Action Selection in Bayesian Reinforcement Learning

Tao Wang

UNIVERSITY OF ALBERTA

ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

UNIVERSITY OF ALBERTA

Background Perspective

- Be *Bayesian* about reinforcement learning
- Ideal representation of uncertainty for action selection

Why are Bayesian approaches not prevalent in RL?

- Computational barriers

My Work

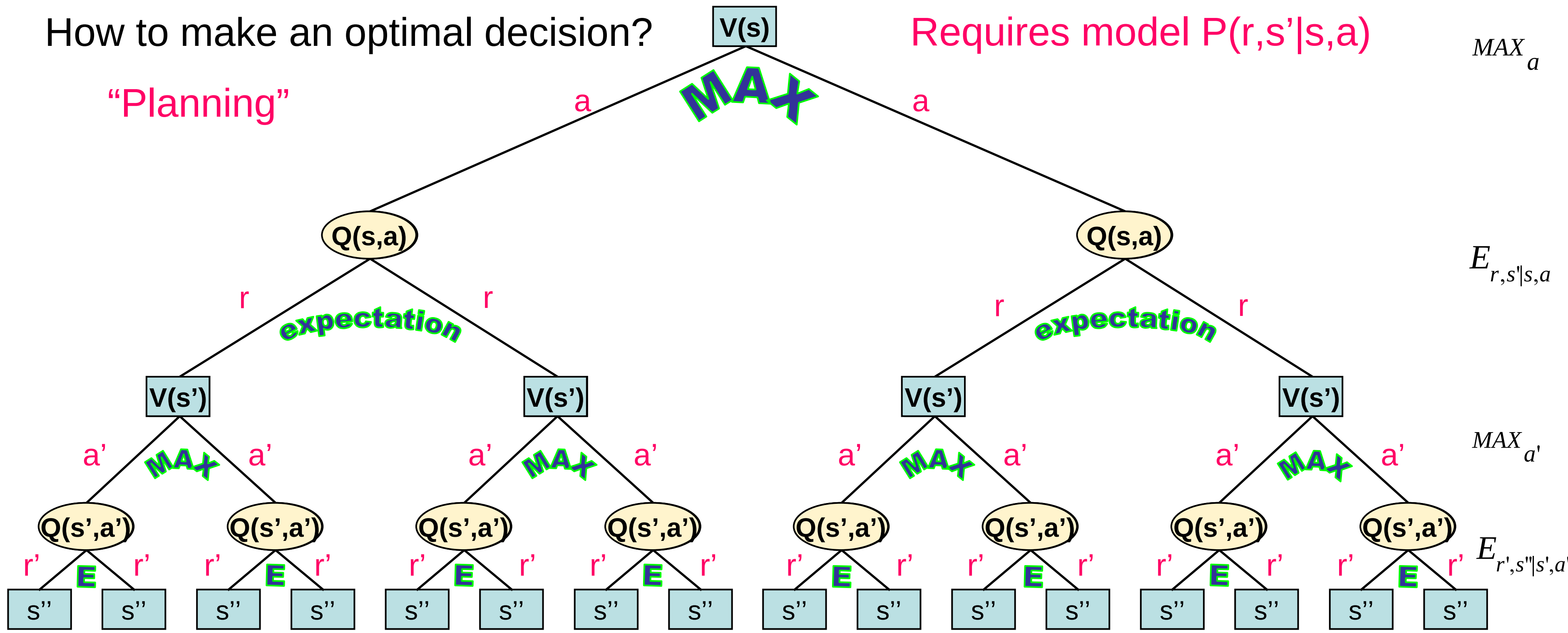
- Practical algorithms for approximating Bayes optimal decision making
- Analogy to *game-tree search*
  - on-line lookahead computation
  - global value function approximation
    (but here expecti-max vs. mini-max)
- Two key parts:
  - build a lookahead tree (ICML 2005)
  - approximate leaf values (AAAI 2006)

Sequential Decision Making

How to make an optimal decision?

Requires model P(r,s’|s,a)

“Planning”



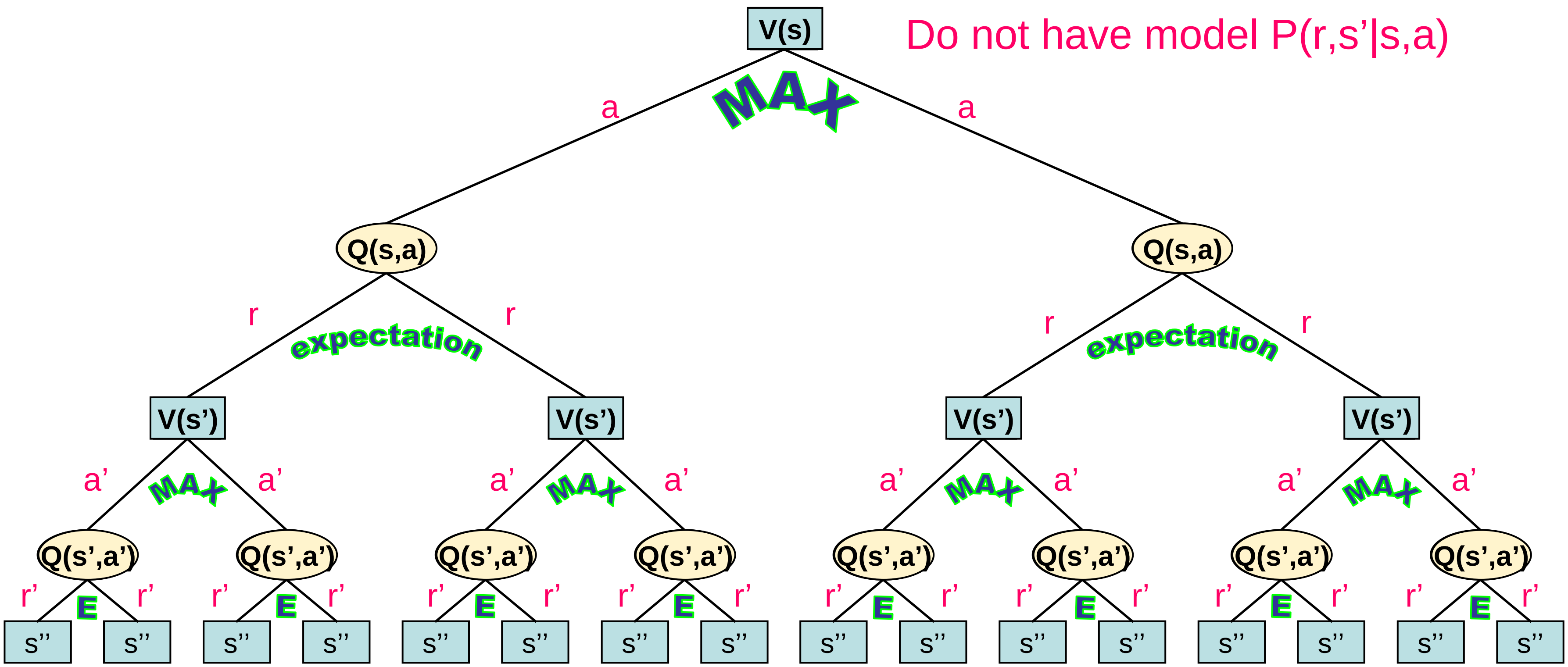
This is: *finite horizon, finite action, finite reward* case

General case: *Fixed point equations:* $V(s) = \sup_a Q(s,a) \quad Q(s,a) = E_{r,s'|s,a}[r + \gamma V(s')]$

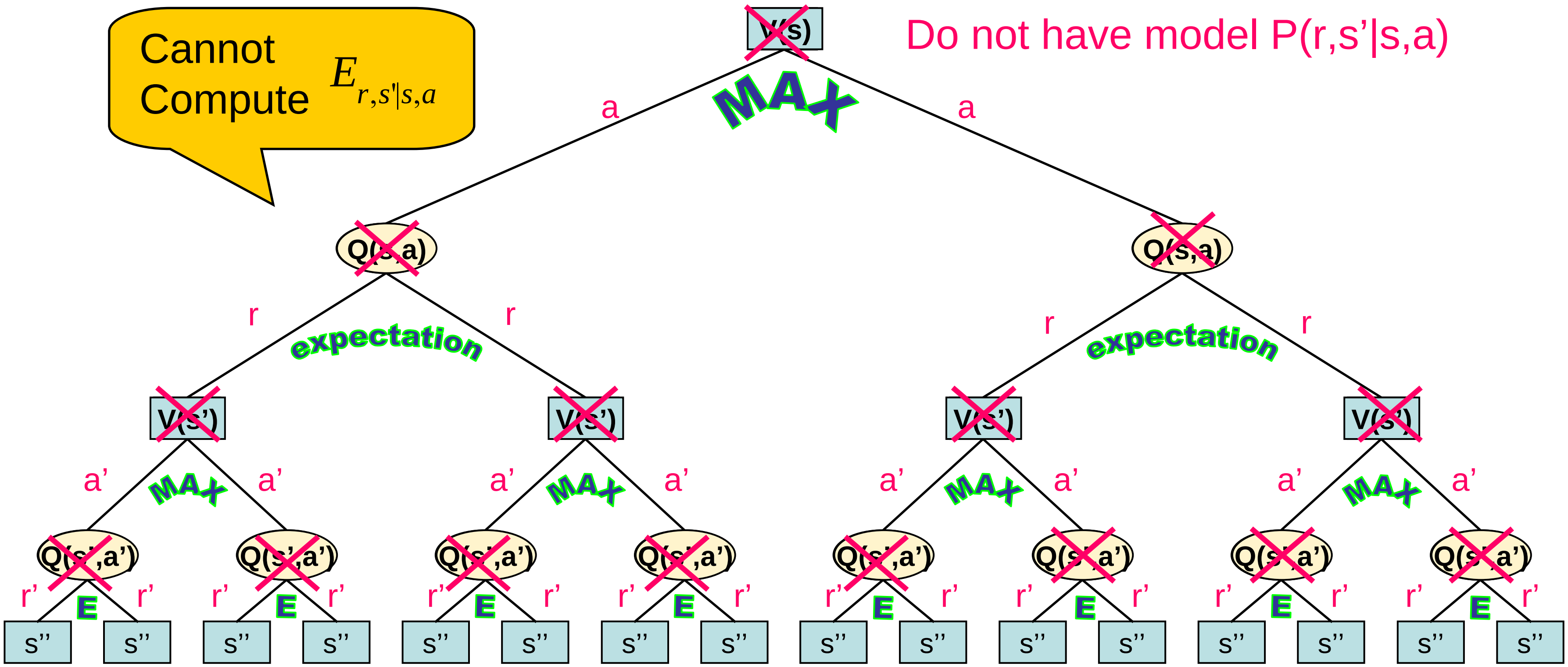
UNIVERSITY OF ALBERTA

Reinforcement Learning

Do not have model P(r,s'|s,a)



Reinforcement Learning



UNIVERSITY OF ALBERTA

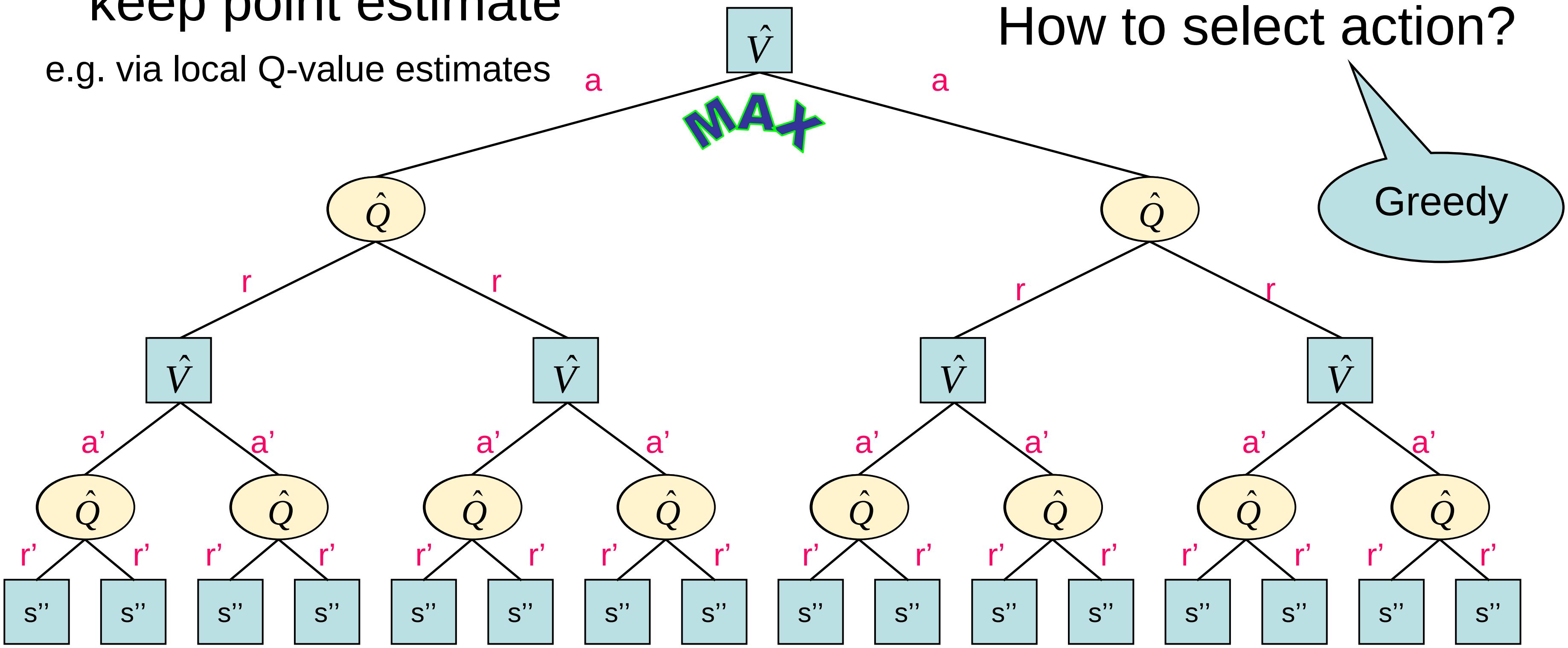
Reinforcement Learning

Standard approach:
keep point estimate

e.g. via local Q-value estimates

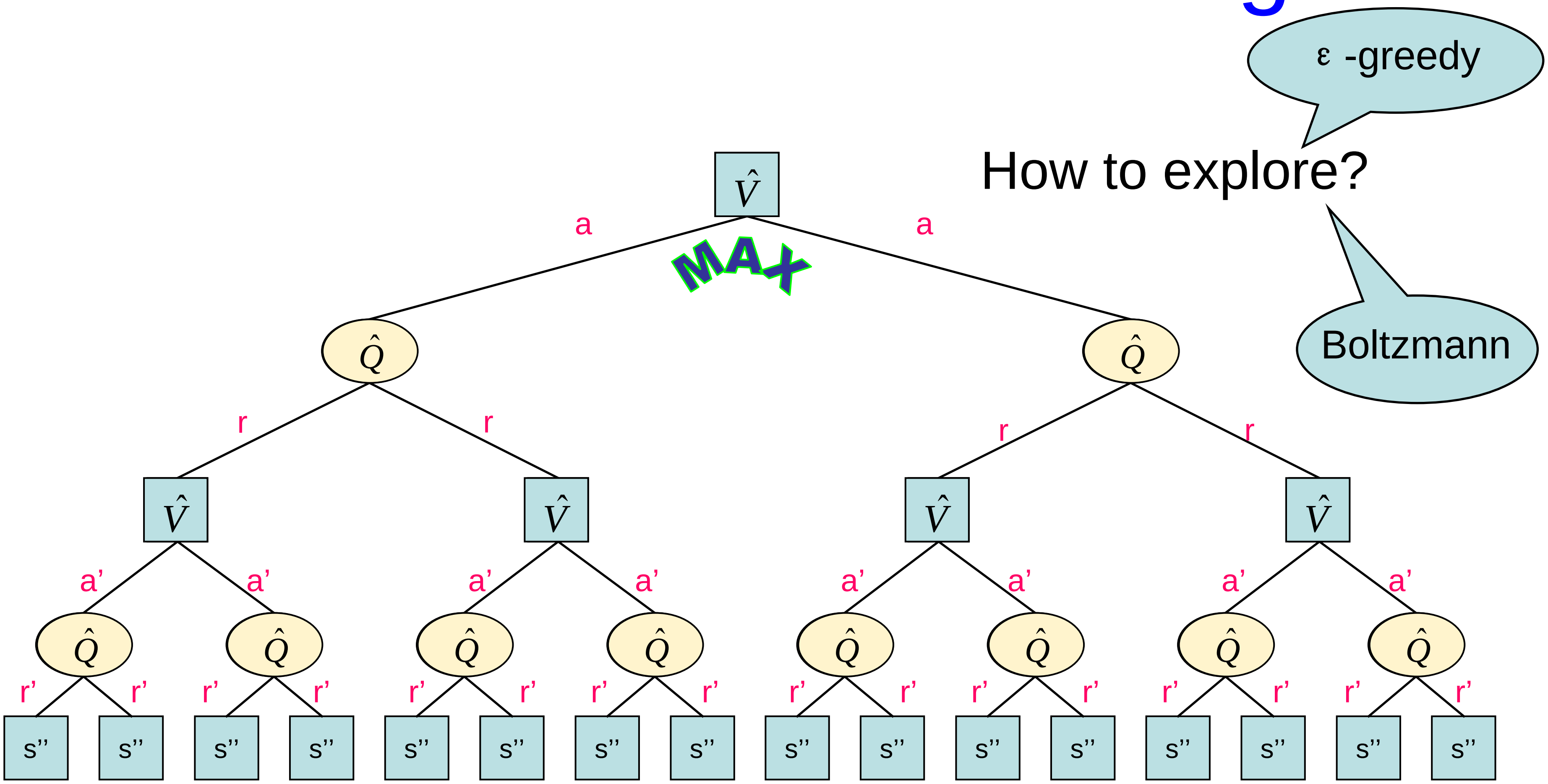
Do not have model P(r,s’|s,a)

How to select action?



Problem: greedy does not explore

Reinforcement Learning



Problem: do not account for uncertainty in estimates

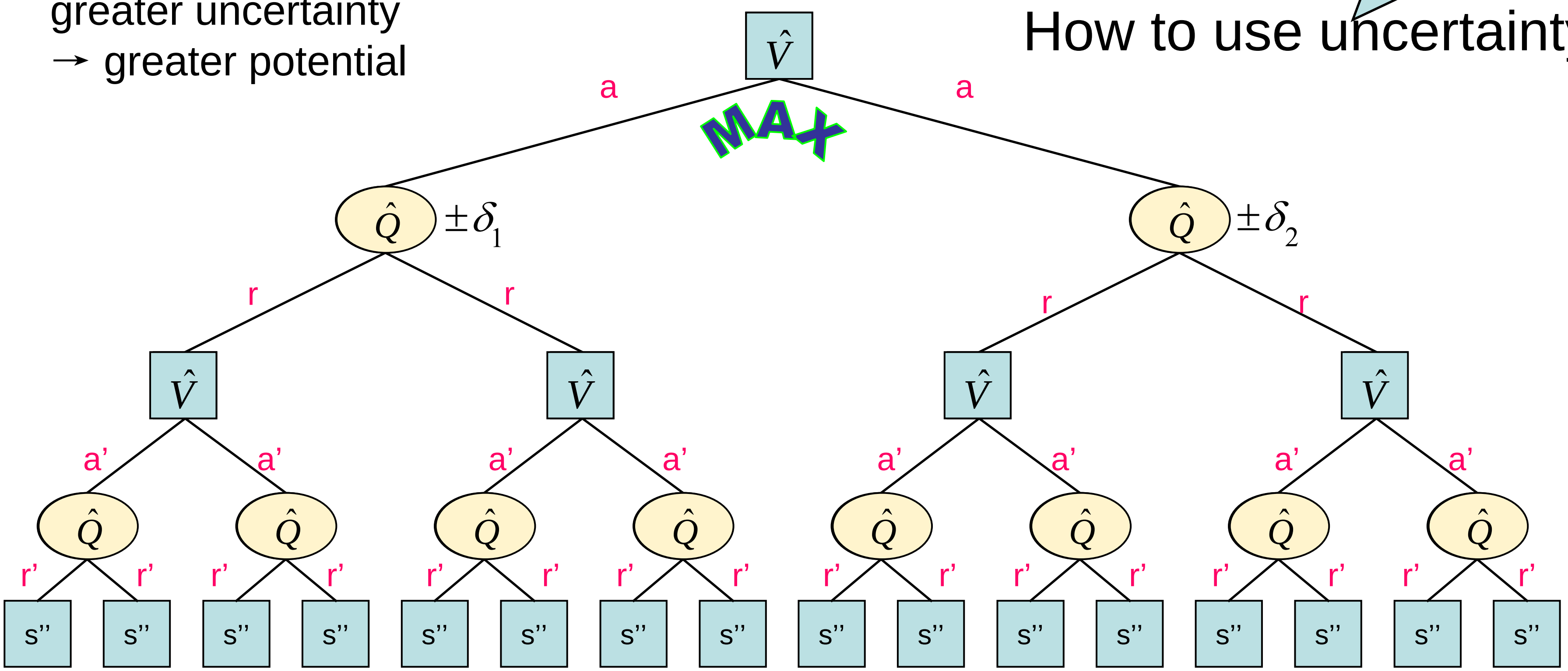
Reinforcement Learning

Interval estimation

Intuition:

greater uncertainty
→ greater potential

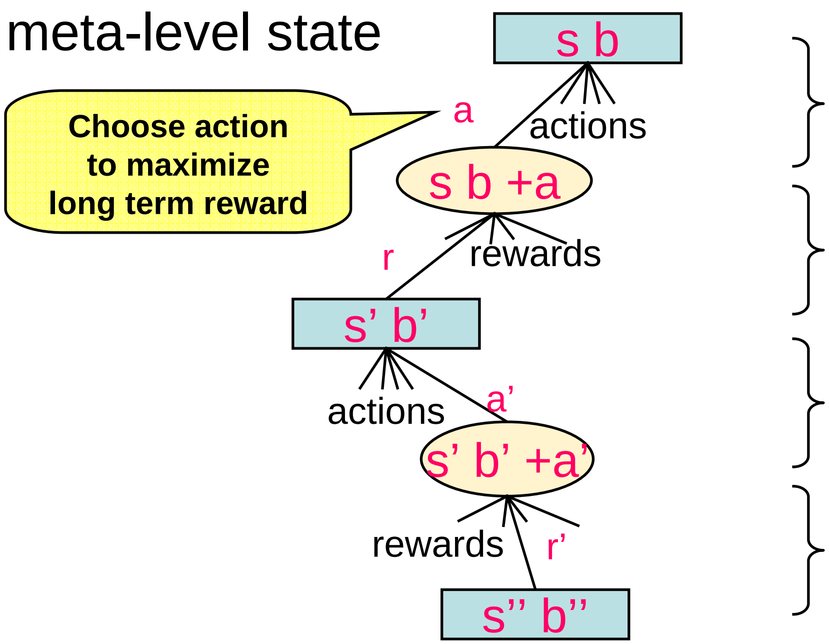
How to use uncertainty?



Problem: δ 's computed myopically: doesn't consider horizon

Bayesian Reinforcement Learning

Prior *P(θ)* on model *P(rs' |sa, θ)* Belief state *b=P(θ)*



Meta-level MDP

decision a	
outcome r, s', b'	Meta-level Model P(r,s'b'|s b,a)
decision a'	
outcome r', s'', b''	Meta-level Model P(r',s''b''|s' b',a')

Have a model for meta-level transitions!

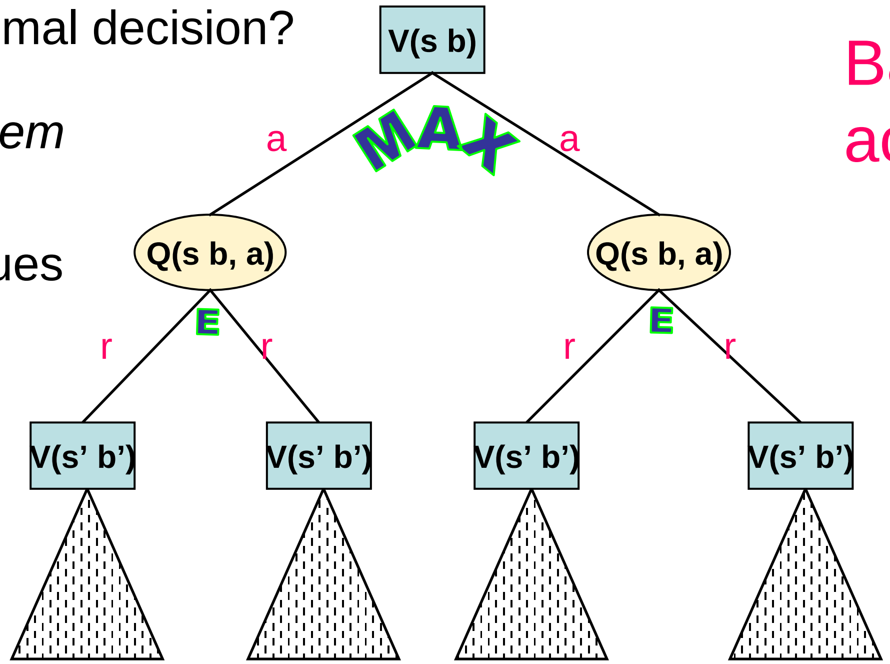
- based on posterior update and expectations over base-level MDPs

Bayesian RL Decision Making

How to make an optimal decision?

Solve planning problem in meta-level MDP:

- Optimal Q,V values



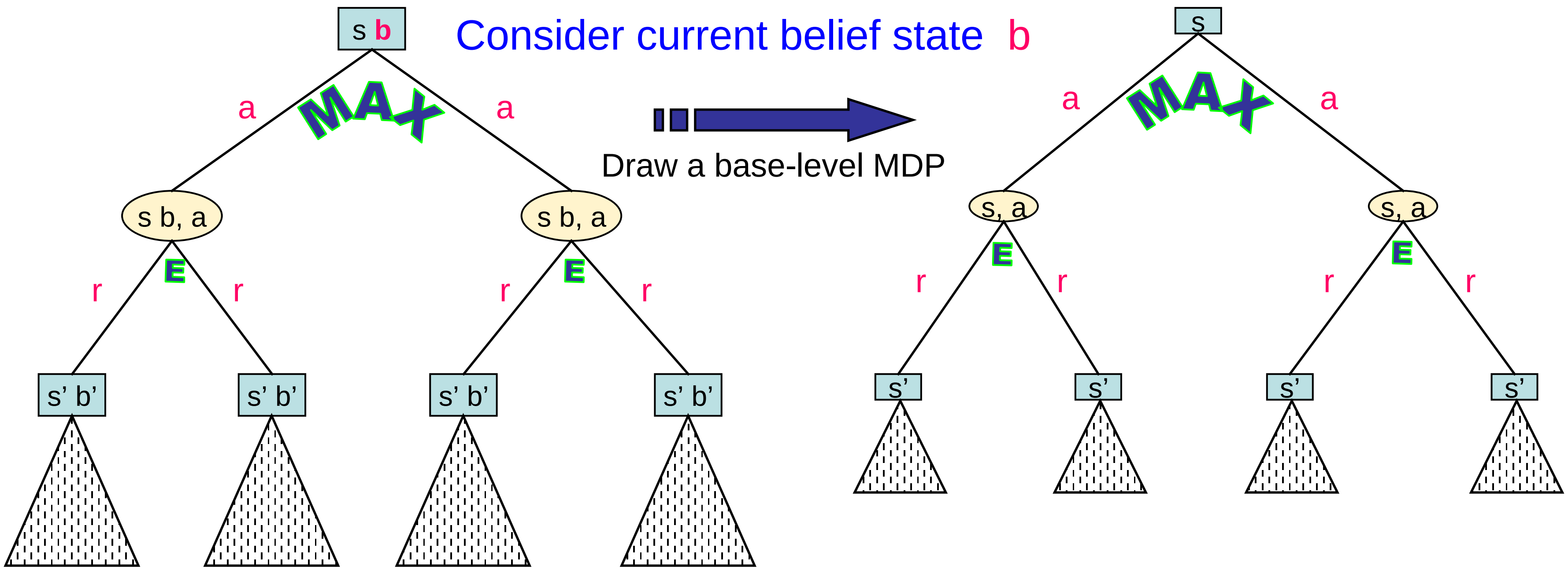
Bayes optimal action selection

Problem: meta-level MDP *much larger* than base-level MDP

Impractical

Bayesian RL Decision Making

Current approximation strategies:



Greedy approach:

current b → mean base-level MDP model

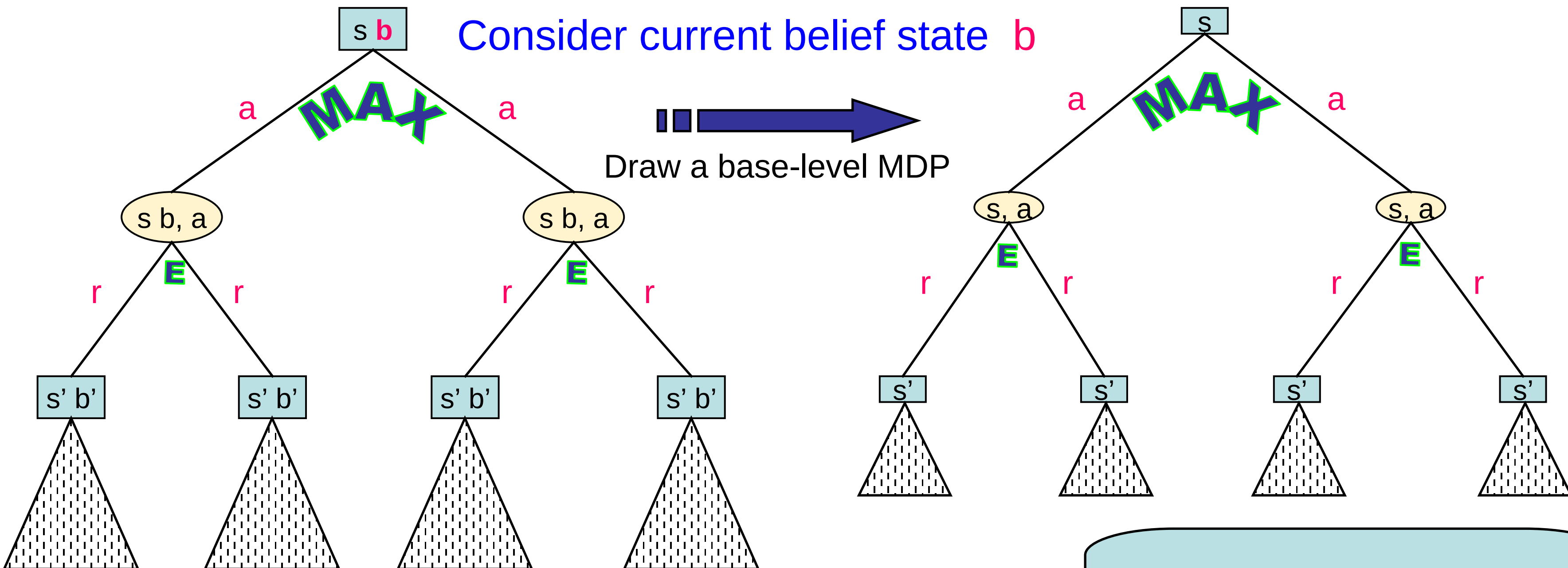
→ point estimate for Q, V

→ choose greedy action

But doesn't consider uncertainty

Bayesian RL Decision Making

Current approximation strategies:



☺ Exploration is based on uncertainty

Thompson approach:

current b → sample a base-level MDP model

→ point estimate for Q, V

(Choose action proportional to probability it is max Q)

But still myopic

Part 1
Bayesian Sparse Sampling

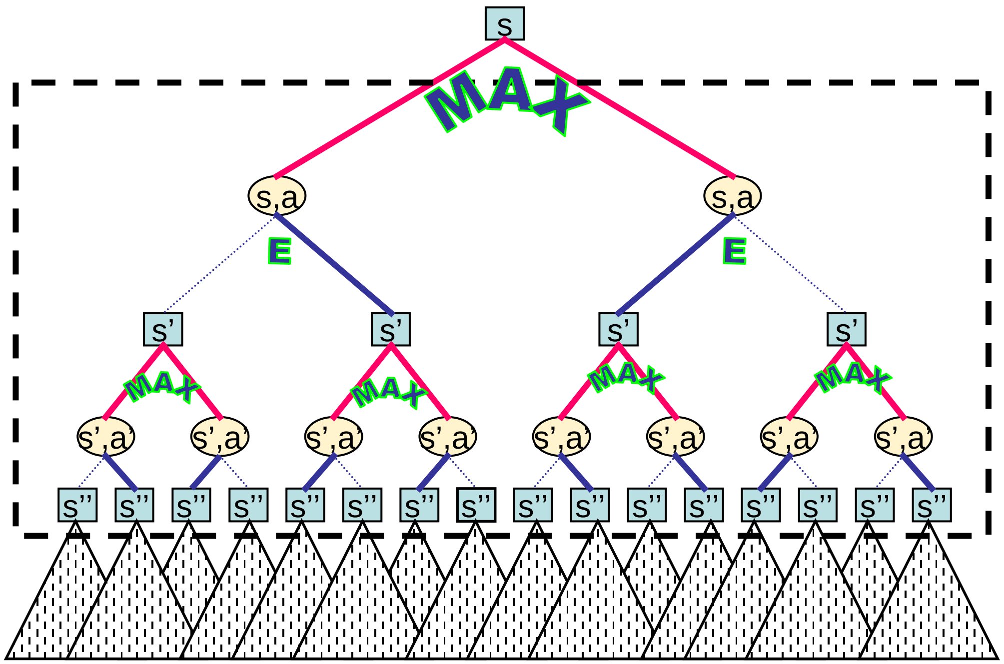
Bayesian Sparse Sampling for On-line Reward Optimization

Tao Wang   Daniel Lizotte   Michael Bowling   Dale Schuurmans

ICML 2005

UNIVERSITY OF ALBERTA

Sparse Sampling



(Kearns, Mansour, Ng 2001)

Approximate values

Enumerate action choices

Subsample action outcomes

Bound depth

Back up approx values

+ Chooses approximately optimal action with high probability (if depth, sampling large enough)

– Achieving guarantees too expensive

+ But can control depth, sampling

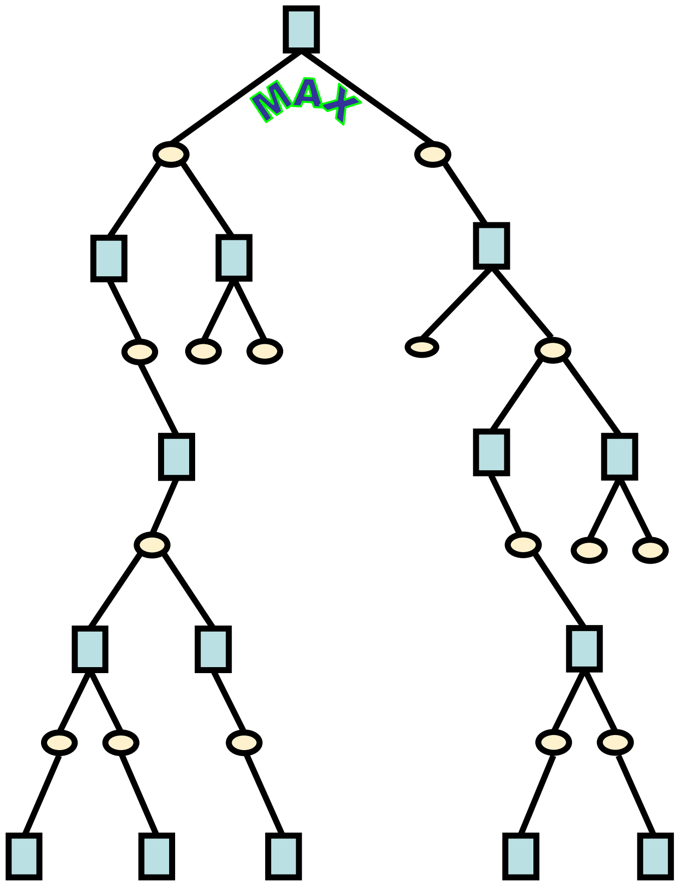
Bayesian Sparse Sampling
Observation 1

- Do not need to enumerate actions in a Bayesian setting
  - Given random variables $Q_1, \ldots, Q_K$
  - and a prior $P(Q_1, \ldots, Q_K)$
  - Can approximate $\max(Q_1, \ldots, Q_K)$
  - Without observing every variable

(Stop when posterior probability of a significantly better Q-value is small)

Bayesian Sparse Sampling

Observation 2



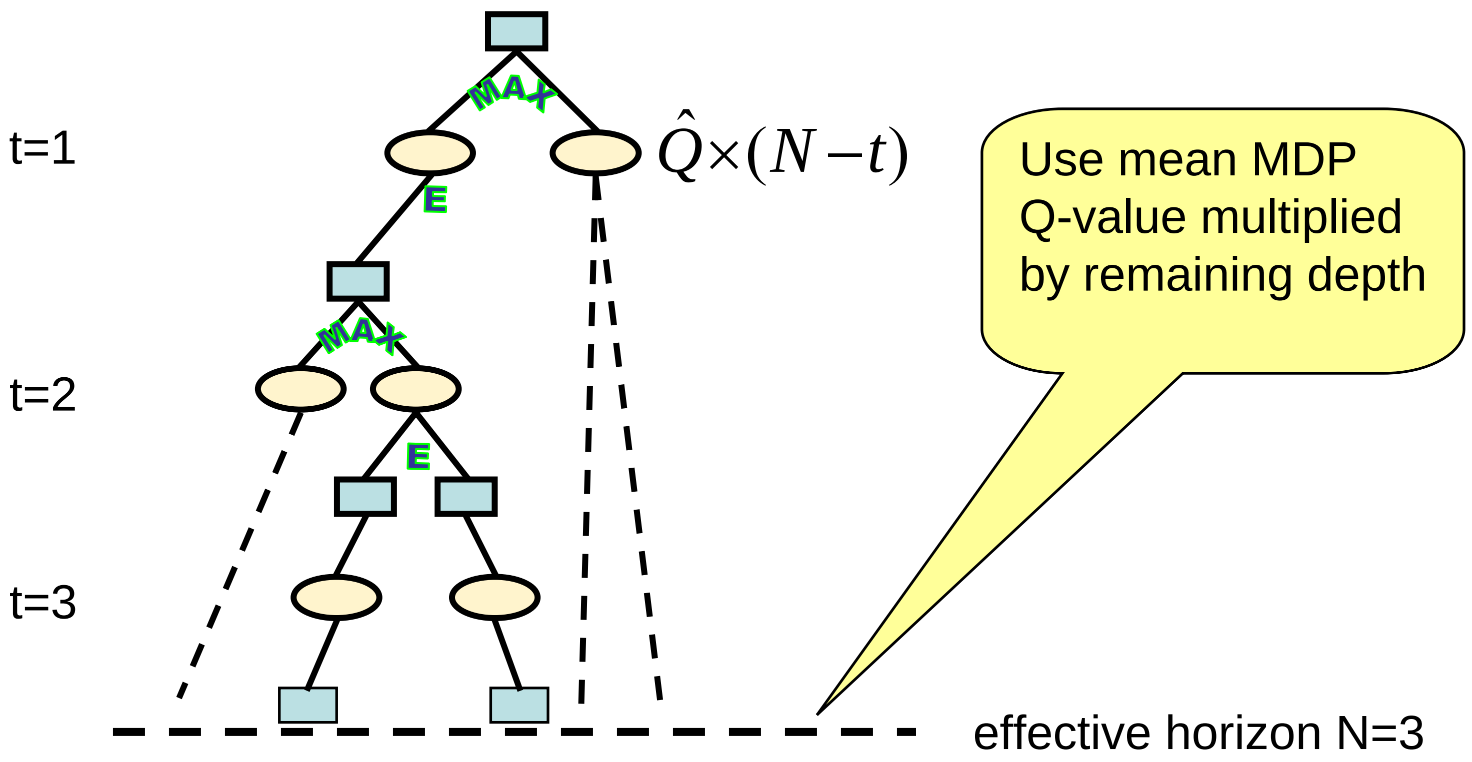
- Action value estimates are not equally important
  - Need better Q value estimates for some actions but not all
  - Preferentially expand tree under actions that might be optimal

Biased tree growth

Use Thompson sampling to select actions to expand

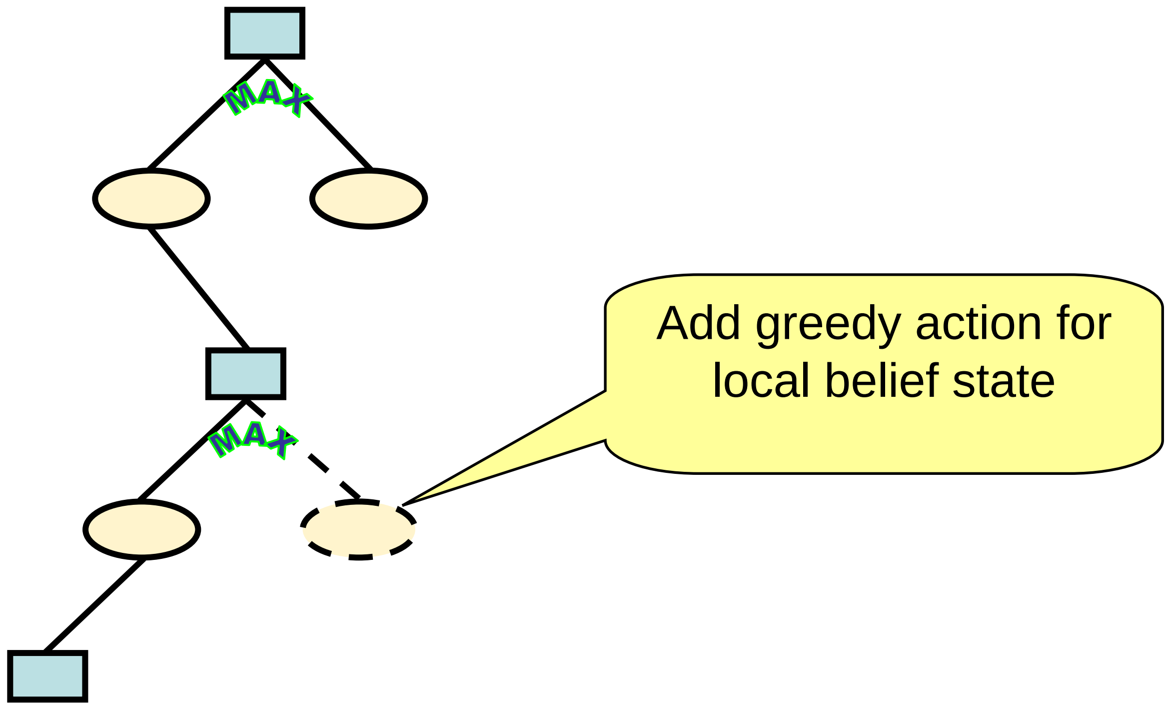
Bayesian Sparse Sampling
Observation 3

Correct leaf value estimates to same depth



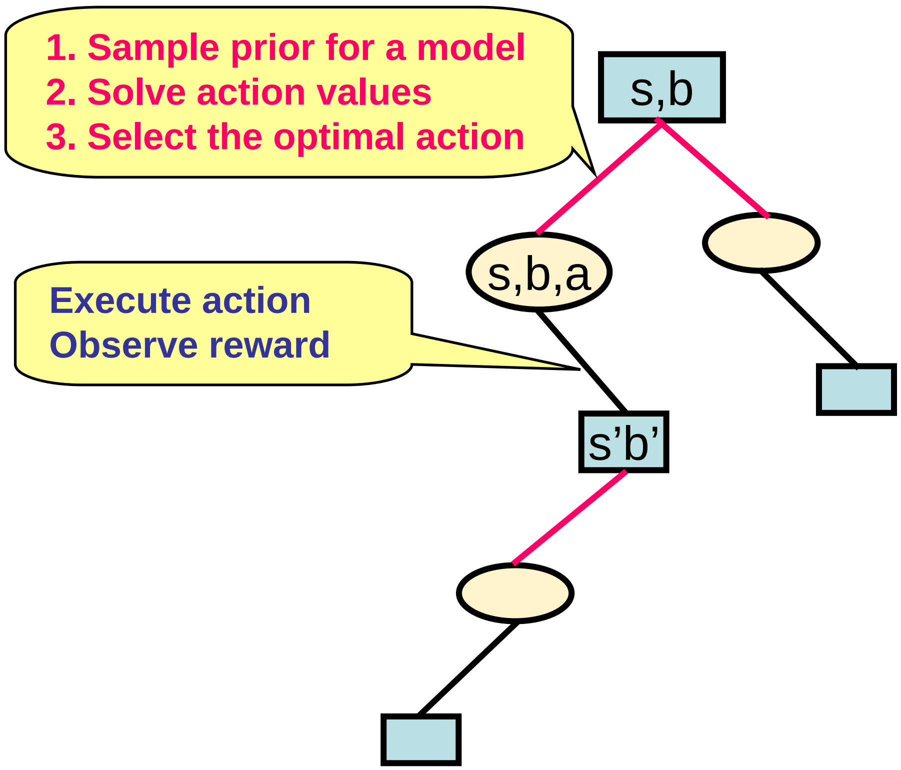
Bayesian Sparse Sampling
Observation 4

Include greedy action at decision nodes (if not sampled)



Bayesian Sparse Sampling

Tree growing procedure

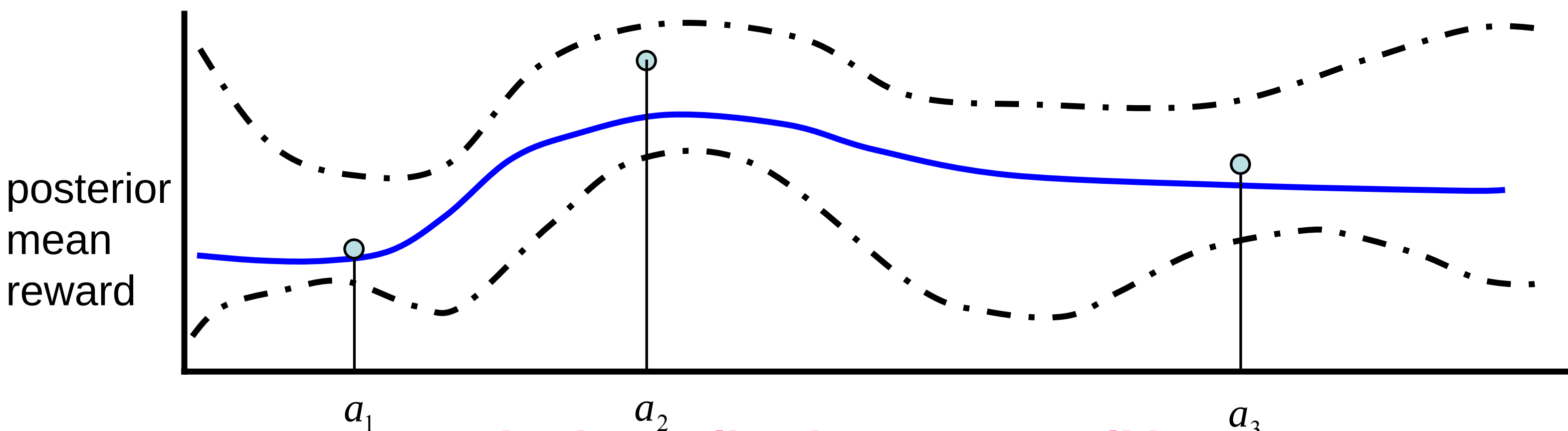


- Descend sparse tree from root
  - **Thompson sample actions**
  - Sample outcome
- Until new node added
- Repeat until tree size limit reached

Control computation by controlling tree size

Application: Gaussian process bandits

- General action spaces
  - Continuous actions, multidimensional actions
- Gaussian process prior over reward models
  - Covariance kernel between actions
- Action rewards correlated
- Posterior is a Gaussian process

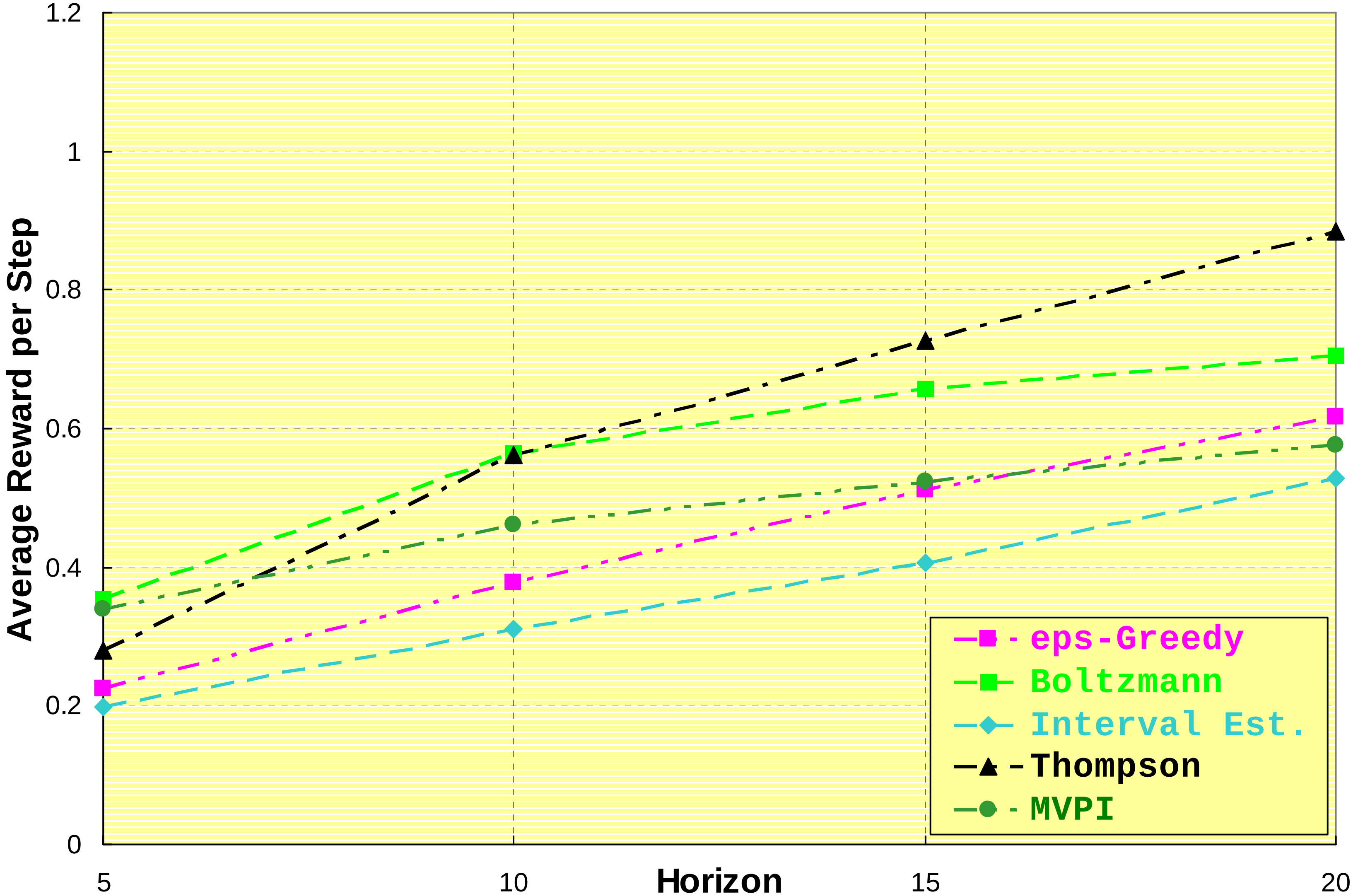


Robotic application: AIBO walking

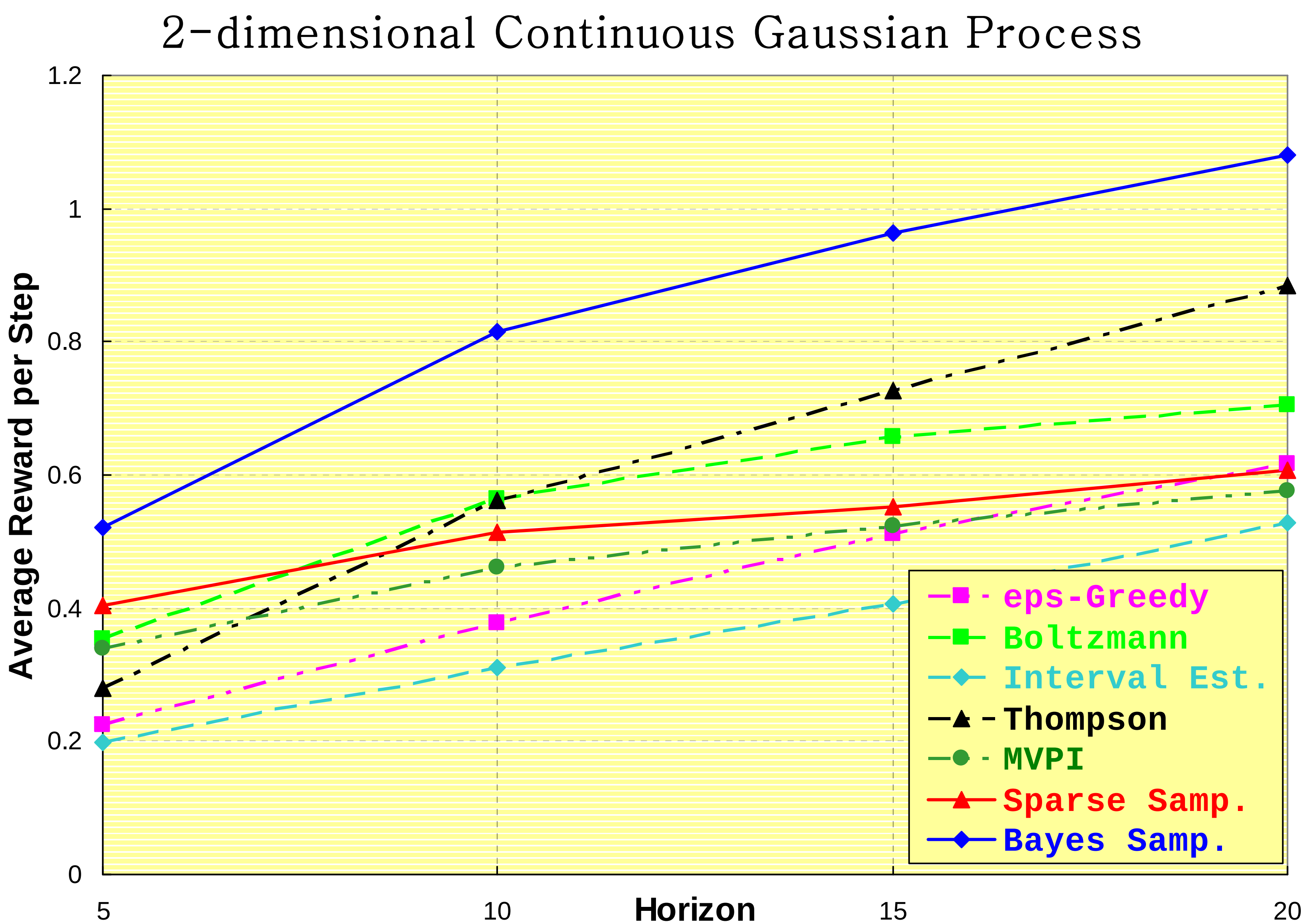
Gaussian process experiments

- 2 dimensional continuous action space
- GP priors RBF kernel
- Sampled model from prior
- Run action selection strategies
- Repeat 3000 times
- Average accumulated reward per step

2-dimensional Continuous Gaussian Process



UNIVERSITY OF ALBERTA


2-dimensional Continuous Gaussian Process
Average Reward per Step
Horizon
1.2
1
0.8
0.6
0.4
0.2
0
5
10
15
20
eps-Greedy
Boltzmann
Interval Est.
Thompson
MVPI
Sparse Samp.
Bayes Samp.

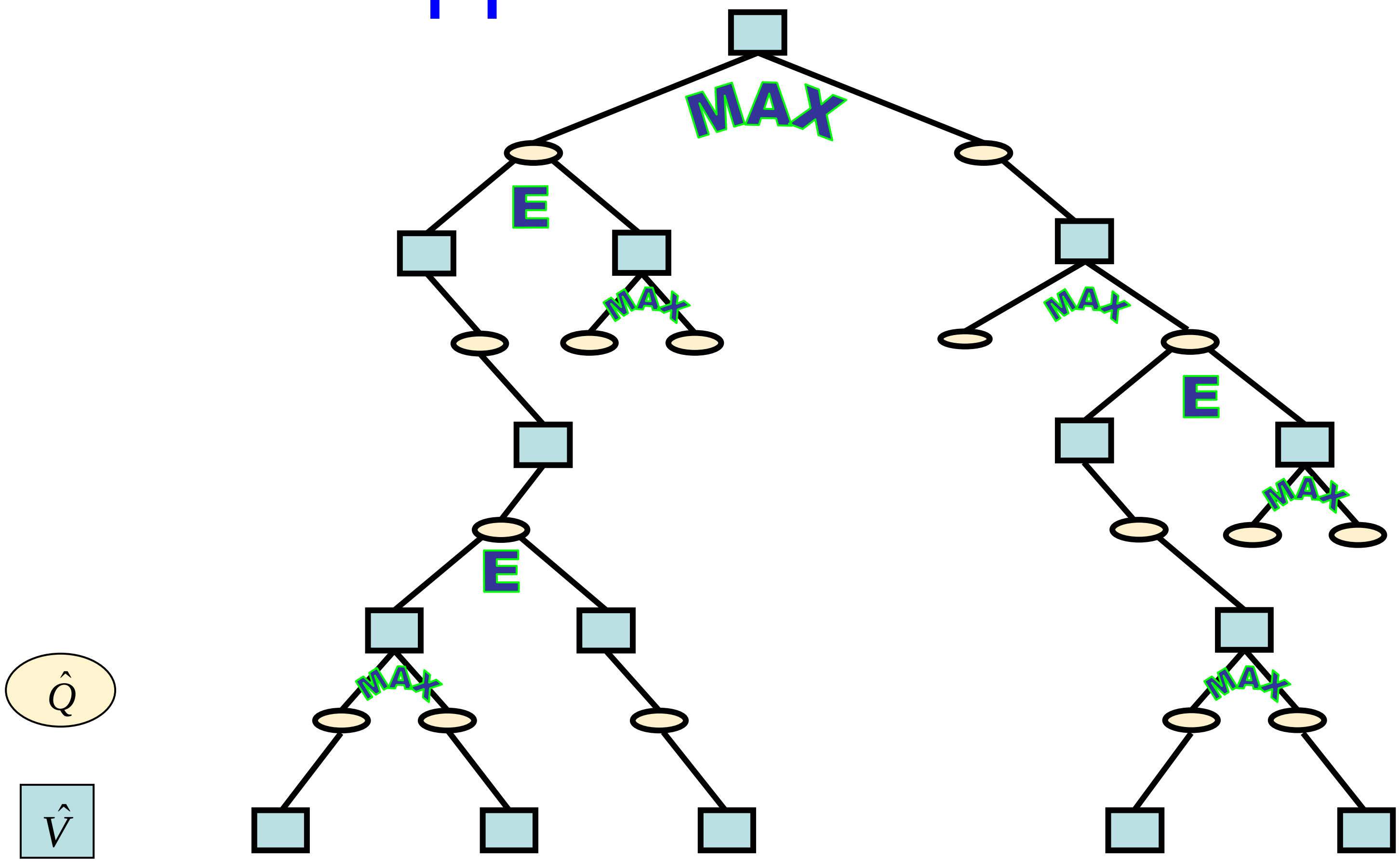
Summary

Bayesian sparse sampling

- Flexible and practical technique for improving action selection
- Reasonably straightforward
- Bandit problems
  - Planning is “easy”

    (at least approximate planning is “easy”)

Question:
How to approximate leaf values?



Part 2
Approximate value function

Compact, Convex Upper Bound Iteration for Approximate POMDP Planning

Tao Wang Pascal Poupart Michael Bowling Dale Schuurmans

AAAI2006

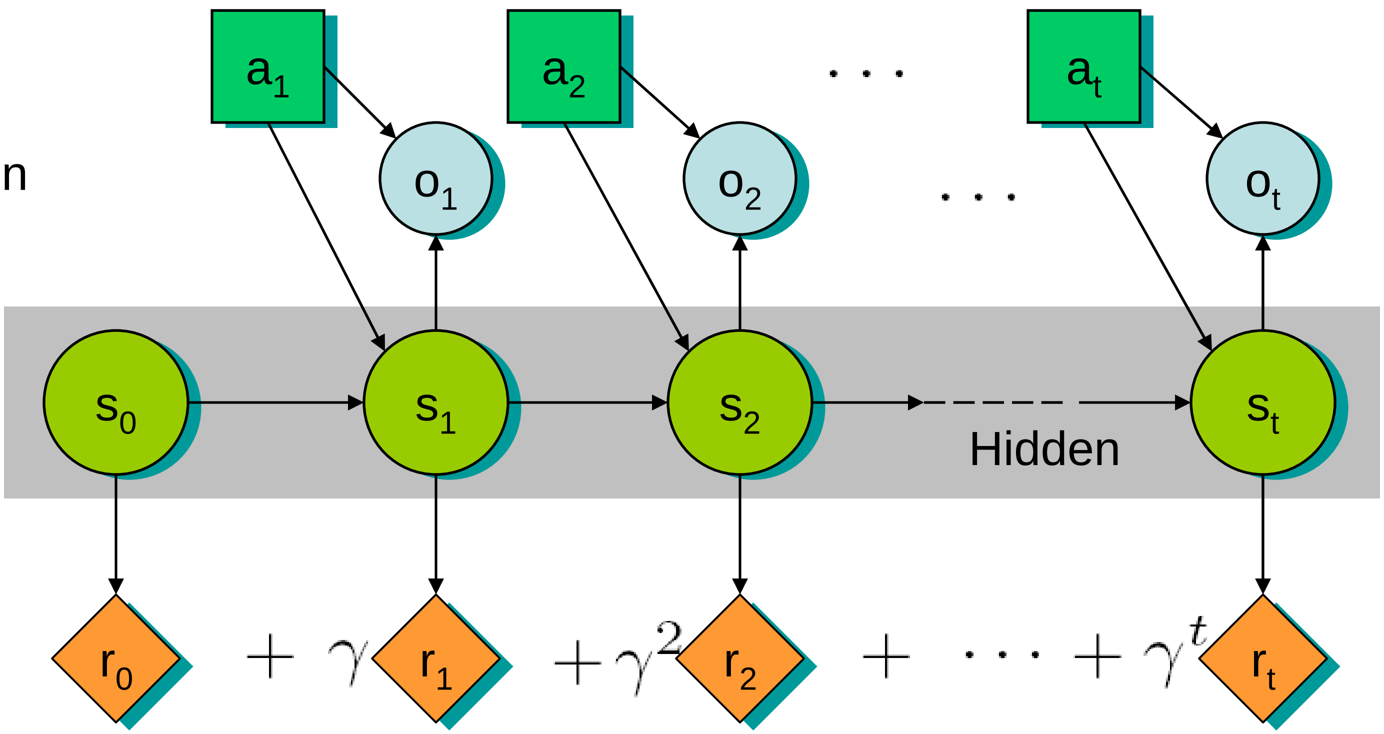
UNIVERSITY OF ALBERTA

POMDP model

Observation function
$p(o'|a, s')$

Transition function
$p(s'|s, a)$

Reward function
$R(s, a)$



Goal: choose actions to maximize long term reward

Belief state

- Probability distribution over underlying states

$$b = \begin{bmatrix} p(s^1) \\ p(s^2) \\ \vdots \\ p(s^{|\mathbb{S}|}) \end{bmatrix}$$

- Sufficient summary of history for decision making $b(\overline{s}_t) = p(\overline{s}_t | b_0 a_1 o_1 a_2 o_2 \cdots a_t o_t)$

POMDP solving

- Value function is expected total discounted future reward starting from each belief state

Optimal Value function

$$V^*(b) \;=\; \max_a r(b,a) + \gamma \sum_{b'} p(b'|b,a) V^*(b')$$

- **Hard to approximate**

POMDP approximation approaches

- Value function approximation (Part 2)
  - Hauskrecht 2000
  - Spaan&Vlassis 2005
  - Pineau et al. 2003
  - Parr&Russell 1995
- Policy based optimization
  - Ng&Jordan 00; Poupart & Boutilier 03,04; Amato et al. 06
- Stochastic sampling (Part 1)
  - Kearns et al. 02; Thrun 00

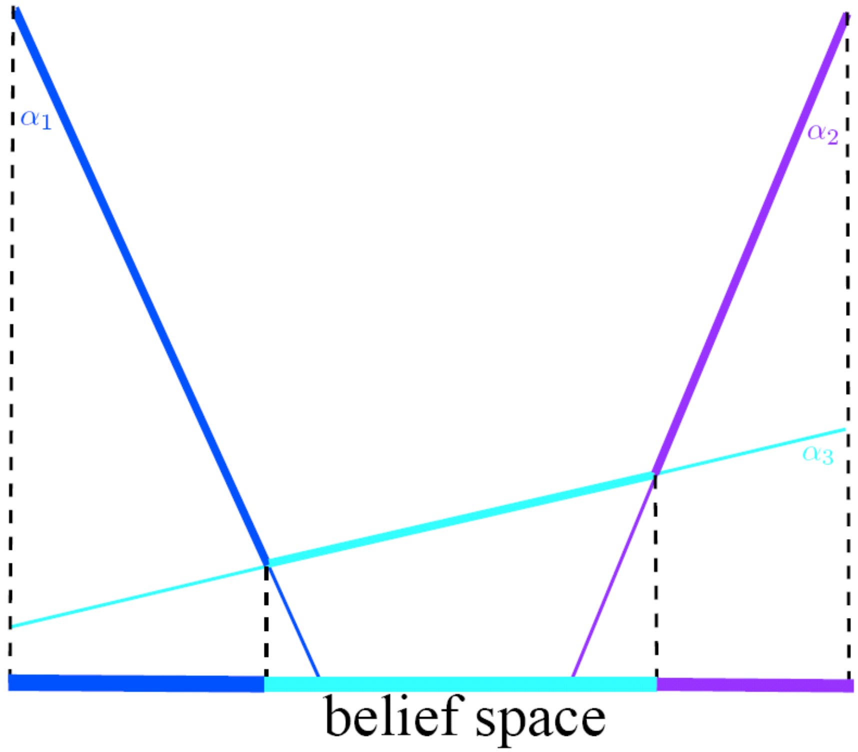
Value function based approaches

- Optimal value function
  (satisfies Bellman equation)

$$
\begin{aligned}
V^*(b) &= \max_a r(b,a) + \gamma \sum_{b'} p(b'|b,a) V^*(b') \\
&= \max_a r(b,a) + \gamma \sum_{o'} p(o'|b,a) V^*(b'_{(b,a,o')})
\end{aligned}
$$

- **Difficulty: belief space is continuous & high dimensional**

Optimal 1-step decision



$$V_1(b) = \max_a b \cdot r_a$$

$$\Gamma_1 = \{r_{a_1}, r_{a_2}, r_{a_3}\} = \{\alpha_1, \alpha_2, \alpha_3\}$$

$$V_1(b) = \max_{\alpha \in \Gamma_1} b \cdot \alpha$$

Optimal value function is **piecewise linear convex**

Optimal n+1-step decision

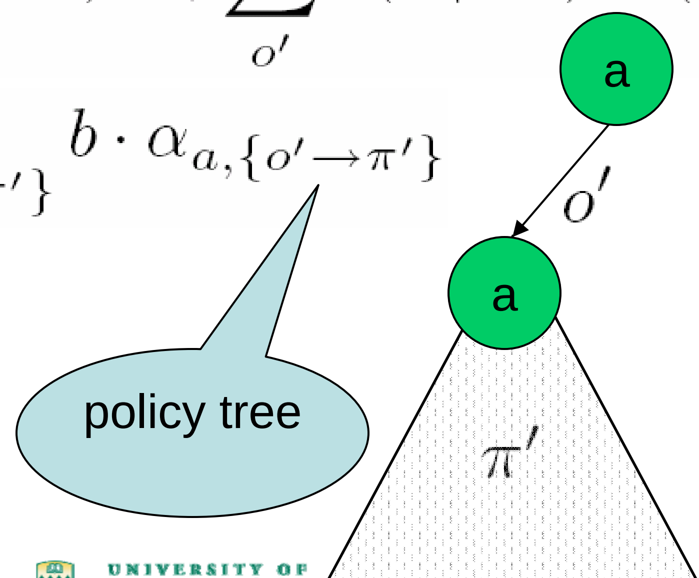
Value function representation

$$V_n(b) \;=\; \max_{\alpha_{\pi'} \in \Gamma_n} b \cdot \alpha_{\pi'}$$

$$b = \begin{bmatrix} p(s_1) \\ p(s_2) \\ \vdots \\ p(s_{|\mathbb{S}|}) \end{bmatrix} \quad \alpha_{\pi'} = \begin{bmatrix} v_{\pi'}(s_1) \\ v_{\pi'}(s_2) \\ \vdots \\ v_{\pi'}(s_{|\mathbb{S}|}) \end{bmatrix} \quad \Gamma_n = \{\alpha_{\pi'} : \pi' \in \Pi_n\}$$

Value function iteration

$$\begin{aligned} V_{n+1}(b) &= \max_a r(b,a) + \gamma \sum_{o'} p(o'|b,a) V_n(b'_{(b,a,o')}) \\ &= \max_{a,\{o' \to \pi'\}} b \cdot \alpha_{a,\{o' \to \pi'\}} \end{aligned}$$



Current approximation strategies

- Grid based approach
  - Gorden 95
  - Hauskrecht 00
  - Zhou & Hansen 01
  - Bonet 02
- Belief point approach
  - Pineau et al. 03
  - Smith & Simmons 05
  - Spaan & Vlassis 05

value function representation

α-vectors

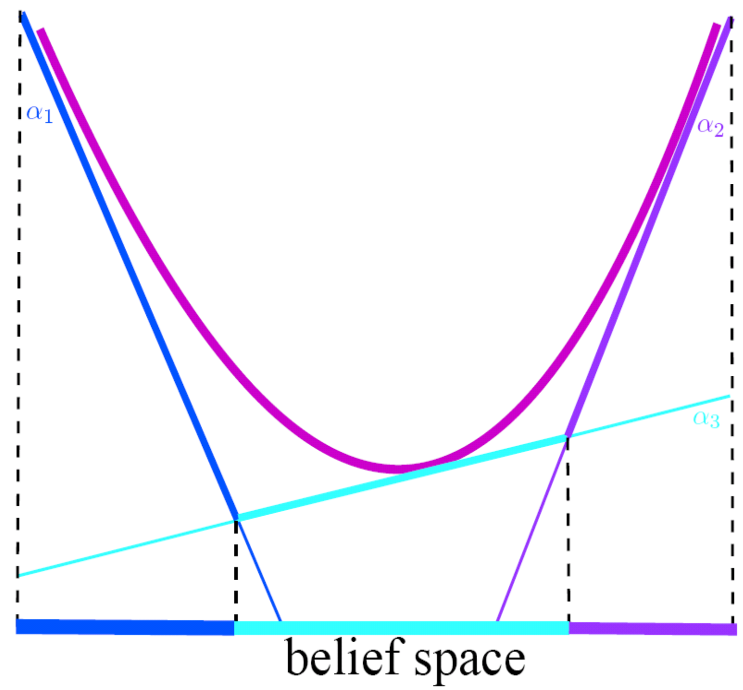
Our idea

Approximate $V^*(b)$ with a **convex quadratic** upper bound

- Maintain compact (bounded size) representation of value approximation

- Can still model multiple α-vectors

- Can be optimized easily



Quadratic approximation

Value function representation

$$\hat{V}(b) \quad = \quad b^\top W b + w^\top b + \omega$$

Would like to enforce

$$\hat{V}_{n+1}(b) \quad \geq \quad \max_a \hat{q}_a(b)$$

Need action-value backup for each action

$$\hat{q}_a(b) \quad = \quad r(b,a) + \gamma \sum_{o'} p(o'|b,a)\hat{V}_n(b'_{(b,a,o')})$$

Quadratic approximation

Combine with belief update

$$b'_{(b,a,o')} = \frac{M_{a,o'} b}{e^\top M_{a,o'} b} \qquad M_{a,o'}(s', s) = p(o'|a, s')p(s'|s, a)$$

Get action-value

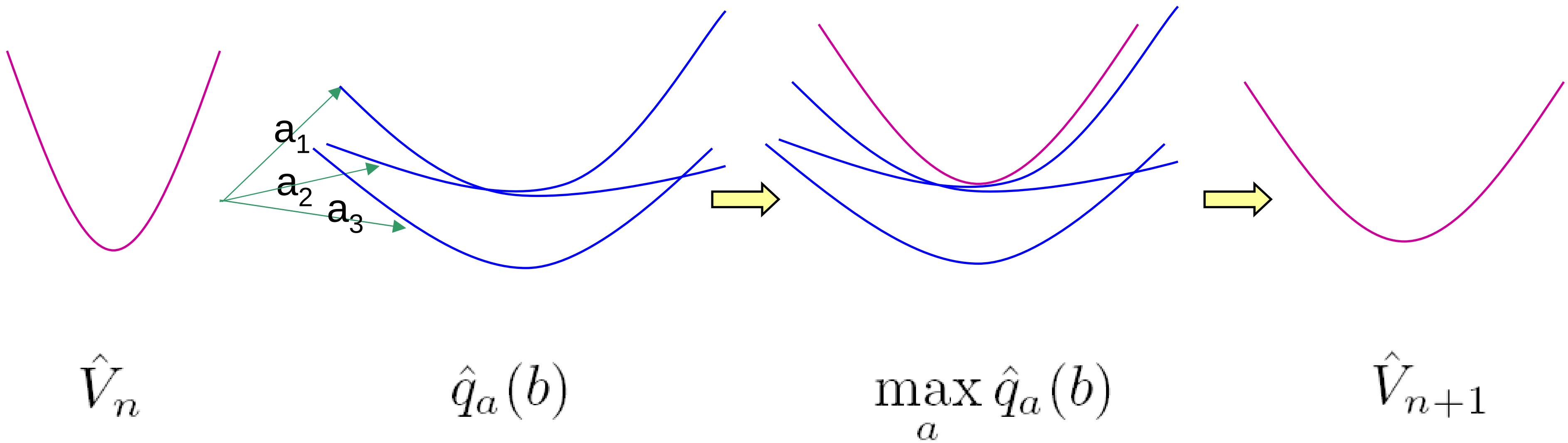
$$q_a(b) = r(b, a) + \gamma \sum_{o'} \frac{b^\top M_{a,o'}^\top W M_{a,o'} b}{e^\top M_{a,o'} b} + (w + \omega e)^\top M_{a,o'} b$$

quadratic
linear

Theorem 1 $q_a(b)$ *is convex in* b.

Corollary 1 $\max_a q_a(b)$ *is convex in* b.

Algorithm



Maintain tight upper bound of the maximum of the action-values

Mathematically

Optimization problem

$$\min_{W,w,\omega} \int_b \left(b^\top W b + w^\top b + \omega\right) \mu(b)\, db$$

a measure over space of possible beliefs

subject to

$$b^\top W b + w^\top b + \omega \geq q_a(b), \quad \forall\, a, b$$

ensure upper bound

$$W \succeq 0$$ (positive semi-definite)

ensure convexity

Two difficulties

- Integral in objective
- Infinite number of linear constraints

But it is a **convex optimization** problem
(SDP plus infinitely many linear constraints)

Integral

Objective $$\int_b \left(b^\top W b + w^\top b + \omega\right) \mu(b) db$$

is equal to $$\langle W, \mathrm{E}[bb^\top]\rangle + w^\top \mathrm{E}[b] + \omega$$ (*linear*)

Assume measure $\mu(b)$ is Dirichlet distribution on b

then $\mathrm{E}[bb^\top]$ and $\mathrm{E}[b]$ have **closed form**

Infinite constraints

Have $b^\top W b + w^\top b + \omega \geq q_a(b), \quad \forall\, a, b$

infinitely many linear constraints on $W \; w \; \omega$

Optimal constraint generation: most violated constraint

$$\min_b b^\top W b + w^\top b + \omega - q_a(b)$$

$$\text{subject to} \quad b \geq 0, \quad \textstyle\sum_s b(s) = 1$$

Unfortunately, not necessarily a convex minimization problem in *b*

Experimental results

- Benchmark problems
- Mean discounted reward & Run time
  - 10 runs
  - 1000 trajectories
- Competitors
  - Perseus (Pineau et al. 2003)
  - PBVI (Spaan & Vlassis 2005)

Problem characteristics

Problems	$\|S\|$	$\|A\|$	$\|O\|$
Maze	20	6	8
Tiger-grid	33	5	17
Hallway	57	5	21
Hallway2	89	5	17
Aircraft	100	10	31

Mean discounted reward

Avg. Reward	CQUB	Perseus	PBVI
Hallway	0.58 ±0.14	0.51 ±0.06	0.53 ±0.03
Hallway2	0.43 ±0.25	0.34 ±0.16	0.35 ±0.03
Tiger-grid	2.16 ±0.02	2.34 ±0.02	2.25 ±0.06
Maze	45.35 ±3.28	30.49 ±0.75	46.70 ±2.00
Aircraft	16.70 ±0.58	12.73 ±4.63	16.37 ±0.42

Compact representation

Size	CQUB	Perseus	PBVI
Maze	231	460	1160
Tiger-grid	595	4422	15510
Hallway	1711	3135	4902
Hallway2	4095	4984	8455
Aircraft	5151	10665	47000

Summary

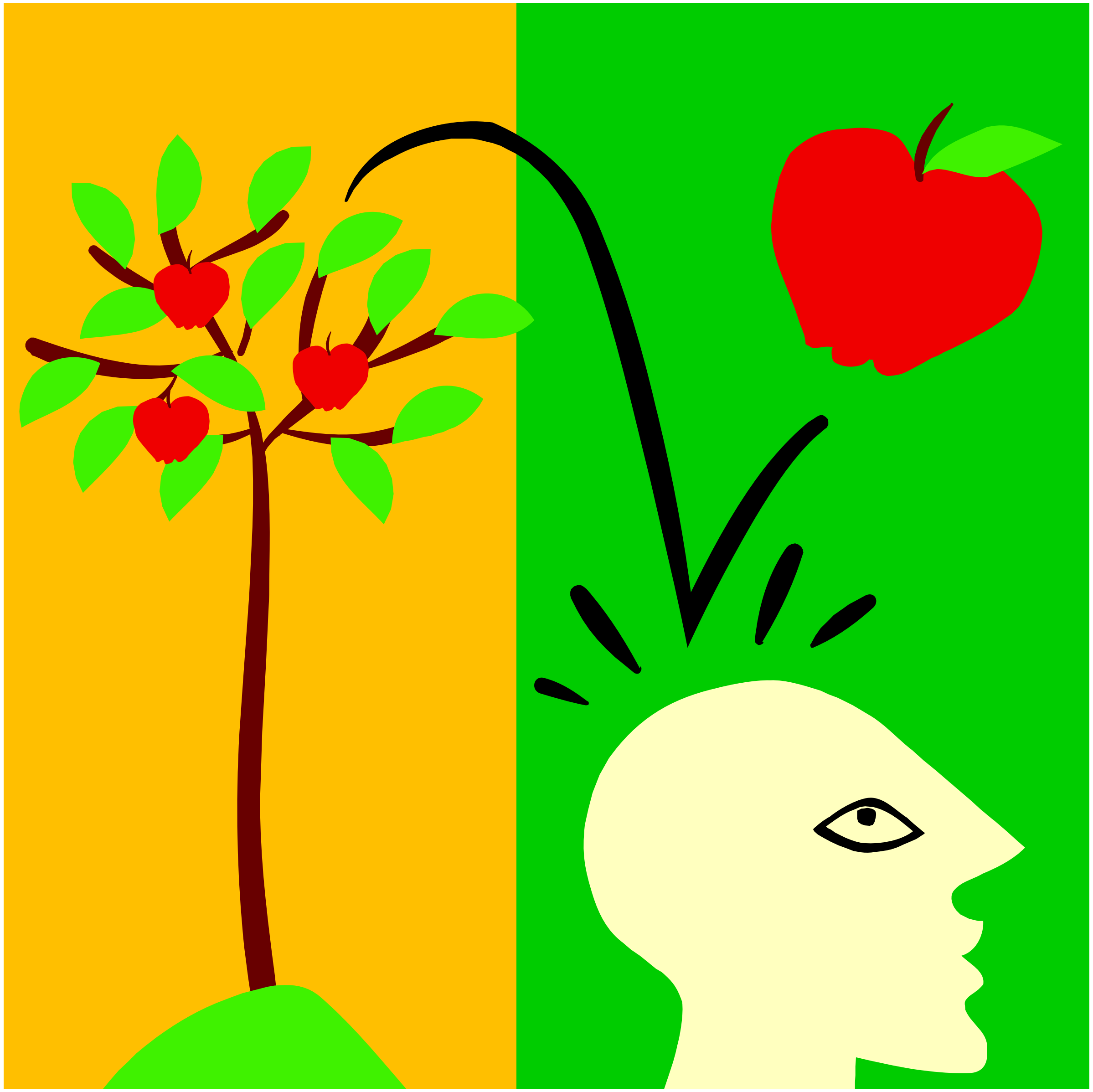
New approximation algorithm

- Quadratic value function approximator
- Compact representation
- Competitive approximation quality
- Provable upper bound on the optimal values
- Computational cost independent of iteration number

Contributions

- Use game-tree search ideas
  - ✓ Bayesian sparse sampling
  - ✓ Approximate POMDP planning
  - ➢ To combine them to solve meta-level MDPs

- Exploit Bayesian modelling tools
  - E.g. Gaussian processes
  - Robotic & game applications



Future work on Bayesian sparse sampling

AIBO dog walking

Opponent modeling (Kuhn poker)

Vendor-bot (Pioneer)

Improve tree search?

Theoretical guarantees?

Cheaper re-planning?

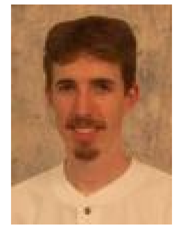
Future work on approximate planning

- Set of quadratics
- Use belief state compression & factored models
- Combine with sampling
- Interpretation: 2nd order Taylor expansion

Research questions

- How to choose actions during reinforcement learning?
- How to scale up solutions for realistic RL problems?
- How to combine Part 1 and Part 2?
  - Challenge: infinite dimensionality
  - Scale up Part 2 (run time)

Acknowledgments



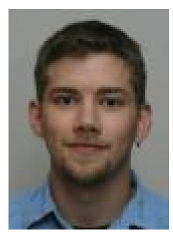
Dale Schuurmans

Mike Bowling

Rich Sutton

Paul Messinger

Robert Holte



Pascal Poupart

Daniel Lizotte

Adam Milstein