An Algebraic Approach to Abstraction in Reinforcement Learning

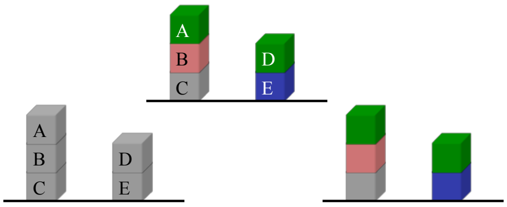
Doctoral Dissertation Defense

Balaraman Ravindran

Advisor: Andrew G. Barto

Committee: Roderic A. Grupen
Sridhar Mahadevan
Neil E. Berthier (Dept. of Psychology)

Abstraction



- Ignore information irrelevant for the task at hand.
- Form simpler representation.

Abstraction

- A key reason that humans are effective problem solvers
  - Learn and plan at a higher level
  - Knowledge transfer
  - c.f. macros, chunks, skills, behaviors, . . .

- **Temporal abstraction** or plan abstraction
- **Spatial abstraction**
- **Combination of the two**

Motivation

- Well studied problem in AI
- Focus of thesis:
  - Decision theoretic setting
    - Markov decision processes
  - General framework
    - Accommodate different notions of abstraction
      - Aggregation, symmetry (Zinkevich and Balch '01, Popplestone and Grupen '00), projections, structured abstractions (Boutilier et al. '94, '95, '01)
  - Formal algebraic framework
    - Group theory, model minimization, operations research
  - Combination of temporal and spatial abstraction
    - Behaviors in a relative frame of reference
      - Efficient knowledge transfer

Outline of Thesis

- Abstraction in decision making
  - Algebraic framework
  - Exploiting symmetry and structure
  - Approximate equivalence
- Abstraction in hierarchical reinforcement learning
  - Hierarchical task decomposition
  - Relativized options
  - Algorithms for dynamic abstraction
    - Choosing transformations
    - Deictic representation

Outline of Thesis

- Abstraction in decision making
  - ➡ Algebraic framework
  - Exploiting symmetry and structure
  - ➡ Approximate equivalence
- Abstraction in hierarchical reinforcement learning
  - Hierarchical task decomposition
  - ➡ Relativized options
  - Algorithms for dynamic abstraction
    - ➡ Choosing transformations
    - Deictic representation

Outline of Talk

- Abstraction in decision making
  - Algebraic framework
  - Approximate equivalence
- Abstraction in hierarchical reinforcement learning
  - Relativized options
  - Algorithms for dynamic abstraction
    - Choosing transformations
  - Summary

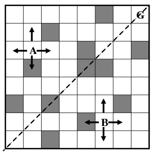
Outline of Talk

- Abstraction in decision making
  - Algebraic framework
    - Markov Decision Processes
    - MDP homomorphisms
    - Some theoretical results
  - Approximate equivalence
- Abstraction in hierarchical reinforcement learning
  - Relativized options
  - Algorithms for dynamic abstraction
    - Choosing transformations
  - Summary

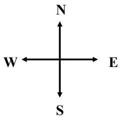
Markov Decision Processes

- MDP, *M,* is the tuple: $M = \langle S, A, \Psi, P, R \rangle$
  - S : set of states.
  - A : set of actions.
  - $\Psi \subseteq S \times A$: set of admissible state-action pairs.
  - $P : \Psi \times S \to [0,1]$: probability of transition.
  - $R : \Psi \to \Re$: expected reward.
- Policy $\pi : S \to A$ (can be stochastic)
- Maximize total expected reward.

Example



$M = \langle S, A, \Psi, P, R \rangle$

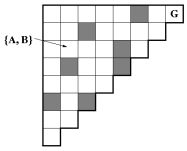


$(\mathrm{A},\mathrm{E}) \equiv (\mathrm{B},\mathrm{N})$

$(\mathrm{A},\mathrm{W}) \equiv (\mathrm{B},\mathrm{S})$

$(\mathrm{A},\mathrm{N}) \equiv (\mathrm{B},\mathrm{E})$

$(\mathrm{A},\mathrm{S}) \equiv (\mathrm{B},\mathrm{W})$



$M' = \langle S', A', \Psi', P', R' \rangle$

Homomorphisms

Group homomorphism

Let G and G' be groups with operations $+$ and $+'$ respectively

$h : G \to G'$ is a group homomorphism iff

$h(x + y) = h(x) +' h(y) \quad \forall x, y \in G$

$$\begin{tikzcd}
G \times G \arrow[r, "+"] \arrow[d, "h \times h"'] & G \arrow[d, "h"] \\
G' \times G' \arrow[r, "+'"] & G'
\end{tikzcd}$$

Homomorphisms (cont.)

Automaton homomorphism

in the autonomous case:

$M = \langle S, \delta \rangle, \quad M' = \langle S', \delta' \rangle$

state set transition function

$h(\delta(s)) = \delta'(h(s))$

$$\begin{tikzcd}
S \arrow[r, "\delta"] \arrow[d, "h"'] & S \arrow[d, "h"'] \\
S' \arrow[r, "\delta'"] & S'
\end{tikzcd}$$

induces equivalence classes $[s]_h$ in S

MDP Homomorphism

MDPs $M=\langle S,A,\Psi,P,R\rangle,\quad M'=\langle S',A',\Psi',P',R'\rangle$

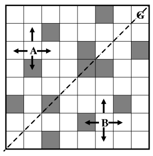
surjection $h:\Psi\rightarrow\Psi'$ defined by $h((s,a))=(f(s),g_s(a))$ where : $f:S\rightarrow S'$, $g_s:A_s\rightarrow A'_{f(s)}$, for all $s\in S$, are surjections such that for all $s,\bar{s}\in S$, and $a\in A_s$:

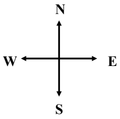
(1) $\quad P'(f(s),g_s(a),f(\bar{s}))=\sum_{t\in[\bar{s}]_f}P(s,a,t)$

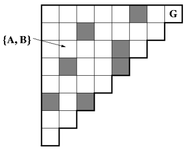
(2) $\quad R'(f(s),g_s(a))=R(s,a)$

$$\begin{tikzcd}
(s,a) \arrow[r, mapsto, "P"] \arrow[d, mapsto, "h"'] & P_{sa} \arrow[d, mapsto, "agg."'] \\
(s',a') \arrow[r, mapsto, "P'"] & P'_{s'a'}
\end{tikzcd}
\qquad
\begin{tikzcd}
(s,a) \arrow[r, mapsto, "R"] \arrow[d, mapsto, "h"'] & \bar{r} \\
(s',a') \arrow[ur, mapsto, "R'"] &
\end{tikzcd}$$

Example







$M=\langle S,A,\Psi,P,R\rangle$

$(\mathrm{A},\mathrm{E})\equiv(\mathrm{B},\mathrm{N})$
$(\mathrm{A},\mathrm{W})\equiv(\mathrm{B},\mathrm{S})$
$(\mathrm{A},\mathrm{N})\equiv(\mathrm{B},\mathrm{E})$
$(\mathrm{A},\mathrm{S})\equiv(\mathrm{B},\mathrm{W})$

$M'=\langle S',A',\Psi',P',R'\rangle$

$$h(A,E)=h(B,N)=(\{A,B\},E)$$

State dependent action recoding

Some Theoretical Results

[generalizing those of Dean and Givan, 1997]

- Optimal Value equivalence:
  If $h(s,a)=(s',a')$ then $Q^*(s,a)=Q^*(s',a')$.
- Corollary:
  If $h(s_1,a_1)=h(s_2,a_2)$ then $Q^*(s_1,a_1)=Q^*(s_2,a_2)$.

Theorem: If M' is a homomorphic image of M, then a policy optimal in M' induces an optimal policy in M.

- Solve homomorphic image and *lift* the policy to the original MDP.

Model Minimization

- Finding reduced models that preserve some aspects of the original model
- Various modeling paradigms
  - Finite State Automata (Hartmanis and Stearns ' 66)
    - Transition Behavior
  - Model Checking (Emerson and Sistla ' 96, Lee and Yannakakis ' 92)
    - Correctness of system models
  - Markov Chains (Kemeny and Snell ' 60)
    - Steady state distribution
  - MDPs (Dean and Givan ' 97, Ravindran and Barto ' 02)
    - Optimal solutions

MDP Minimization

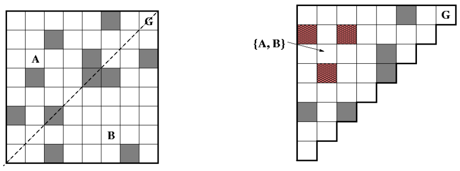
- In general, NP-hard
  - Polynomial time algorithm for computing homomorphic image, under certain assumptions
    - Extends Dean and Givan ' 97, Lee and Yannakakis ' 92
- State dependent action recoding
  - Greater reduction in problem size
  - Model symmetries
    - Reflections, rotations, permutations

Outline of Talk

- Abstraction in decision making
  - Algebraic framework
  - ➡ Approximate equivalence
    - Approximate homomorphisms
    - Error bounds
    - Bounded parameter approximations
- Abstraction in hierarchical reinforcement learning
  - Relativized options
  - Algorithms for dynamic abstraction
    - Choosing transformations
  - Summary

Approximate Notions of Equivalence

- Complete and exact equivalence often do not exist.
- Approximate equivalence.
  - "Equivalent" state-action pairs have nearly same behavior.



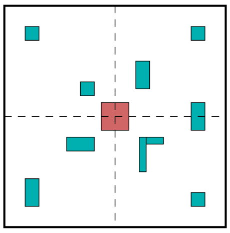
Approximate Homomorphisms

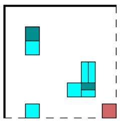
- Use averages
- Relax homomorphism criteria:
  - $P'(f(s), g_s(a), f(\bar{s})) = \sum_{t\in[\bar{s}]_f} P(s,a,t)$
  - Compute $\sum_{t\in[\bar{s}]_f} P(s,a,t)$ for all (s, a)

$$\boxed{P'(f(s), g_s(a), f(\bar{s})) = \frac{1}{|[(s,a)]_h|}\sum_{(q,b)\in[(s,a)]_h}\ \sum_{t\in[\bar{s}]_f} P(q,b,t)}$$

  - Similar computation for the reward function.

Example





Approximate Image

Task is to reach red goal area.

Error Bound

- Approximate homomorphism between arbitrarily different MDPs!
- Useful when loss in performance is acceptable.
- Bound the maximum difference in optimal value function in *M* and the value of the lifted optimal policy.
  - Specializes Whitt '78.
  - Function of maximum difference in the probabilities and rewards that are averaged.

Error bound (cont.)

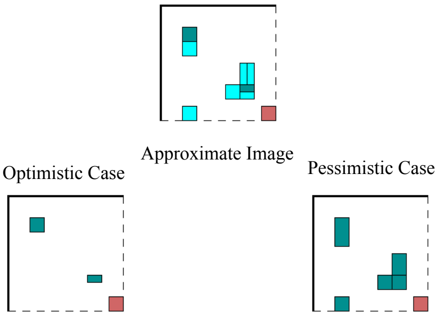
- K_p – maximum difference between $P'(f(s), g_s(a), f(\bar{s}))$ and $\sum_{t\in[\bar{s}]_f} P(s,a,t)$
- K_r – corresponding difference in reward
- Δ – the range of the reward function
- γ – the discount factor, $0 \le \gamma < 1$

$$\boxed{\left\| V^* - V'^* \right\|_{\max} \le \frac{2}{1-\gamma}\left(K_r + \frac{\gamma}{1-\gamma}\Delta\frac{K_p}{2}\right)}$$

Bounded Parameter Approximation

- Model as a map onto a *Bounded-parameter MDP* (Givan, Leach and Dean '00)
  - Transition probabilities and rewards given by bounded intervals
  - Upper and lower bounds on optimal values of states
  - Loose bounds

Example Revisited



Outline of Talk

- Abstraction in decision making
  - Algebraic framework
  - Approximate equivalence
- Abstraction in hierarchical reinforcement learning
  - Relativized options
    - Semi-Markov decision processes
    - Options framework
    - Relativized options
  - Algorithms for dynamic abstraction
    - Choosing transformations
  - Summary

(discrete-time) semi-Markov Decision Process

- SMDP, *M,* is the tuple: $M = \langle S, A, \Psi, P, R \rangle$
  - S : set of states.
  - A : set of actions.
  - $\Psi \subseteq S \times A$: set of admissible state-action pairs.
  - $P : \Psi \times S \times \mathrm{N} \rightarrow [0,1]$: transition probabilities.
  - $R : \Psi \times \mathrm{N} \rightarrow \Re$: expected reward.
- Policy (stationary, stochastic): $\pi : \Psi \rightarrow [0,1]$
- Maximize expected return.
- Generalize MDP homomorphism.

Hierarchical Reinforcement Learning

Options (Sutton, Precup, & Singh, 1999): **A generalization of actions to include temporally-extended courses of action**

An option is a triple $o =< I, \pi_o, \beta >$
- $I \subseteq S$ is the set of states in which o may be started
- $\pi_o : \Psi \rightarrow [0,1]$ is the (stochastic) policy followed during o
- $\beta : S \rightarrow [0,1]$ is the probability of terminating in each state

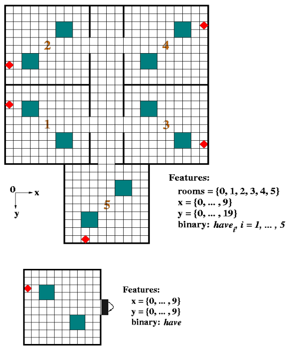
Example: robot docking

I **:** all states in which charger is in sight

?,??: pre-defined controller

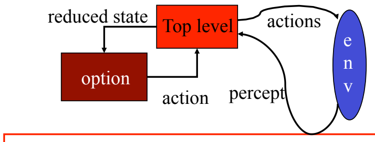
? : terminate when docked or charger not visible

Sub-goal Options



- Task is to collect all objects in the world
- 5 options – one for each room.
- Markov, subgoal options
- Implicitly define option policy
- Employ option specific abstraction

Relativized Options



Relativized option:

$$O = \langle h, M_O, I, \beta \rangle$$

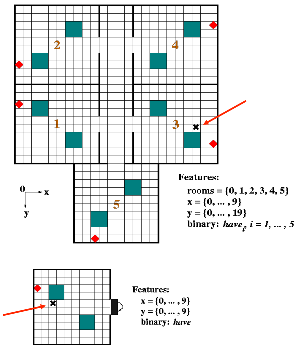
h : Option homomorphism.

M_O: Option SMDP. (Image of h.)

$I \subseteq S$: Initiation set.

$\beta : S_O \rightarrow [0,1]$: Termination criterion.

Rooms world task



- Task is to collect all objects in the world
- 5 options – one for each room
- Single relativized option – *get-object-exit-room*
- Partial homomorphism
- Especially useful when learning option policy
  - Speed up.
  - Knowledge transfer.

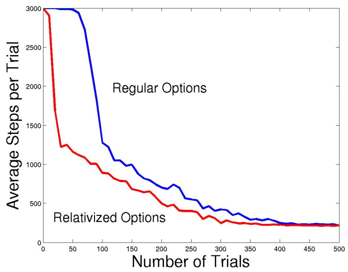
Experimental Setup

- Regular Agent
  - 5 options, one for each room
  - Option reward of +1 on exiting room with object
- Relativized Agent
  - 1 relativized option, known homomorphism
  - Same option reward
- Global reward of +1 on completing task
- Actions fail with probability 0.1

Learning Algorithm

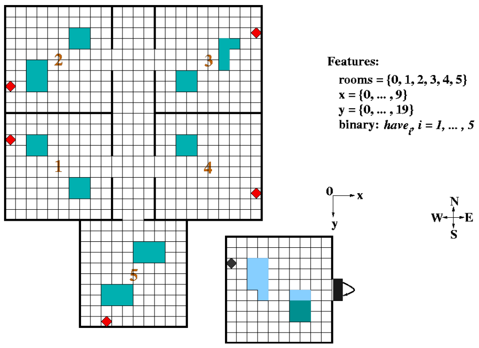
- Hierarchical SMDP Q-learning (Dietterich ' 00b)
  - Q-learning at the lowest level (Watkins ' 89)
  - SMDP Q-learning at the higher levels (Bradtke and Duff ' 95)
- Simultaneous learning at all levels
  - Converges to recursively optimal policy
    - Using results from Dietterich ' 00a

Results

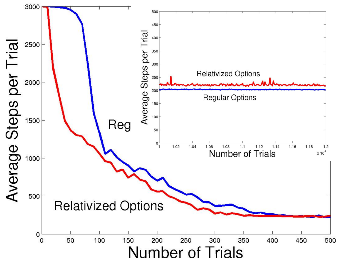


- Average over 100 runs

Asymmetric Testbed


2
3
1
4
5
Features:
rooms = {0, 1, 2, 3, 4, 5}
x = {0, ... , 9}
y = {0, ... , 19}
binary: *have*$_i$, *i = 1, ... , 5*
0
x
y
N
W
E
S

Results – Asymmetric Testbed



- Still significant speed up in initial learning
- Asymptotic performance slightly worse

Outline of Talk

- Abstraction in decision making
  - Algebraic framework
  - Approximate equivalence
- Abstraction in hierarchical reinforcement learning
  - Relativized options
  - Algorithms for dynamic abstraction
    - ➡ Choosing transformations
  - Summary

Choosing Transformations

Motivation

- Relax prior knowledge requirement
  - Unknown homomorphism
- Option SMDP and policy can be viewed as a *policy schema* (Schmidt ' 75, Arbib ' 95)
  - Template of a policy
  - Acquire schema in a prototypical setting
  - Learn bindings of sensory inputs and actions to schema
- Assume set of possible bindings available

Choosing Transformations

Problem Formulation

- Given:
  - M_O, I, β of a relativized option
  - H , a family of transformations
- Identify the option homomorphism h
- Formulate as a parameter estimation problem
  - One parameter, takes values from H
  - Samples: $\langle s_1, a_1, s_2, a_2, \cdots \rangle$
  - Bayesian learning

Choosing Transformations

Algorithm

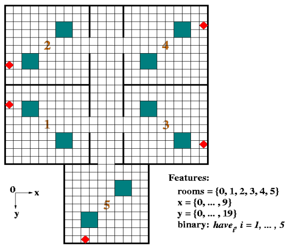
- Assume uniform prior: $p_0(h, \bar{s})$
- Experience: $\langle s_n, a_n, s_{n+1} \rangle$

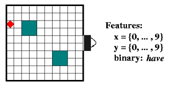
$$P(\langle s_n, a_n, s_{n+1} \rangle | h, \bar{s}) = P_O(f(s_n), g_{s_n}(a_n), f(s_{n+1}))$$

- Update Posteriors:

$$p_n(h, \bar{s}) = \frac{P_O(f(s_n), g_{s_n}(a_n), f(s_{n+1})) \cdot p_{n-1}(h, \bar{s})}{\text{Normalizing Factor}}$$

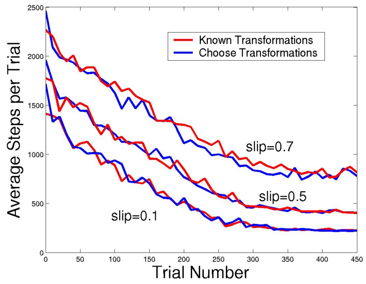
Rooms world task





- Train in room 1
- 8 candidate transformations
  - Reflections about x and y axes and the x=y and x=-y lines
  - Rotations by integer multiples of 90 degrees

Results – Speed of convergence



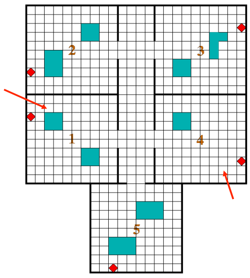
- Not much of a difference since the task is too easy
- Correct transformation identified in 15 iterations

Choosing Transformations

Approximate Equivalence

- More complex domains
- Problem with Bayesian update
  - Use prototypical room as option schema
  - Susceptible to incorrect samples
- Use a heuristic lower bound

Example



Features:

rooms = {0, 1, 2, 3, 4, 5}
x = {0, … , 9}
y = {0, … , 19}
binary: *have*$_i$, *i* = *1*, … , *5*

$$p_n(h,\bar{s}) = \frac{P_O\big(f(s_n), g_{s_n}(a_n), f(s_{n+1})\big)\cdot p_{n-1}(h,\bar{s})}{\text{Normalizing Factor}}$$

Choosing Transformations

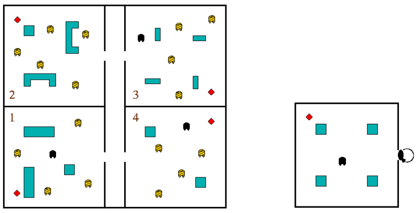
Heuristic Update Rule

- Use a heuristic update rule:

$$\boxed{w_n(h,\bar{s}) = \frac{\bar{P}\big(f(s_n), g_{s_n}(a_n), f(s_{n+1})\big)\cdot w_{n-1}(h,\bar{s})}{\text{Normalizing Factor}}}$$

where, $\bar{P}(s,a,s') = \max\big(\nu, P_O(s,a,s')\big)$
and ν is a small positive constant.

Complex Game World



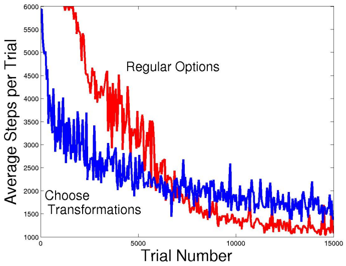
- Gather all 4 diamonds in the world
- 25×10^{55} states
- 40 transformations
  - 8 spatial transformations combined with 5 projections

Experimental Setup

- Regular agent
  - 4 sub-goal options
- Relativized agent
  - Uses option MDP shown earlier
  - Chooses from 40 transformations
- Room 2 has no right transformation
- Hierarchical SMDP Q-learning

Results

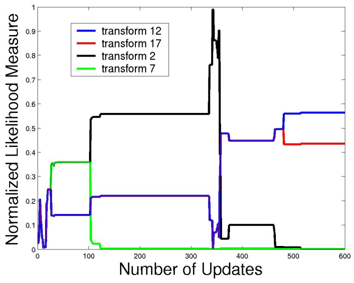
Speed of Convergence



- Learning the policy is more difficult than learning the correct transformation!

Results

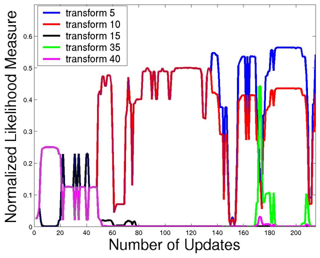
Transformation Weights in Room 4



- Transformation 12 eventually converges to 1

Results

Transformation Weights in Room 2



- Weights oscillate a lot
- Some transformation dominates eventually
  - Changes from one run to another

Choosing Transformations

- Related work
  - Multiple forward models (Haruno et al. ’01, Doya et al. ‘02)
  - Dynamic control models (Coelho and Grupen ’98)
  - Variably bound controllers (Huber and Grupen ’99)
- Representations can be designed to implicitly perform transformations
  - Formalizes such representations
  - E.g. Deictic representations

Summary of Contributions

- Developed an abstraction framework for MDPs
  - Introduced MDP homomorphisms
    - State dependent action recoding
  - Theoretical results
- Approximate homomorphisms
  - Bound maximum loss
  - Upper and lower bound performance

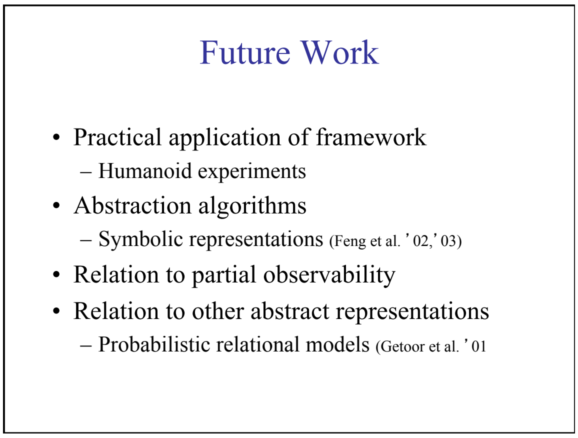
Summary of Contributions

(cont.)

- Abstraction in hierarchical systems
  - Relativized options
    - An option defined in a relative frame of reference
    - Uses partial homomorphisms
  - Policy schema
    - Policy template
  - Bayesian algorithm for choosing the right bindings
    - Heuristic modification for approximate equivalence
    - Complex game domain

Other Contributions

- Exploiting structure and symmetry
  - Structured morphisms
  - Symmetry groups
    - Reflections, rotations and permutations
  - Polynomial time algorithm
- Hierarchical decomposition framework
  - Based on SMDP homomorphisms
  - Relation to safe state abstraction (Dietterich ’00a)
- Deictic option schema
  - Representation based on pointers (Agre ’88)
  - Modification of Bayesian algorithm



Future Work

- Practical application of framework
  - Humanoid experiments
- Abstraction algorithms
  - Symbolic representations (Feng et al. ' 02,' 03)
- Relation to partial observability
- Relation to other abstract representations
  - Probabilistic relational models (Getoor et al. ' 01