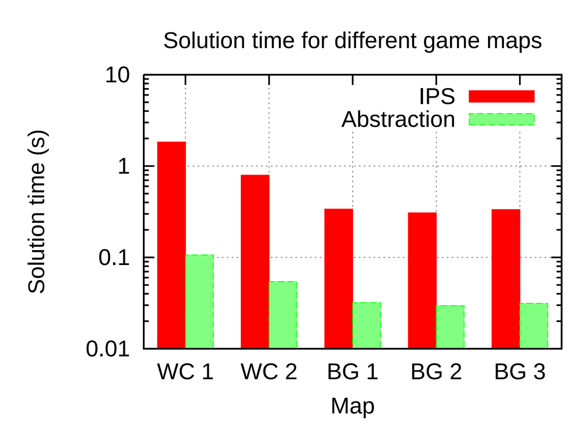
Subobtimality versus the solution time ratio as compared to IPS for different parameter configurations. The dominant configurations are shown for different levels of abstraction.



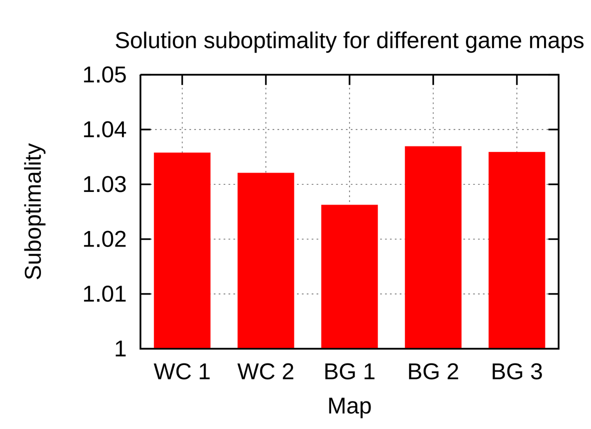
Subobtimality versus the solution time ratio as compared to IPS for different values of P.



Cost versus solution time for IPS and abstraction at different values of P.



Solution times for several game maps.



Solution suboptimalities for several game maps.