Hierarchy Discovery in Partially Observable Domains

April 9th, 2010

Reinforcement Learning Workshop, Barbados

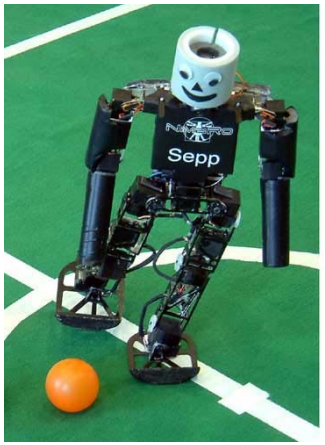
Presented by Pascal Poupart

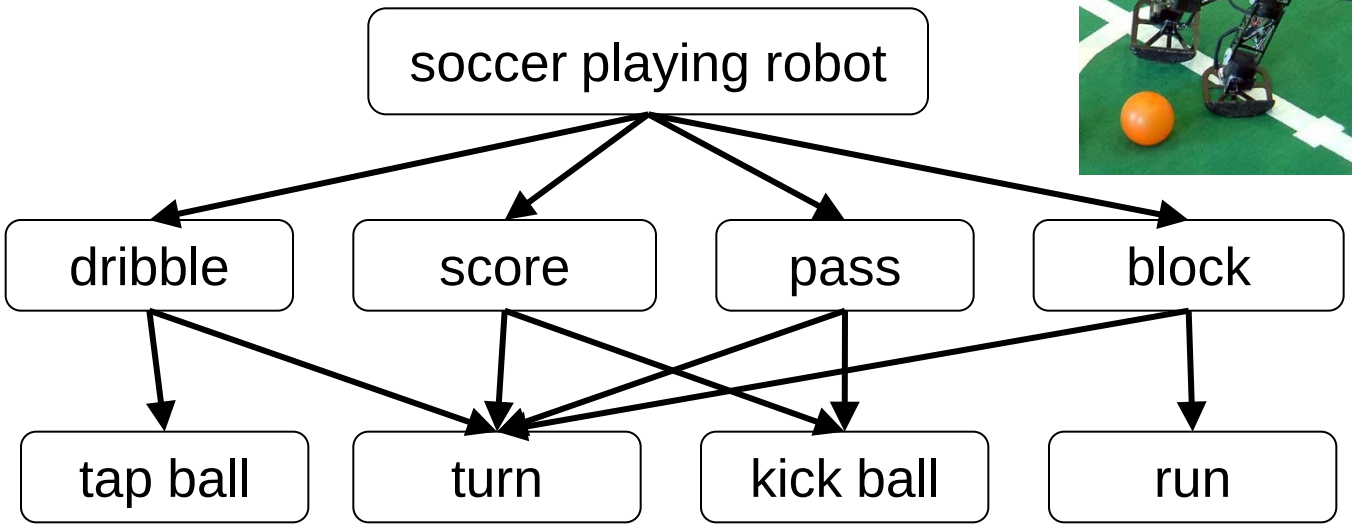
University of Waterloo, Canada

Joint work with Marc Toussaint and Laurent Charlin

Hierarchies in Planning

- Idea: task decomposed into subtasks arranged hierarchically





- Benefits: temporal/spatial abstraction, sub-policy reuse, intuitive policy representation

Where do hierarchies come from?

- Most of the time: expert knowledge
- Can be discovered automatically:
  - Loc… traversed or con…
  - Con… states or satisfy necessary conditions

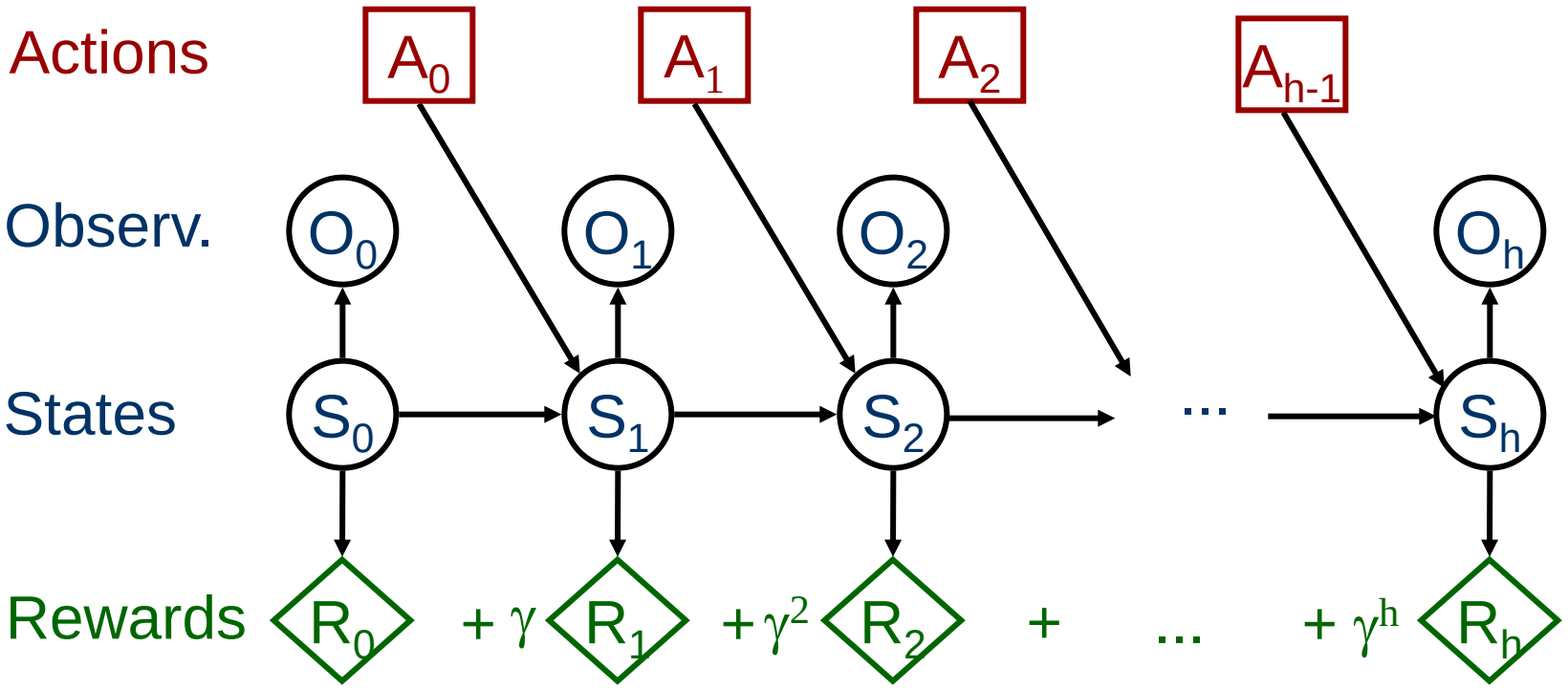
Common assumption: observable states/conditions

- How can we discover hierarchies in partially observable domains?

Outline

- Background:
  - Partially observable Markov decision processes
  - Finite state controllers
  - Hierarchical controllers
- Hierarchy discovery by
  - Non-convex optimization
  - Likelihood maximization
- Experiments
- Conclusion

POMDP Graphical Representation



Solution: policy π maximizes expected total rewards

Policy Optimization

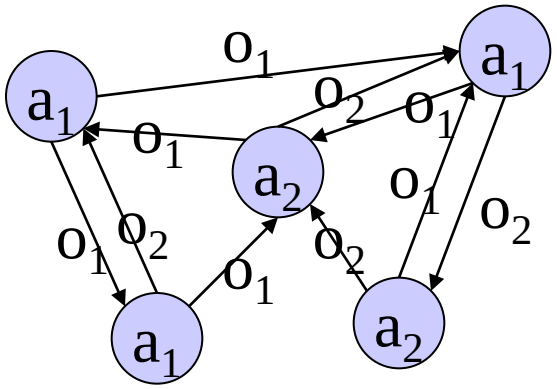
- Beliefs b: distribution over states
  - Bayes filtering: $b_{ao'}(s') = k \Sigma_s b(s) \Pr(s'|s,a) \Pr(o|s',a)$
  - Sufficient statistic of past observations and actions

- Policy $\pi : B \rightarrow A$
  - mapping from beliefs to actions

- Value function $V^{\pi}(b) = \Sigma_t \gamma^t E_{b_t|\pi} [R]$
- Optimal policy $\pi^*: V^*(b) \geq V^{\pi}(b)$ for all π, b
  - $V^*(b) = \max_a E_b[R] + \gamma \Sigma_{o'} \Pr(o'|s,a) V^*(b_{ao'})$

Finite State Controllers

- Alternative policy representation: controllers
  - Action mapping: α: N → A or Pr(a|n)
  - Next node mapping: σ: N x O → N or Pr(n'|a,n)



- Policy optimization: select best α and σ

Controller Optimization

- Policy search: alternate between
  - Policy evaluation
  - Policy improvement

- Optimization problem (Amato et al., 2007)

$$\max_{x,y} \quad \sum_s b_o(s) \underbrace{V_{n_o}(s)}_{y}$$

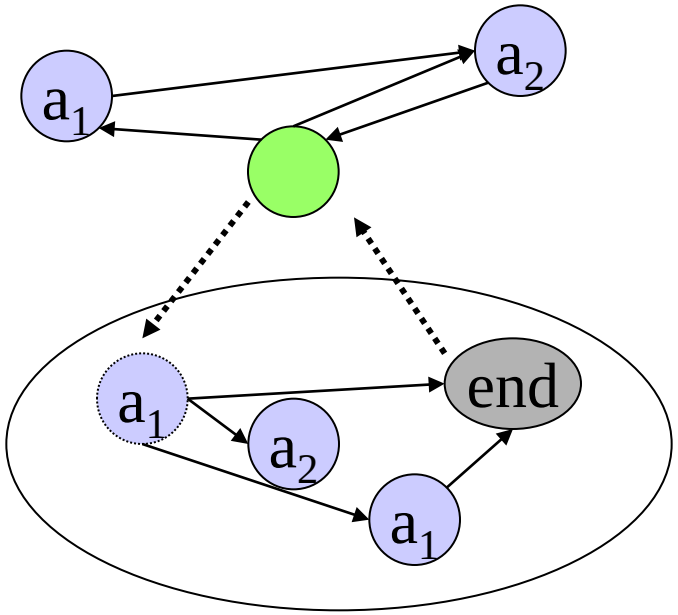
$$\text{s.t.} \quad \underbrace{V_n(s)}_{y} = \sum_{a,n'} \left[\underbrace{\Pr(a, n' | n, o_k)}_{x} R(s,a) + \sum_{s',o} \Pr_\gamma(s'|s,a) \Pr(o|s',a) \underbrace{\Pr(n', a|n, o)}_{x} \underbrace{V_{n'}(s')}_{y} \right] \quad \forall n, s$$

$$\underbrace{\Pr(n', a|n, o)}_{x} \geq 0 \quad \forall n', a, n, o \qquad \sum_{n',a} \underbrace{\Pr(n', a|n, o)}_{x} = 1 \quad \forall n, o$$

$$\sum_{n'} \underbrace{\Pr(n', a|n, o)}_{x} = \sum_{n'} \underbrace{\Pr(n', a|n, o_k)}_{x} \quad \forall a, n, o$$

Hierarchical Controllers

- Hansen & Zhou, 2005: let some nodes be sub-controllers
  - Action mapping: α: N → A or Pr(a|n)
  - Next node mapping: σ: N x O → N or Pr(n'|n,o)
  - Child mapping (abstract nodes): ϕ: N → N or Pr(n'|n^{par})
  - Special exit nodes



Hierarchical Controllers

- Local reward function: $R(s,\hat{a}) = \Sigma_n \Pr(n|n^{par})\ V_n(s)$
  - $V_n(s) = \Sigma_a \Pr(a|n)\ [R(s,a) + \gamma \Sigma_{s'n'o'} \Pr(s'|s,a) \Pr(o'|a,s') \Pr(n'|n,o')\ V_{n'}(s')]$
- Local transition prob: $\Pr(s'|\hat{a},s) = oc(s',n^{end})$
  - Discounted occupancy frequency
  - $oc(s',n') = \Pr(n'|n^{par}) + \gamma \Sigma_{snao'} oc(s,n) \Pr(a|n) \Pr(s'|s,a) \Pr(o'|a,s') \Pr(n'|n,o')$

Optimization Problem

- Charlin, Poupart & Shioda, 2006

$$\max_{w,x,y,z} \quad \sum_{s \in S} b_0(s) \underbrace{V_{n_0}(s)}_{y} \tag{3}$$

$$\text{s.t.} \quad \underbrace{V_n(s)}_{y} = \sum_{a,n'} \left[\underbrace{\Pr(n', a | n, o_k)}_{x} R(s,a) + \sum_{s',o} \Pr_\gamma(s'|s,a) \Pr(o|s',a) \underbrace{\Pr(n', a|n,o)}_{x} \underbrace{V_{n'}(s')}_{y} \right] \quad \forall s, n \tag{4}$$

$$\underbrace{V_{\bar{n}}(s)}_{y} = \sum_{n_{beg}} \underbrace{\Pr(n_{beg}|\bar{n})}_{z} \left[\underbrace{V_{n_{beg}}(s)}_{y} + \sum_{s_{end},a,n'} \underbrace{\mathrm{oc}(s_{end}, n_{end} | s, n_{beg})}_{w} \left[\underbrace{\Pr(n', a|\bar{n}, o_k)}_{x} R(s_{end}, a) \right.\right.$$

$$\left.\left. + \sum_{s',o} \Pr_\gamma(s'|s_{end}, a) \Pr(o|s',a) \underbrace{\Pr(n', a|\bar{n}, o)}_{x} \underbrace{V_{n'}(s')}_{y} \right] \right] \quad \forall s, \bar{n} \tag{5}$$

$$\underbrace{oc(s', n' | s_0, n_0)}_{w} = \delta(s', n', s_0, n_0) + \sum_{s,o,a} \Bigg[\tag{6}$$

$$\left. \sum_n \underbrace{oc(s, n | s_0, n_0)}_{w} \Pr_\gamma(s'|s,a) \Pr(o|s',a) \underbrace{\Pr(n', a|n,o)}_{x} \right] \Bigg\} \; n \text{ concrete (6a)}$$

$$\left. \begin{array}{l} + \sum_{s_{end}, n_{beg}, \bar{n}} \underbrace{oc(s, \bar{n} | s_0, n_0)}_{w} \Pr_\gamma(s'|s_{end}, a) \Pr(o|s',a) \\ \underbrace{oc(s_{end}, n_{end} | s, n_{beg})}_{w} \underbrace{\Pr(n', a|\bar{n}, o)}_{x} \underbrace{\Pr(n_{beg}|\bar{n})}_{z} \Big] \quad \forall s_0, s', n_0, n' \end{array} \right\} \; \bar{n} \text{ abstract (6b)}$$

$$\Pr(\bar{n}'|\bar{n}) = 0 \text{ if } label(\bar{n}') \leq label(\bar{n}), \forall \bar{n}, \bar{n}' \tag{7}$$

Optimization Techniques

- Non-convex quartic optimization problem
  - Non-convex solvers do not scale
- Approximations:
  - Mixed-integer non-linear programming
  - Bounded hierarchical policy iteration
  - Do not scale either

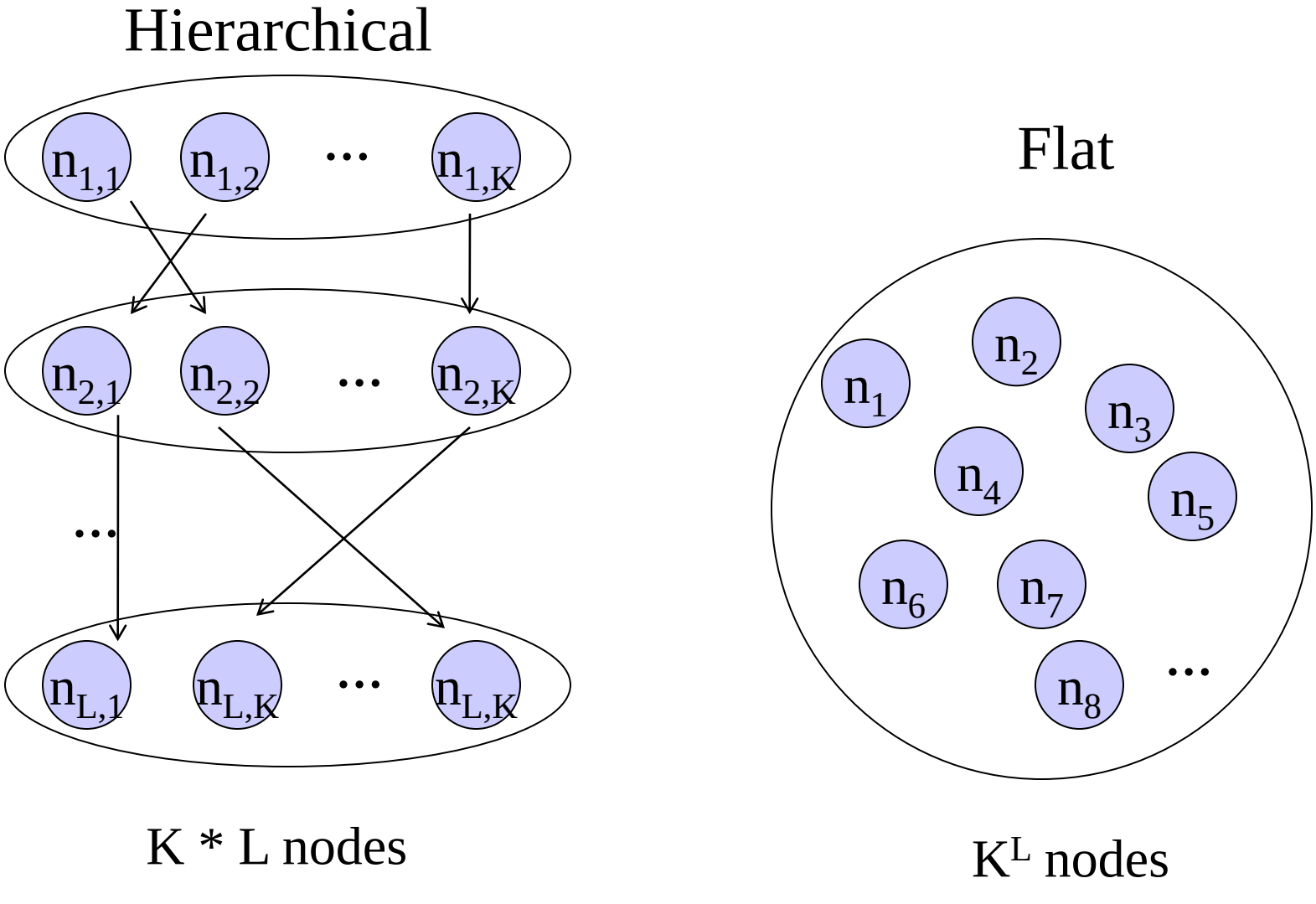
Hierarchy Discovery

- Why discover the hierarchy since this seems to make the problem harder?

- Problem complexity remains the same
  - Search in a different policy space
  - Bias the search towards hierarchical policies
  - May find exponentially more compact policies

- Reveal interesting structure
  - Facilitate policy explanation for humans

Hierarchical vs Flat Controllers

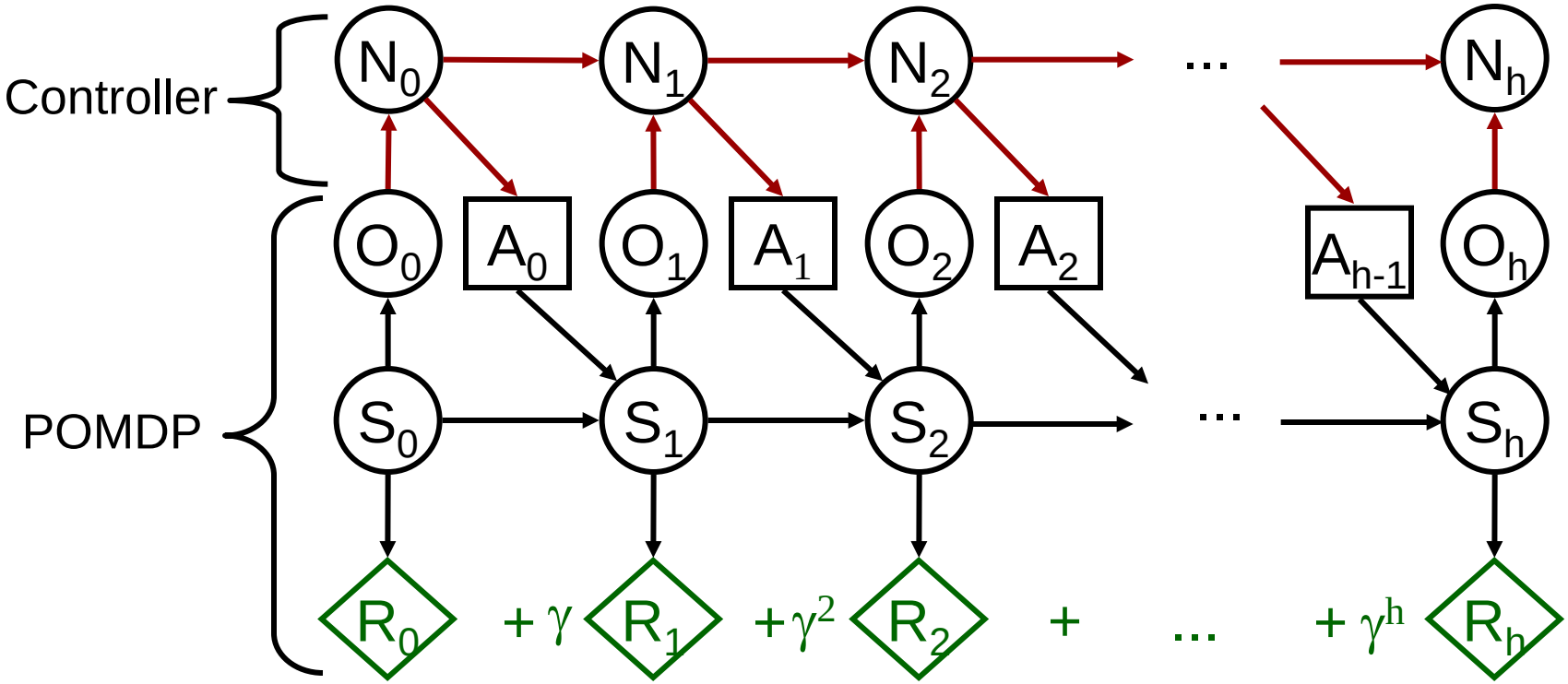


Controllers as DBNs

- Alternative: planning as inference
  - Convert POMDP with hierarchical controller into a mixture of DBNs
  - Reduce policy optimization to a maximum likelihood estimation problem
  - Use EM
  - Scalable!

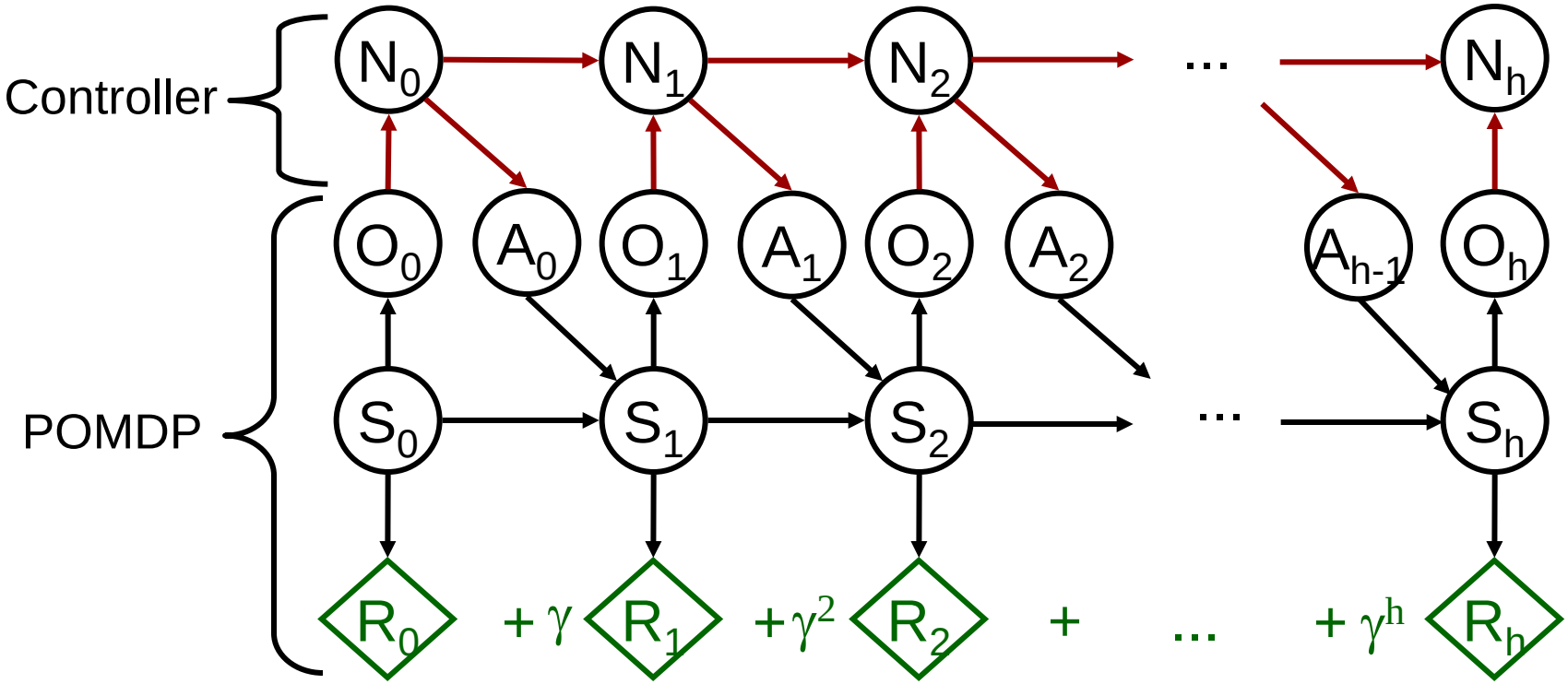
Graphical Model

- Meuleau et al., 1999



Graphical Model

- Toussaint et al., 2006:
  - Optimize $Pr(A_t|N_t)$ & $Pr(N_{t+1}|N_t,O_{t+1})$ by EM

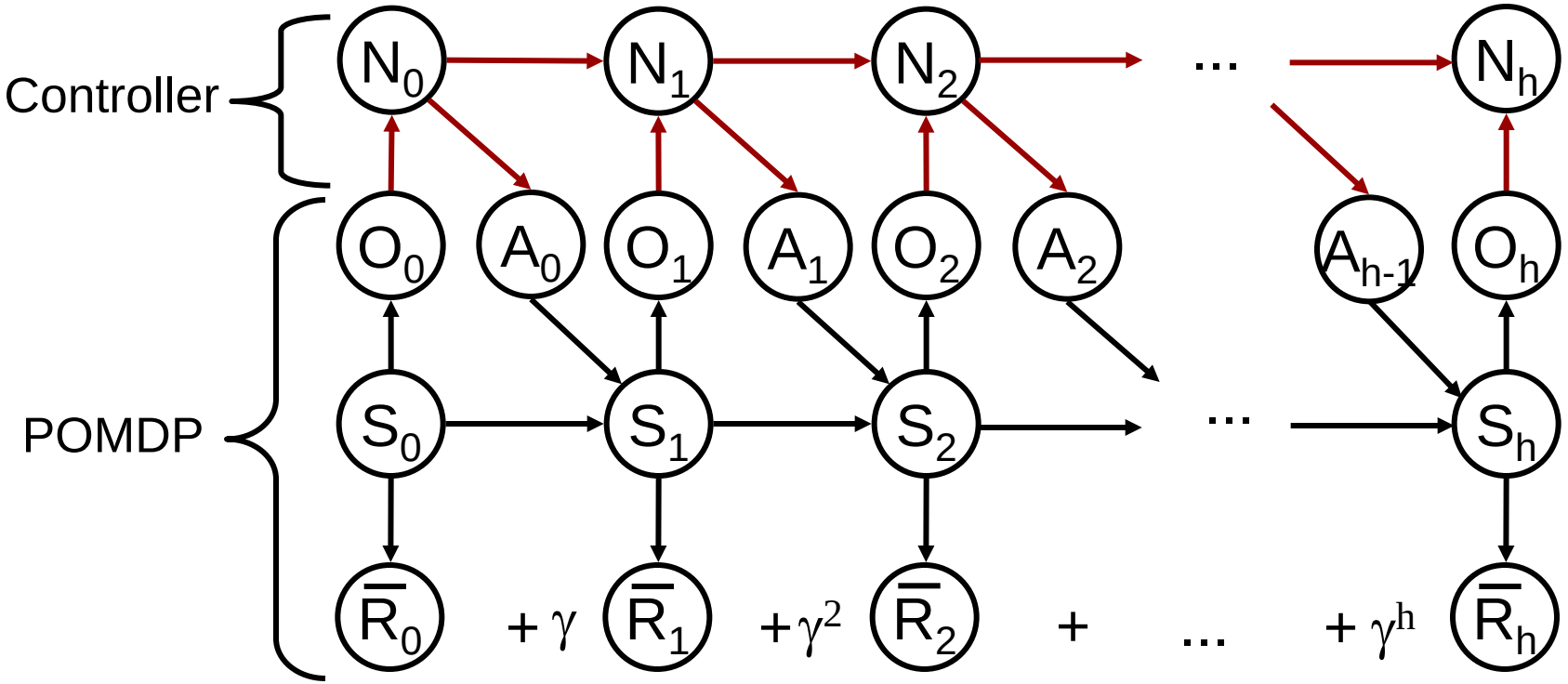


Graphical Model

- Normalize rewards

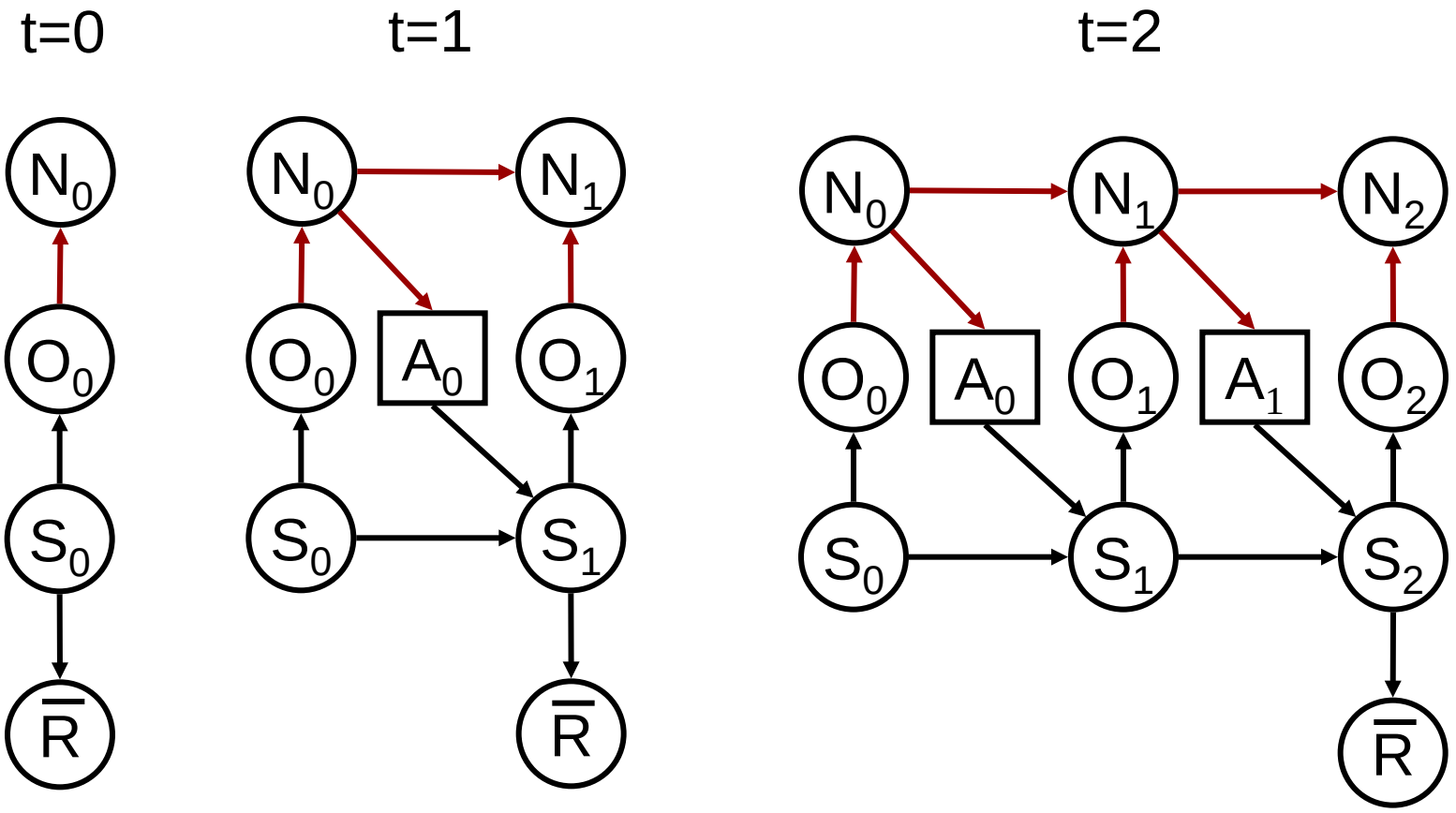
  – $\overline{R}_t \in \{0,1\}$

$$\Pr(\overline{R}_t=1) = \frac{R(s_t) - r_{min}}{r_{max} - r_{min}}$$



Graphical Model

- Mixture of DBNs: $Pr(t) = k\ \gamma^t$



Controllers as DBNs

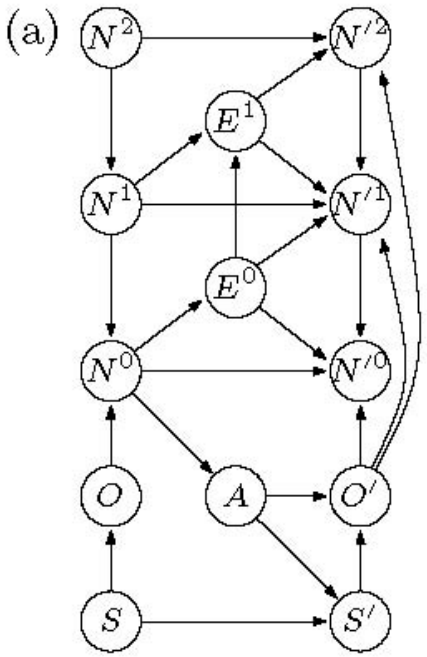
- Policy optimization
  - Maximize likelihood of $\bar{R}$=1:
    - $\text{argmax}_{Pr(a|n),Pr(n'|n,o')} Pr(\bar{R}=1)$
  - Expectation maximization

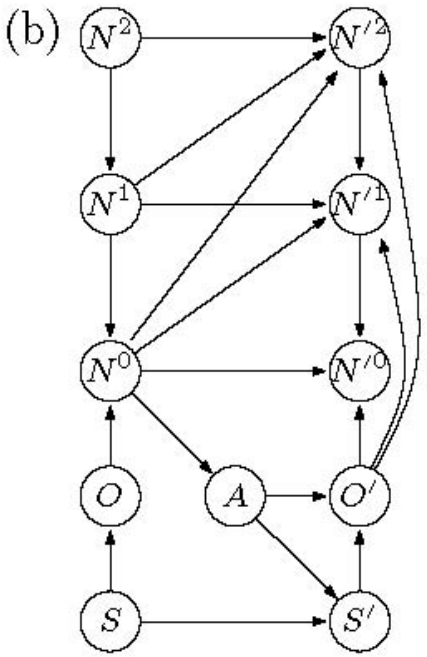
- **Advantage: any inference algorithm can be used**

Factored Controllers

- Toussaint, Charlin & Poupart, 2008
  - Hierarchical controllers ⇔ DBN mixture
  - More generally: factored controllers

Hierarchical controller





Factored controller

Factored Controllers

- Hierarchy discovery & policy optimization
  - Maximize likelihood of $\overline{R}=1$
  - Expectation Maximization

- Same problem as non-convex quartic opt. prob.
  - EM much faster than optimization-based techniques
  - But EM gets stuck in local optima more easily

Results

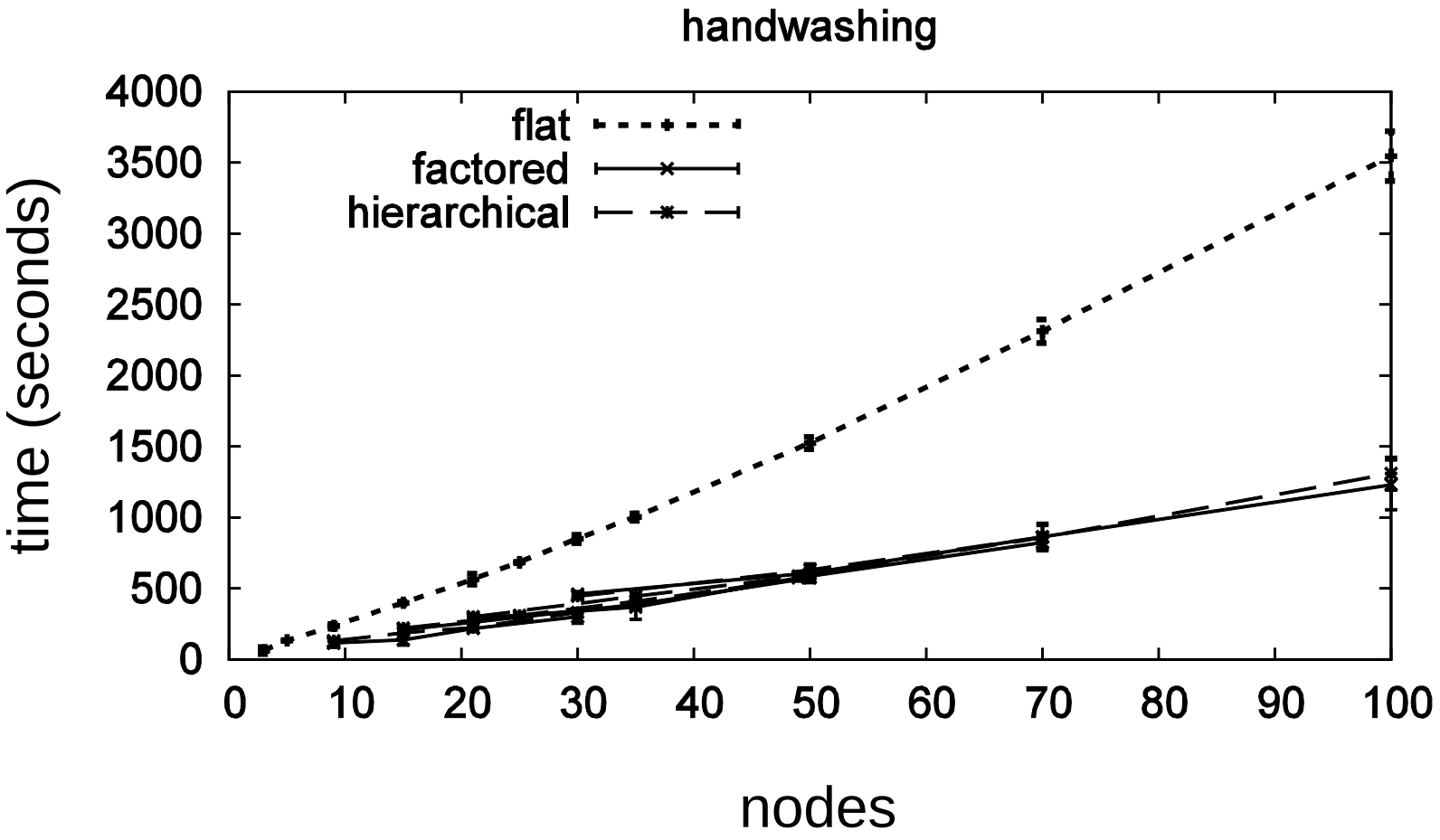
- Benchmark problems

Non-convex optimization (NEOS server)

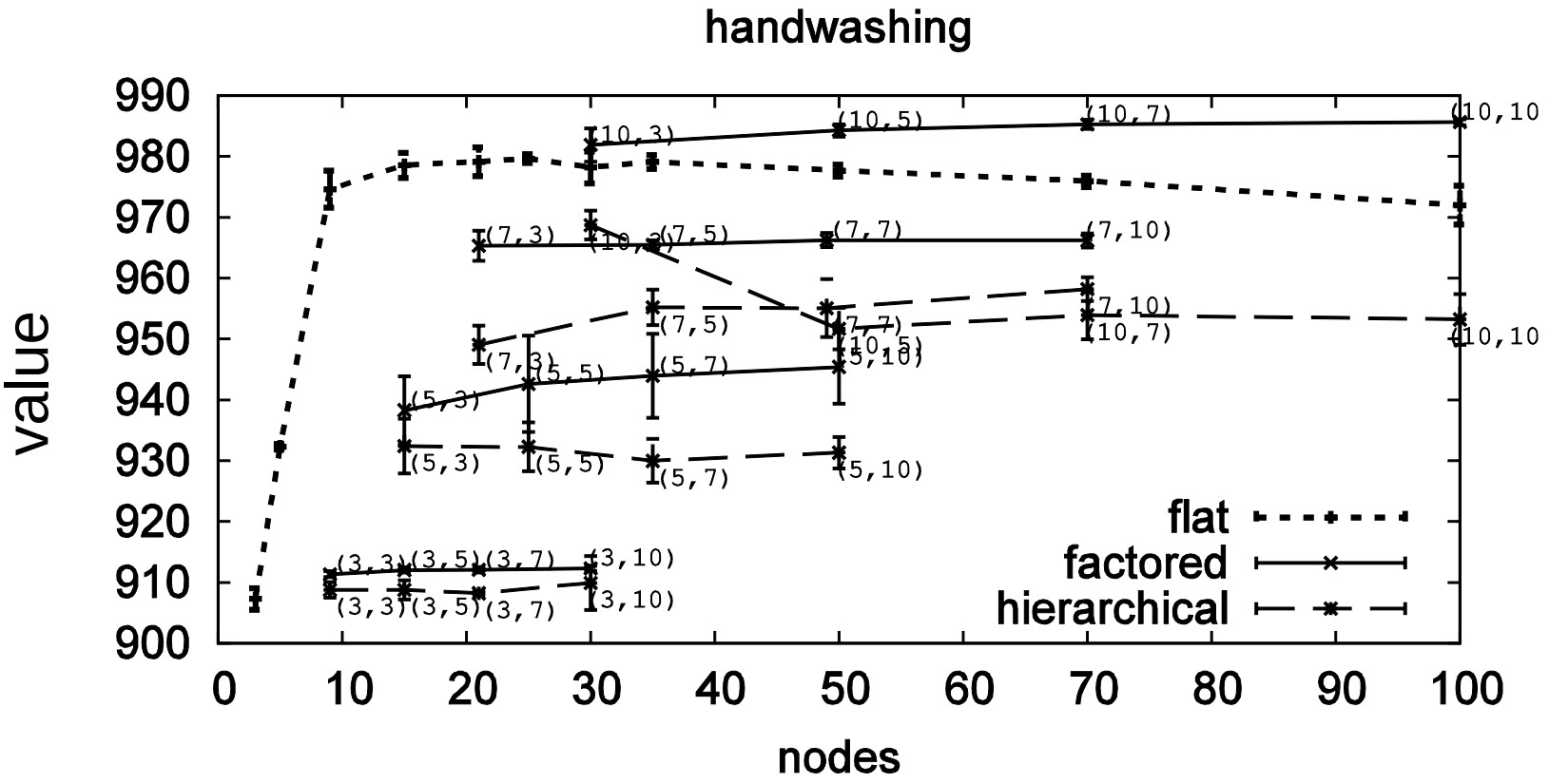
EM

Problem	$\|\mathcal{S}\|, \|\mathcal{A}\|, \|\mathcal{O}\|$	V^*	HSVI2 V	Best results from [4] nodes	t(s)	V	ML approach (avg. over 10 runs) nodes	t(s)	V
paint	4, 4, 2	3.28	3.29±0.04	(1,3)	<1	3.29	(5,3)	0.96±0.3	3.26±0.004
shuttle	8, 3, 5	32.7	32.9±0.8	(1,3)	2	31.87	(5,3)	2.81±0.2	31.6±0.5
4x4 maze	16, 4, 2	3.7	3.75±0.1	(1,2)	30	3.73	(3,3)	2.8±0.8	$3.72 \pm 8e-5$
chain-of-chains	10, 4, 1	157.1	157.1±0	(3,3)	10	0.0	(10,3)	6.4±0.2	151.6±2.6
handwashing	84, 7, 12	⩽1052	N/A			N/A	(10,5)	655±2	984±1
cheese-taxi	33, 7, 10	⩽5.3	2.53±0.3			N/A	(10,3)	311±14	−9±11(2.25*)

Results (EM)

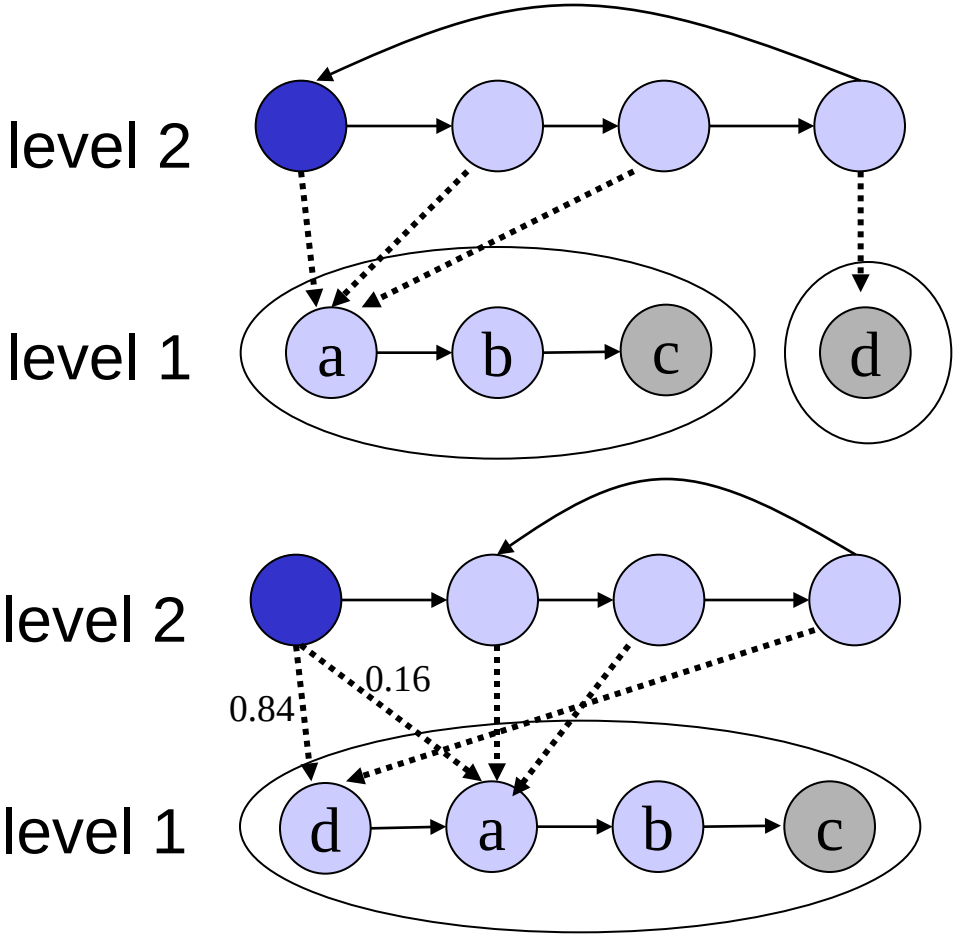

handwashing
flat
factored
hierarchical
time (seconds)
nodes
0
500
1000
1500
2000
2500
3000
3500
4000
0
10
20
30
40
50
60
70
80
90
100

Results


handwashing
value
nodes
990
980
970
960
950
940
930
920
910
900
0
10
20
30
40
50
60
70
80
90
100
flat
factored
hierarchical
(10,3)
(10,5)
(10,7)
(10,10
(7,3)
(7,5)
(7,7)
(7,10)
(5,3)
(5,5)
(5,7)
(5,10)
(3,3)
(3,5)
(3,7)
(3,10)

Non-intuitive hierarchy

- Simple unobservable POMDP:
  - Reward when executing d after doing a-b-c three times
  - Controller found:



Conclusion

- Simultaneous hierarchy discovery and planning
  - Direct optimization: not scalable
  - Planning as inference with EM: scalable but subject to local optima
  - Hierarchies discovered are unintuitive
- Future work:
  - Technique to escape local optima
  - Reinforcement learning (Vlassis & Toussaint, 2009)

Selected work relevant to workshop

- Feature extraction for Walker User Behaviour Recognition
  - UAI submission
- Bayesian RL
  - An Analytic Solution to Discrete Bayesian Reinforcement Learning, ICML 2006
- Value-directed information bottleneck
  - Value-directed Compression of POMDPs, NIPS 2002
- Policy explanation
  - Minimal Sufficient Explanations for Factored Markov Decision Processes, ICAPS 2009