Analyzing and Escaping Local Optima in Planning as Inference for Partially Observable Domains

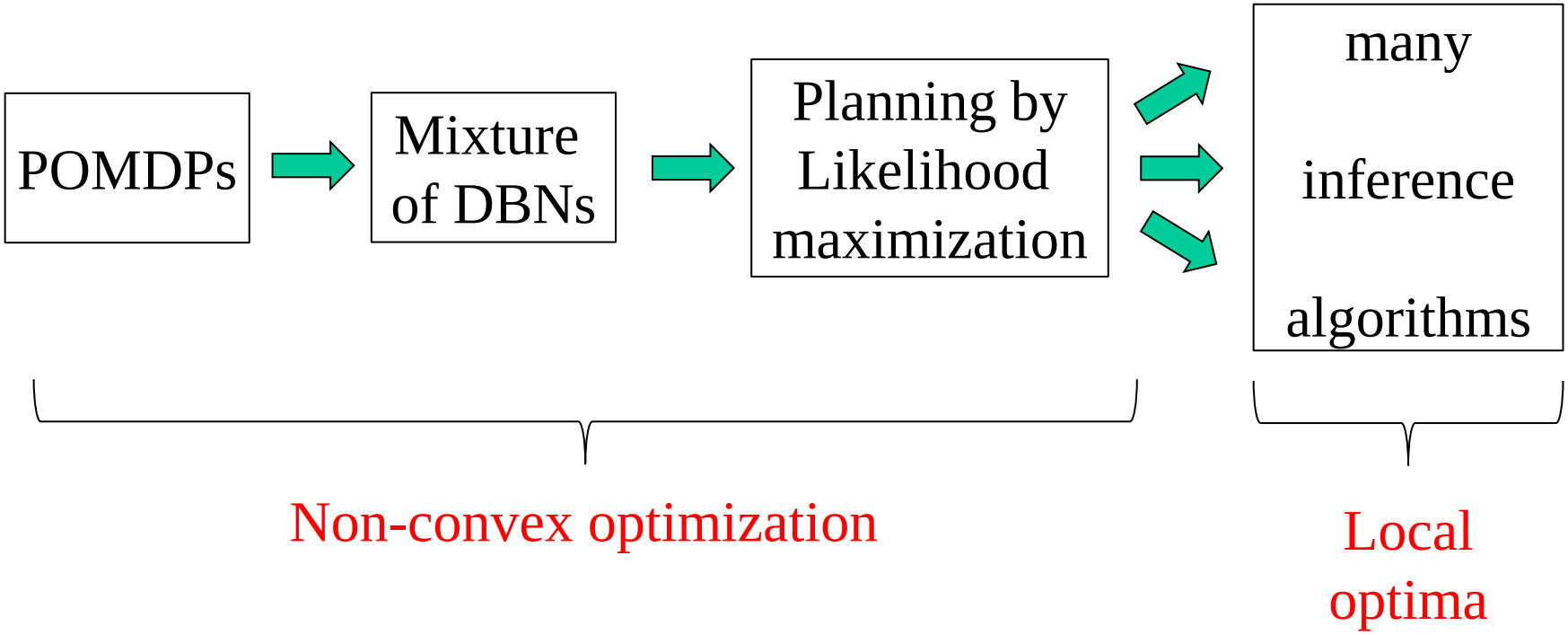
September 8th, 2011
ECML-PKDD, Athens, Greece

Presented by Pascal Poupart
University of Waterloo, Canada

Joint work with Tobias Lang and Marc Toussaint
FU Berlin, Germany

Introduction

- Planning as inference

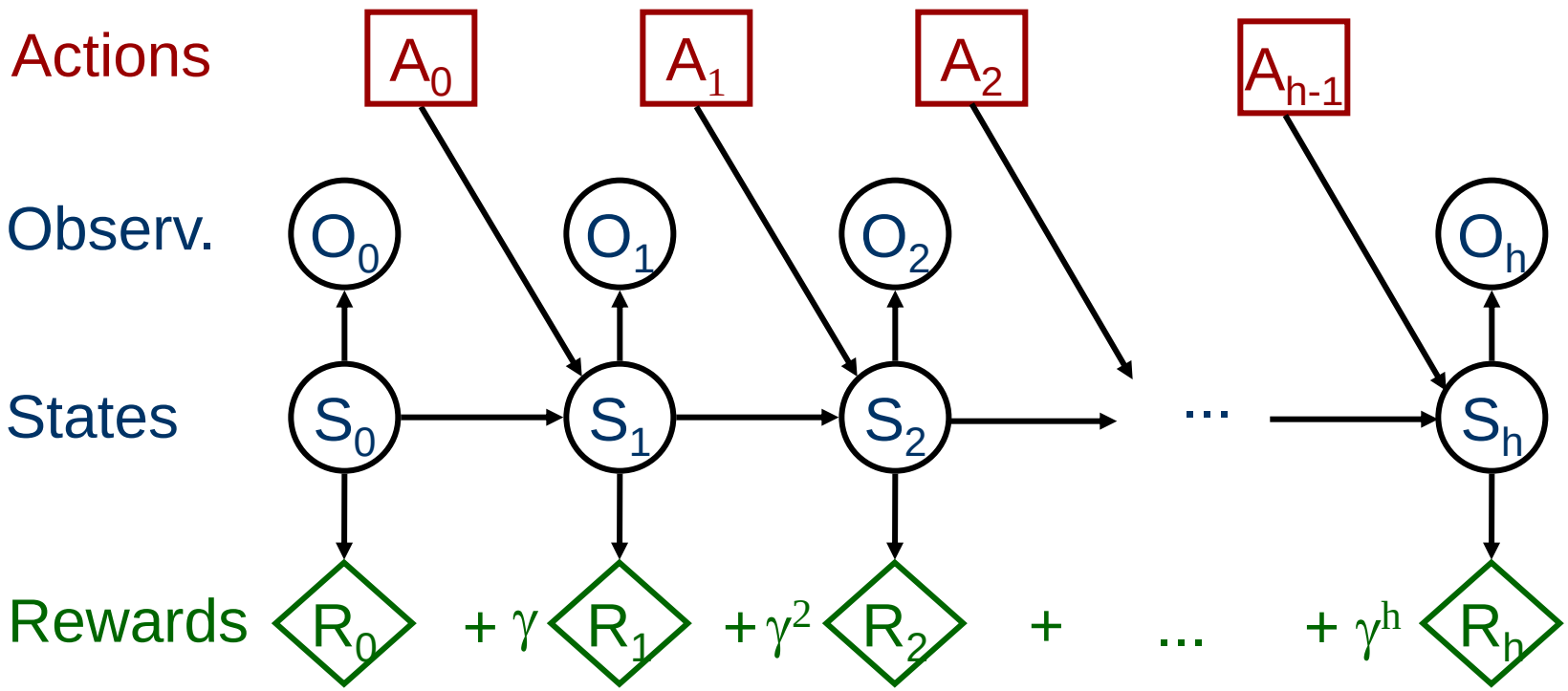


- Contributions:
  - Analysis and interpretation of EM's local optima
  - Two escape techniques

Outline

- Background:
  - POMDPs
  - Planning as inference
- Local Optima Interpretation
  - one-step look ahead optimality
- Escape Techniques
  - Forward search
  - Node Splitting
- Experiments
- Conclusion

POMDP Graphical Representation



Solution: policy π maximizes expected total rewards

Policy Optimization

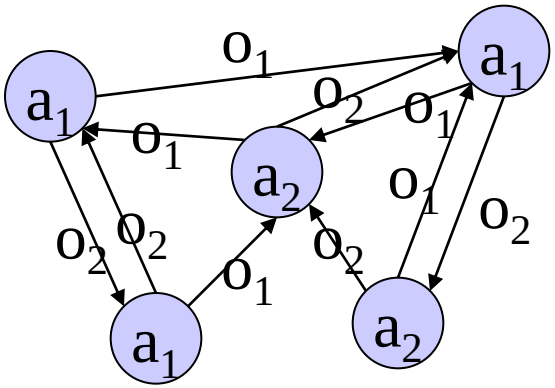
- Beliefs b: distribution over states
  - Bayes filtering: $b_{ao'}(s') \propto \sum_s b(s) \Pr(s'|s,a) \Pr(o'|s'a)$
  - Sufficient statistic of past observations and actions

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

- Value function $V^\pi(b) = \sum_t \gamma^t E_{b_t|\pi}[R_t]$
- Optimal policy $\pi^* : V^*(b) \geq V^\pi(b)\ \forall \pi, b$
  - Bellman eqn: $V^*(b) = \max_a E_b[R] + \gamma \sum_{o'} \Pr(o'|b,a) V^*(b_{ao'})$

Finite State Controllers

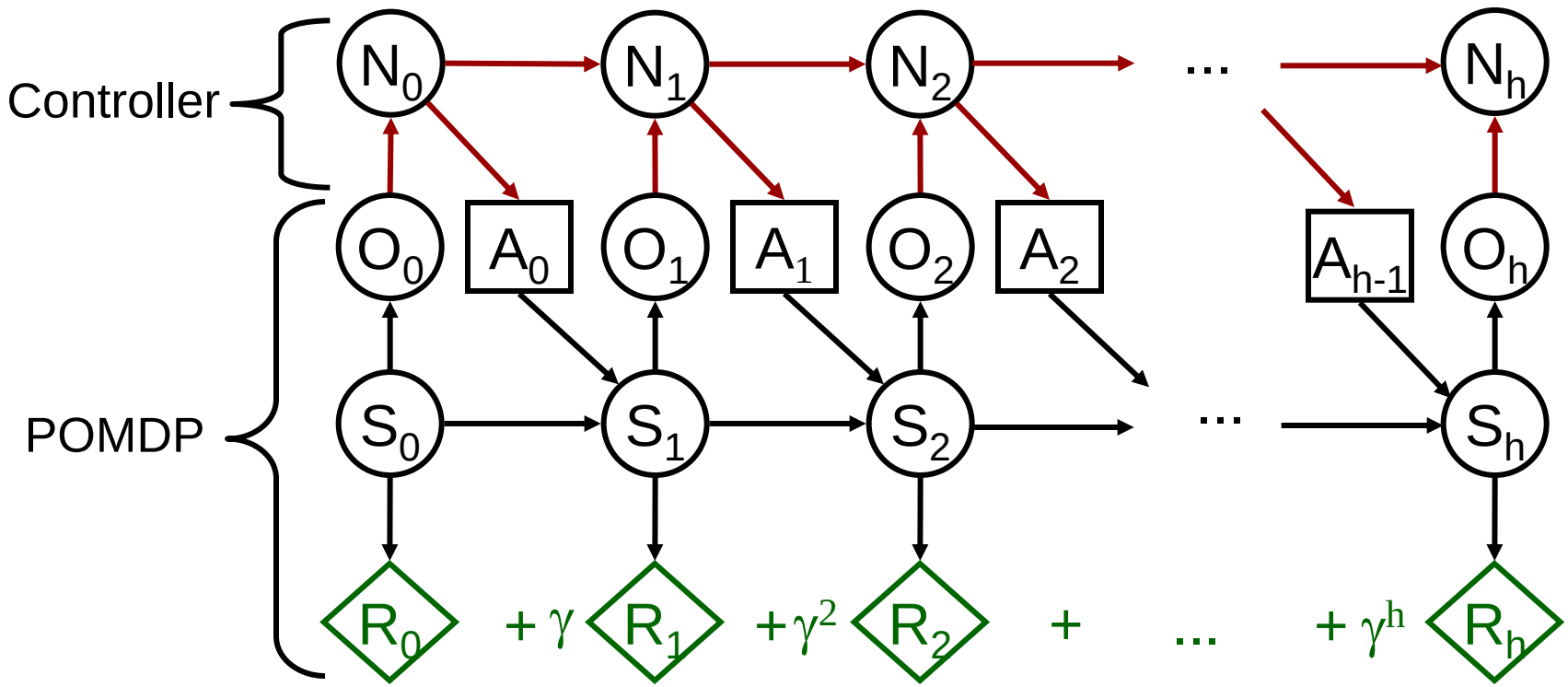
- Alternative policy representation: controllers
  - Action mapping: $\delta: N \leftarrow A$ or $\Pr(a|n)$
  - Next node mapping: $\sigma: N \times O \rightarrow N$ or $\Pr(n'|n, o')$



- Policy optimization: select best δ and σ

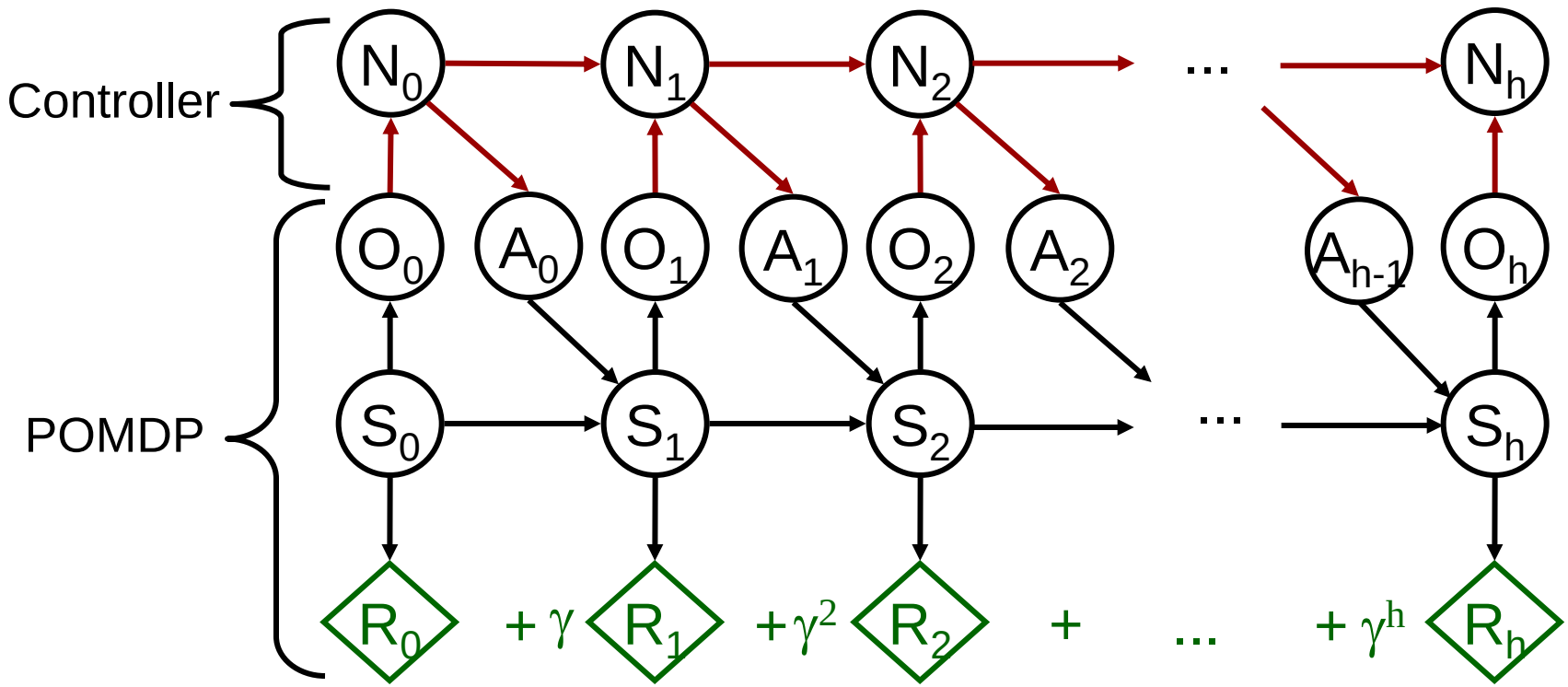
Graphical Model

- 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

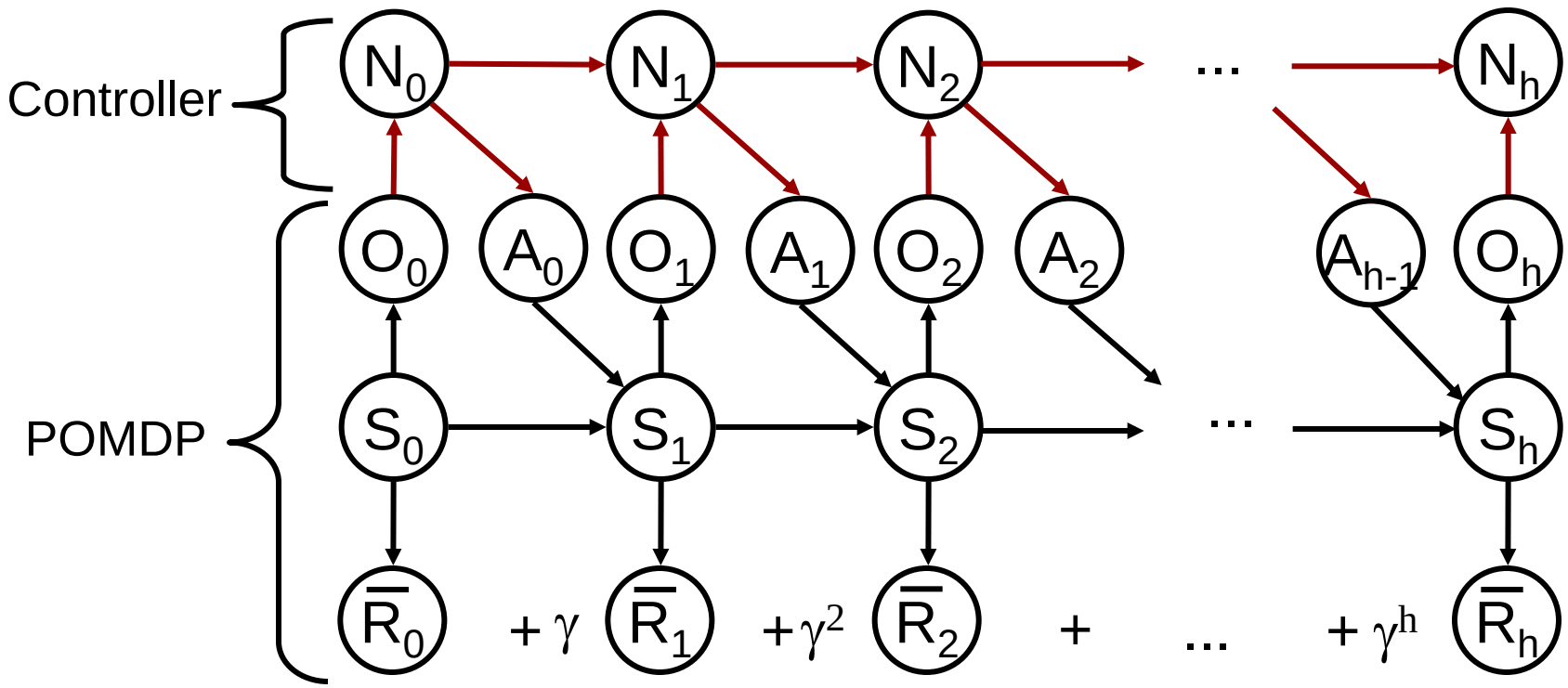


Graphical Model

- Normalize rewards

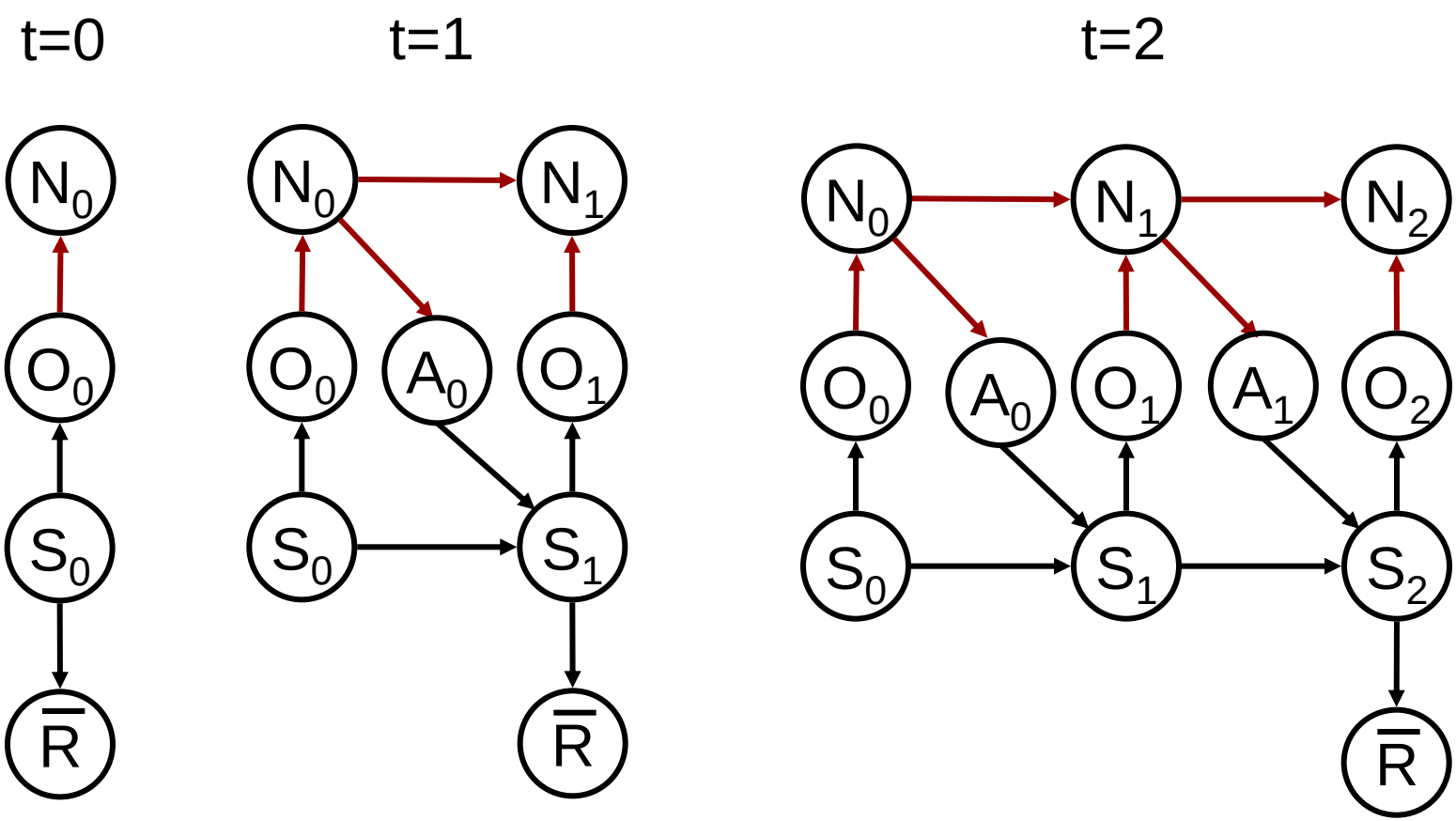
$$\bar{R}_t \in \{0,1\}$$

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



Graphical Model

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



Planning as Inference

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

- **Advantages:**
  - Any inference algorithm can be used
  - Versatile:
    - hierarchical planning [Toussaint et al. 2008],
    - continuous states and actions [Hoffman et al. 2009],
    - reinforcement learning [Vlassis et al. 2009],
    - multi-agent systems [Kumar et al. 2010]

Local Optima

- EM local optima: necessary, but not sufficient optimality conditions
  - One-step lookahead optimality
- Escaping local optima
  - Multi-step lookahead forward search
  - Node splitting

EM details

- Parameter updates:

$$\delta^{i+1}(a|n) \propto \delta^{i}(a|n) \underbrace{\sum_{ss'o'n'} \alpha_{sn}[\Pr(\bar{R}=1) + \gamma \Pr(s',o'|s,a)\, \sigma^{i}(n'|o',n)\beta_{n's'}]}_{g_{an}}$$

$$\sigma^{i+1}(n'|o',n) \propto \sigma^{i}(n'|o',n) \underbrace{\sum_{ss'a} \alpha_{sn}[\delta(a|n) \Pr(s',o'|s,a)\, \beta_{n's'}]}_{h_{n'o'n}}$$

where

$$\alpha^{t}_{s'n'} = b(s')1_{n'=n_0} + \gamma \sum_{asno'} \alpha^{t-1}_{sn} \delta(a|n) \Pr(s',o'|s,a)\, \sigma(n'|o',n)$$

$$\beta^{t}_{sn} = \sum_{as'n'o'} \delta(a|n)[\Pr(\bar{R}=1) + \gamma \Pr(s',o'|s,a)\, \sigma(n'|o',n)\beta^{t-1}_{s'n'}]$$

Local Optima Conditions

- Theorem 1: If a policy $\pi =< \delta, \sigma >$ is a stable fixed point of EM then:

$$\forall an \text{ if } \delta(a|n) \neq 0 \text{ then } g_{an} = \max_{\tilde{a}} g_{\tilde{a}n}$$

$$\forall n'o'n \text{ if } \sigma(n'|o'n) \neq 0 \text{ then } h_{n'o'n} = \max_{\tilde{n}'} h_{\tilde{n}'o'n}$$

Optimality

- Local optimality (EM):

$$g_{an} = \max_{\tilde{a}} \sum_{ss'o'n'} \alpha_{sn} [\Pr(\bar{R} = 1) + \gamma \Pr(s', o' | s, \tilde{a}) \sigma(n' | o', n) \beta_{n's'}] \; \forall n$$

$$h_{n'o'n} = \max_{\tilde{n}'} \sum_{ss'a} \alpha_{sn} [\delta(a|n) \Pr(s', o' | s, a) \beta_{\tilde{n}'s'}] \forall o'n$$

- Global optimality (Bellman Eqn):

$$a^n = argmax_{\tilde{a}} \sum_{ss'o'n'} b_n(s) \left[R(s, \tilde{a}) + \gamma \Pr(s', o' | s, \tilde{a}) V_{n^{nbo'}s'} \right] \forall nbo'$$

$$n^{nbo'} = argmax_{\tilde{n}'} \sum_{ss'} b_n(s) [\Pr(s'o' | s, a) V_{\tilde{n}'s'}] \; \forall o'nba$$

Local Optima Conditions

- Theorem 2: the conditions

$$\forall an \text{ if } \delta(a|n) \neq 0 \text{ then } g_{an} = max_{\tilde{a}}\, g_{\tilde{a}n}$$

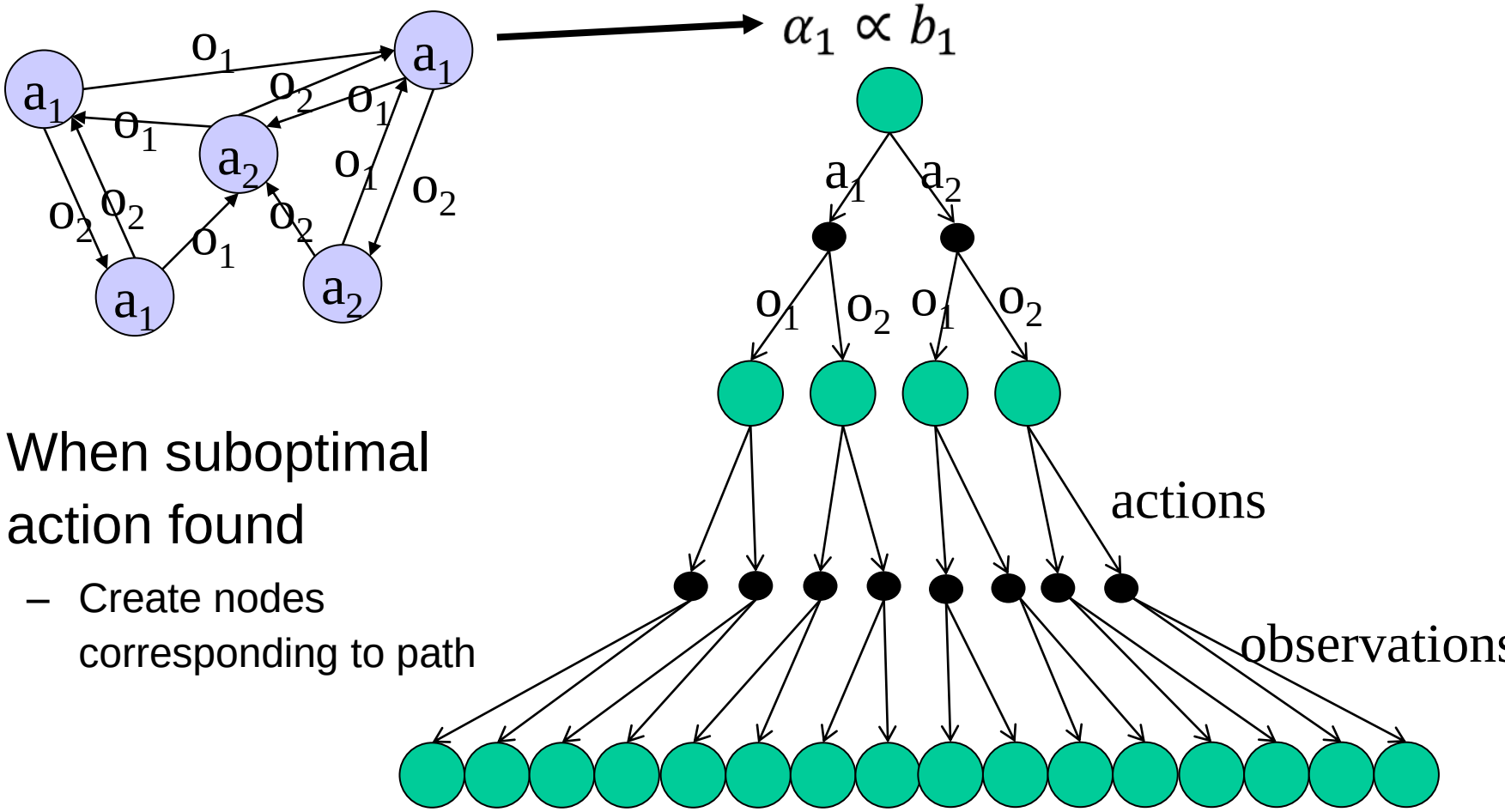
$$\forall n'o'n \text{ if } \sigma(n'|o'n) \neq 0 \text{ then } h_{n'o'n} = max_{\tilde{n}'}\, h_{\tilde{n}'o'n}$$

  are necessary, but not sufficient for global optimality

- EM ensures only one-step lookahead optimality

- Escape local optima by
  - multistep forward search
  - Node splitting

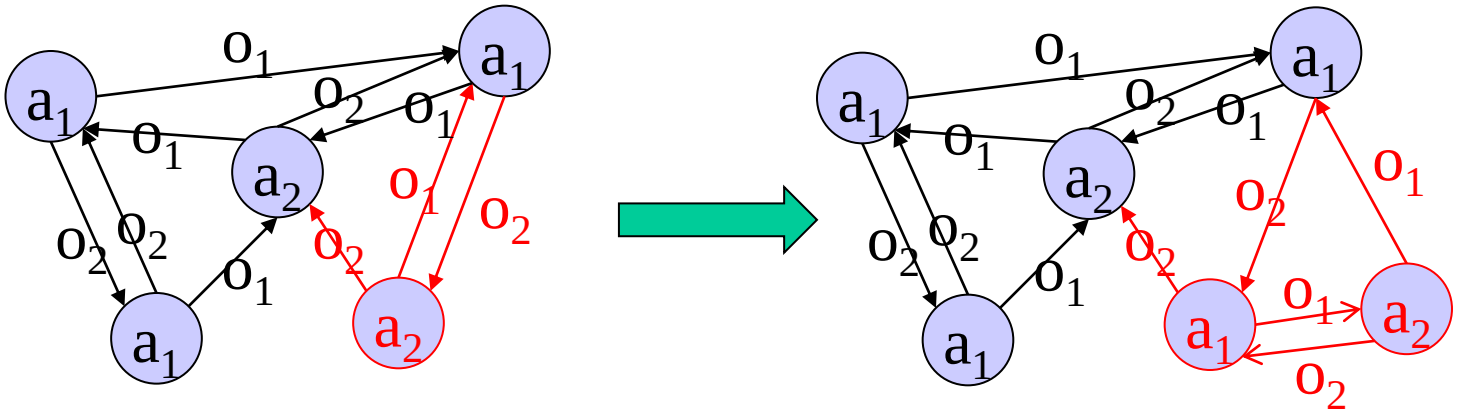
Multi-step forward search



- When suboptimal action found
  - Create nodes corresponding to path

Node Splitting

- Split node
- Re-optimize corresponding parameters by EM

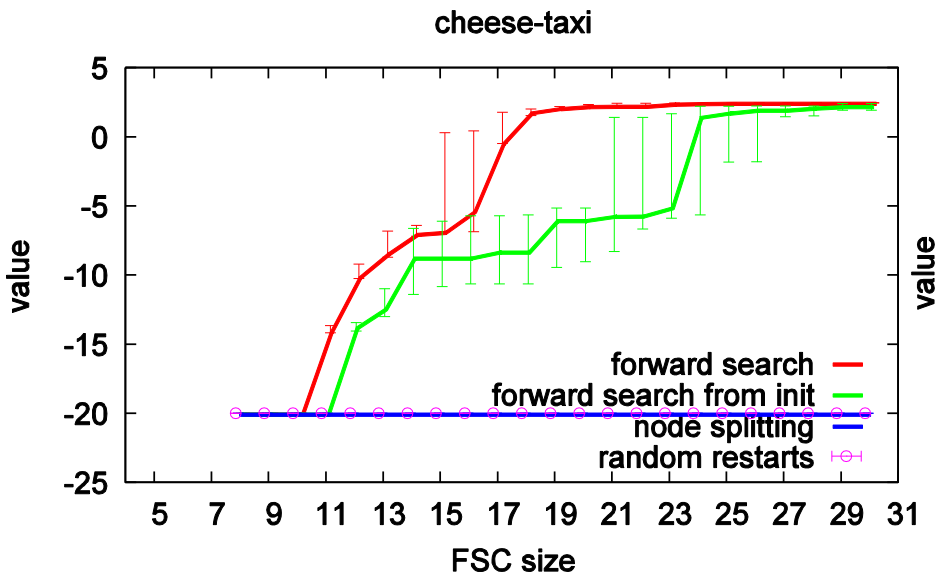


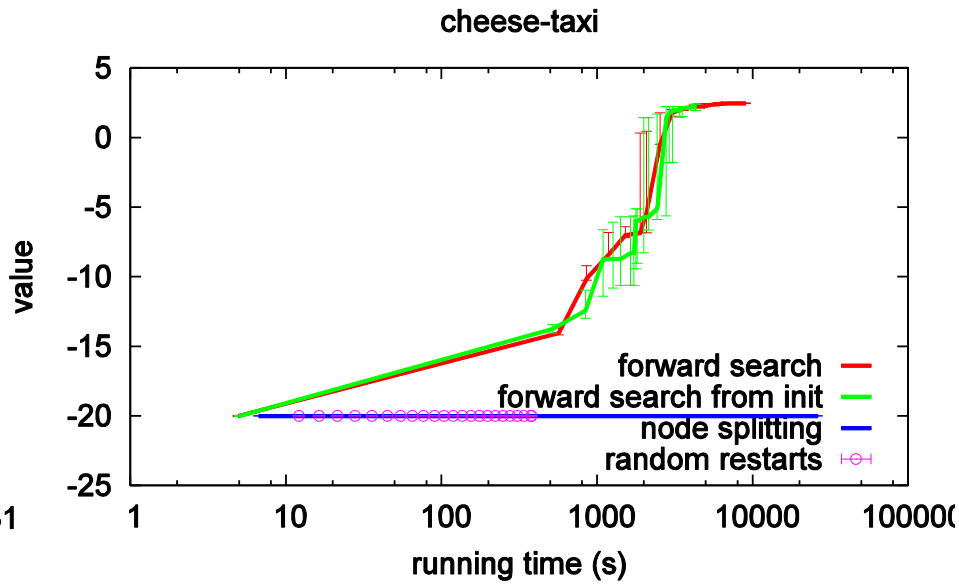
Complexity

- Complexity with respect to
  - Controller size: $|N|$
  - Search depth: d

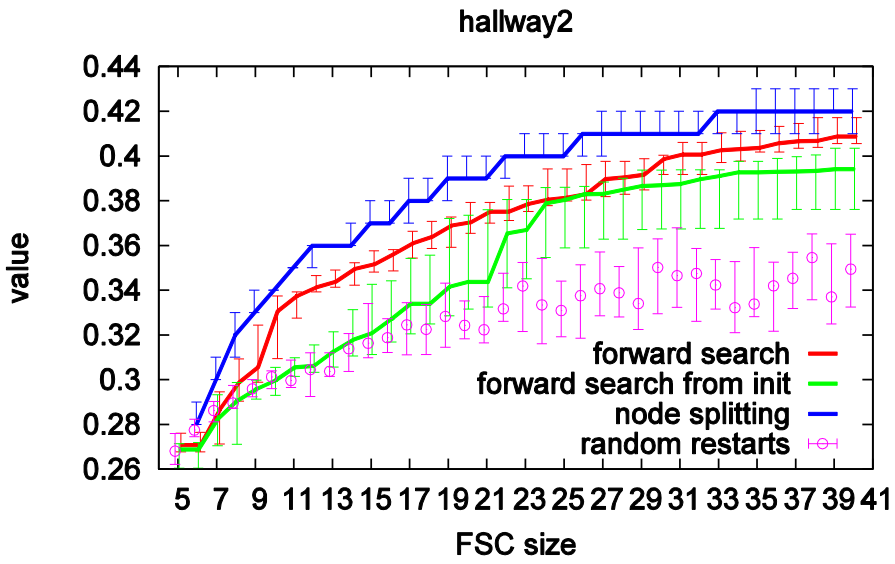
Algorithms	Complexity
EM	$O(\|N\|^2)$
Node Splitting	$O(\|N\|^4)$
Forward Search	$O(\|N\|^3(\|A\|\|O\|)^d)$

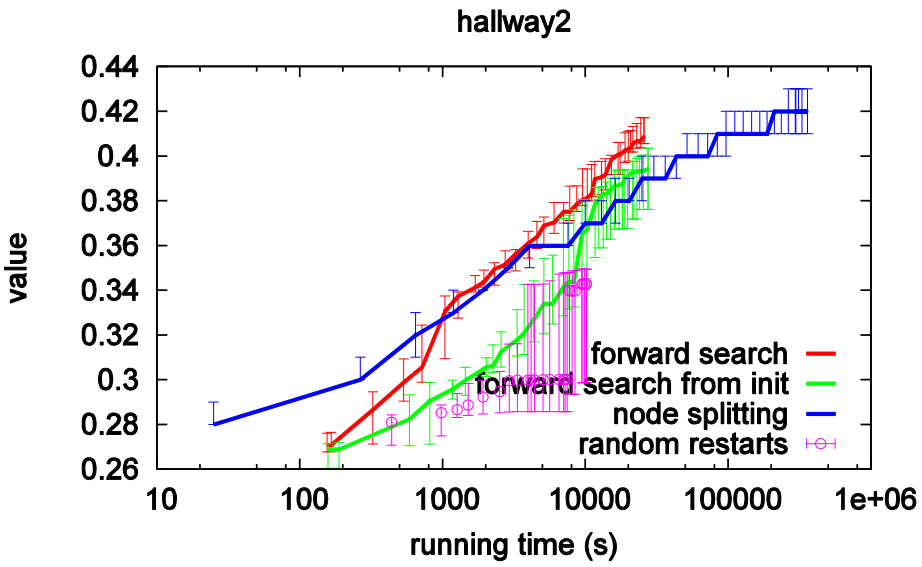
Experiments





Experiments





Experiments

Techniques	cheeseT	heavenH	chainOC	hallway2	hallway	machine
Upper bound	2.48	8.64	157.1	0.88	1.18	66.1
SARSOP (10^5 sec)	2.48(168)	8.64(1720)	157.1(10)	0.44(3295)	1.01(4056)	63.2(1262)
SARSOP	-6.38(40)	0.45(55)	**157.1(10)**	0.11(50)	0.15(49)	35.7(42)
Biased-BPI +escape	2.13(30)	3.50(30)	40.0(30)	0.41(40)	0.94(40)	63.0(30)
QCLP	n.a.	n.a.	n.a.	n.a.	0.72(8)	61.0(6)
BBSLS	n.a.	7.65(?)	n.a.	n.a.	0.80(10)	n.a.
Forward Search	**2.47(19)**	**8.64(16)**	157.1(11)	0.41(40)	0.92(40)	62.6(19)
Node Splitting	-20.0(30)	0.00(30)	157.1(23)	**0.43(40)**	**0.95(40)**	**63.0(16)**

Conclusion

- Analysis:
  - EM ensures one-step look ahead optimality
- Escaping local optima
  - multi-step forward search
  - Node splitting
- Future work:
  - Factored implementation
  - Decentralized POMDPs