Coordination in Multiagent Reinforcement Learning: A Bayesian Approach

Georgios Chalkiadakis & Craig Boutilier

Department of Computer Science
University of Toronto

Coordination / Equilibrium Selection

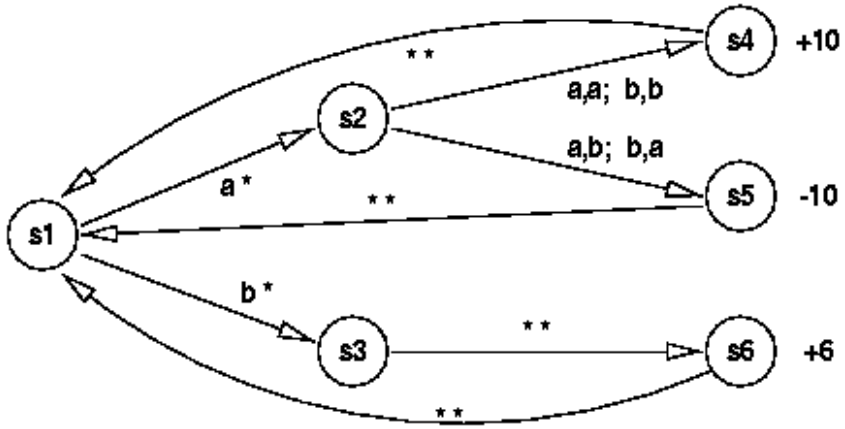
- *Coordination* of agent activities an important focus of MARL
  - (Identical interest) stochastic games provide a useful model for studying such problems
- A repeated game example: *The Penalty Game…the equilibrium selection problem (curse of multiple equilibria)*

	a0	a1	a2
b0	10	0	k
b1	0	2	0
b2	k	0	10

[Claus & Boutilier 1998:
The dynamics of reinforcement learning in cooperative multiagent systems]

- A stochastic game example:

The Opt in or out Game…



To Avoid the Suboptimal Equilibria?

- A number of (mainly heuristic) methods proposed to avoid convergence to suboptimal equilibria [CB96, LR00, KK02, WS02]
  - Generally, adopt *an optimistic bias* in exploration, in an attempt to reach optimal equilibrium
  - Some methods even guarantee convergence to optimal equilibrium
  - *Tradeoff*: Is the price paid – penalties, lost opportunities – worth the gain offered by convergence to optimal equilibrium?
    - Depends on discount factor, horizon, odds of converging to specific equilibrium, etc.

An "optimal" MARL exploration method should be able to address this tradeoff

Bayesian Perspective on MARL

- *Single-agent* Bayesian RL: Bayesian update of distributions over possible rewards and transition dynamics models [Dearden et al.]
- We have adopted this point of view for Bayesian exploration in *multi-agent* RL settings
- However, new components required:
  - Priors over models include opponents' strategies
  - Action selection is formulated as a POMDP:
    - Value of information includes what is learned about opponents' strategies
    - Object level value includes how action choice will impact what the opponents will do

Basic Setup

- Assume a stochastic game as a framework for MARL
  - States S, fully observable
  - Players i $\in$ {1,…,N}
  - Action sets A_i , joint action set $A = \times A_i$
  - Transition dynamics $Pr(s, \boldsymbol{a}, t)$
  - Stochastic reward functions R_i
  - Strategies σ_i, strategy profiles σ, σ_{-i}
- Each agent's experience is a tuple $<s, \boldsymbol{a}, \boldsymbol{r}, t>$

Agent Belief State

- Belief state: $b = <P_M , P_S , s, h>$
  - P_M : density over space of possible models (games)
  - P_S : density over space of opponents' strategies
  - s : current state of the game
  - h : relevant history
- Update belief state given experience <s, **a**, **r**, t>
  - $b' = b(<s, \mathbf{a}, \mathbf{r}, t>) = <P'_M , P'_S , t, h'>$
  - Densities obtained by Bayes rule
    - $P'_M (m) = z\, Pr(t, \mathbf{r} \mid \mathbf{a}, m)\, P_M(m)$
    - $P'_S (\sigma_{-i}) = z\, Pr(\mathbf{a}_{-i} \mid s, h, \sigma_{-i})\, P_S(\sigma_{-i})$
    - This combines Bayesian RL and Bayesian strategy learning

Simplifying Assumptions

- P_M factored into independent local models $P_D^{s,a}$, $P_R^{s,a}$
  - Assume local densities are Dirichlet,
  - This allows for easy updating of P_M
- Some convenient prior P_S
  - We use simple fictitious play beliefs (no history)
  - More general models are feasible
  - Interesting question: what are reasonable, feasible classes of opponent models?

Tradeoffs in Optimal Exploration

- Given belief state b, each action a_i:
  - has expected object level value
  - provides info which can subsequently be exploited
- Object level value:
  - immediate reward
  - predicted state transition (expected value)
  - impact on future opponent action selection
- Value of information:
  - what you learn about transition model & reward
  - what you could learn about opponent strategy
  - how this info impacts own future decisions

POMDP Formulation

- Tradeoff can be made implicitly by considering long-term impact of actions on belief states and associating value with belief states:

$$Q(a_i,b) = \sum_{a_{-i}} \Pr(a_{-i}|b) \sum_t \Pr(t|a_i o a_{-i},b)$$
$$\sum_r \Pr(r|a_i o a_{-i},b)[r+\gamma V(b(<s, a, r, t>))]$$
$$V(b) = \max_{a_i} Q(a_i,b)$$

where

$$\Pr(a_{-i}|b) = \int_{\sigma_{-i}} \Pr(a_{-i}|\sigma_{-i}) P_S(\sigma_{-i})$$
$$\Pr(t|a,b) = \int_m \Pr(t|s,a,m) P_M(m)$$
$$\Pr(r|b) = \int_m \Pr(r|s,m) P_M(m)$$

- These equations describe the solution to the POMDP that represents the multiagent exploration-exploitation problem.
- Solving this belief state MDP is intractable

Computational Approximations I

- Myopic Q-function equations, assuming a *fixed* distribution over models and strategies:

$$Q_m(a_i,b) = \sum_{a_{-i}} \Pr(a_{-i} \mid b) \sum_t \Pr(t \mid a_i o a_{-i}, b)$$
$$\sum_r \Pr(r \mid a_i o a_{-i}, b)[r+\gamma V_m(b(\langle s, a, r, t\rangle))]$$
$$V_m(b) = \max_{a_i} \int_m \int_{\sigma_{-i}} Q(a_i, s \mid m, \sigma_{-i})\, P_M(m)\, P_S(\sigma_{-i})$$

- What they mean intuitively: *One step lookahead* in belief space followed by evaluation of the expected value of these successor states…

Computational Approximations II

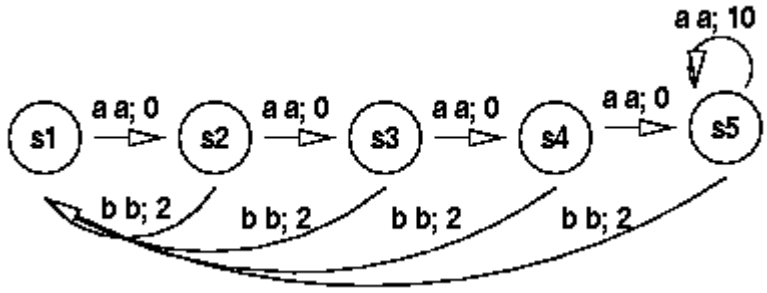
- Other approaches include using the rather different *naïve Q-value sampling approach* to estimating EVOI [Dearden et al.] (see paper for details)
  - some number of models can be sampled
  - the MDPs are solved
  - Q-values are estimated by averaging over the results
  - decision is made on whether these values are sufficient to change the optimal action choice at the current state

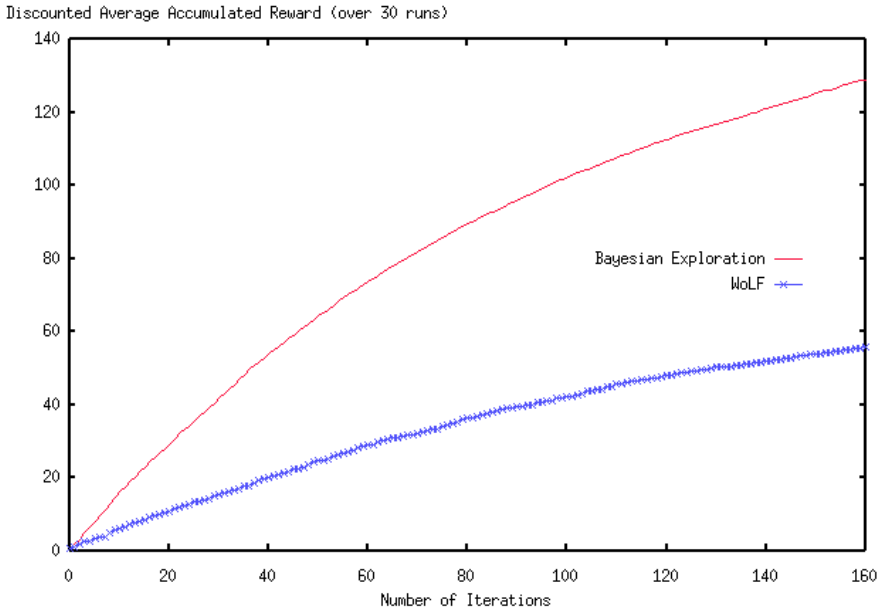
Experiments

- Tested the Bayesian approach using both
  - One-step lookahead using expectations for strategies - for single state games (BOL)
  - Naïve VPI sampling (BVPI)
- Compared on several repeated and (multi-state) stochastic games to several algorithms:
  - KK (Kapetanakis & Kudenko, AAAI-02)
  - (model-based versions of) OB & CB (Claus & Boutilier ’98)
  - WoLF-PHC (Bowling & Veloso, IJCAI-01)
    - Much more general algorithm
- Compared using total discounted reward accrued

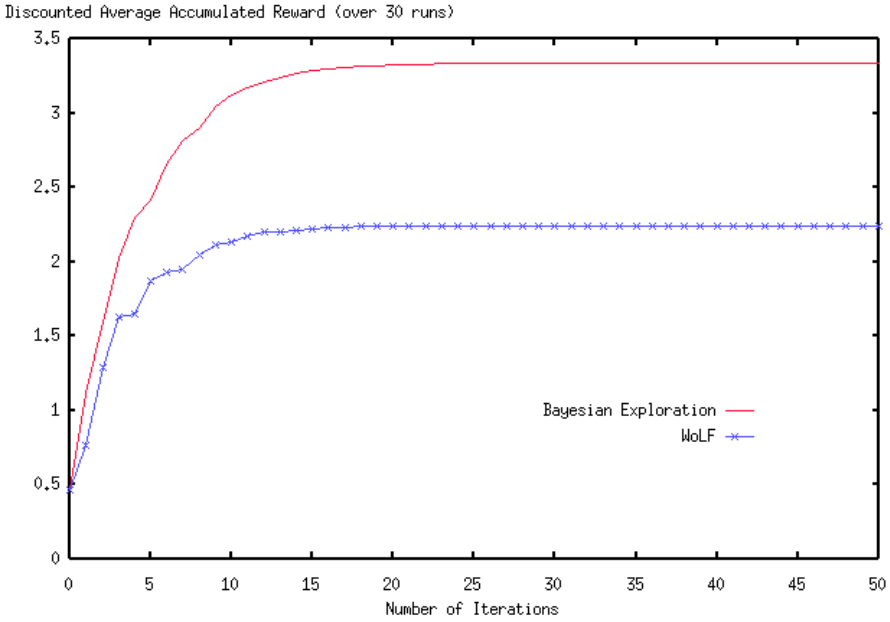
“Chain World” Results

Chain World Domain





$\gamma = 0.99$

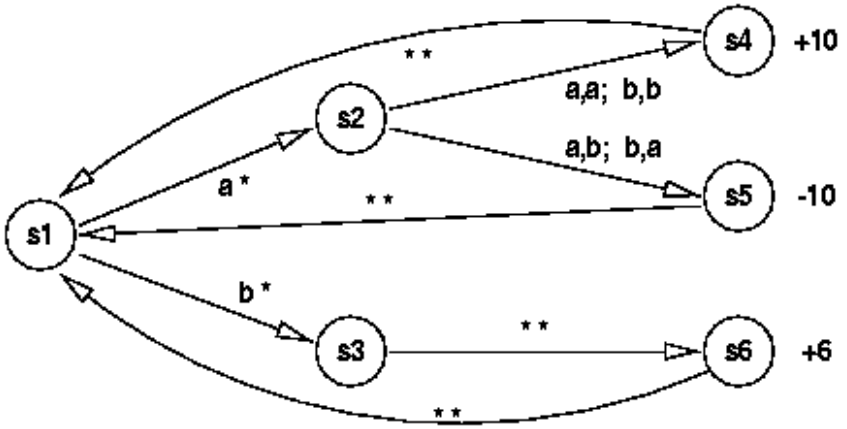


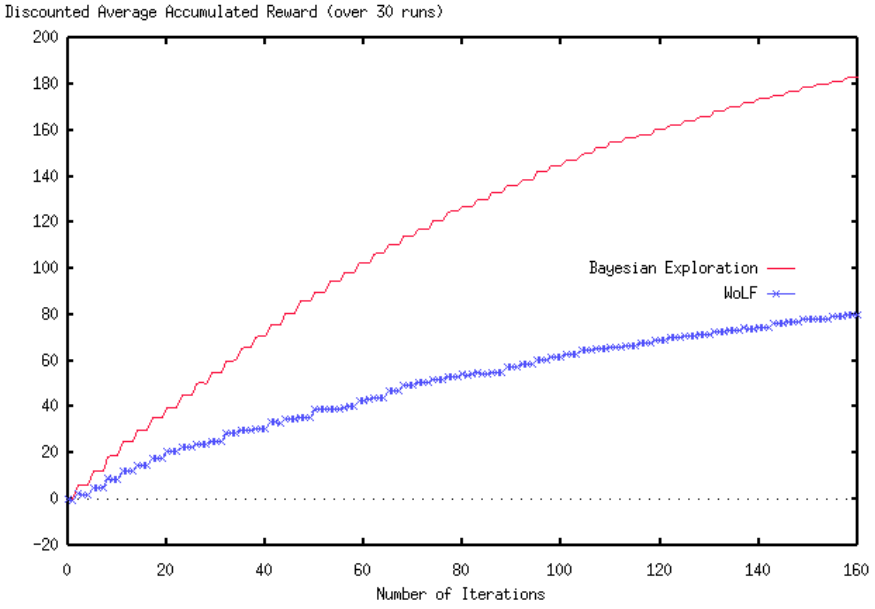
$\gamma = 0.75$

“Opt in or out” Results I

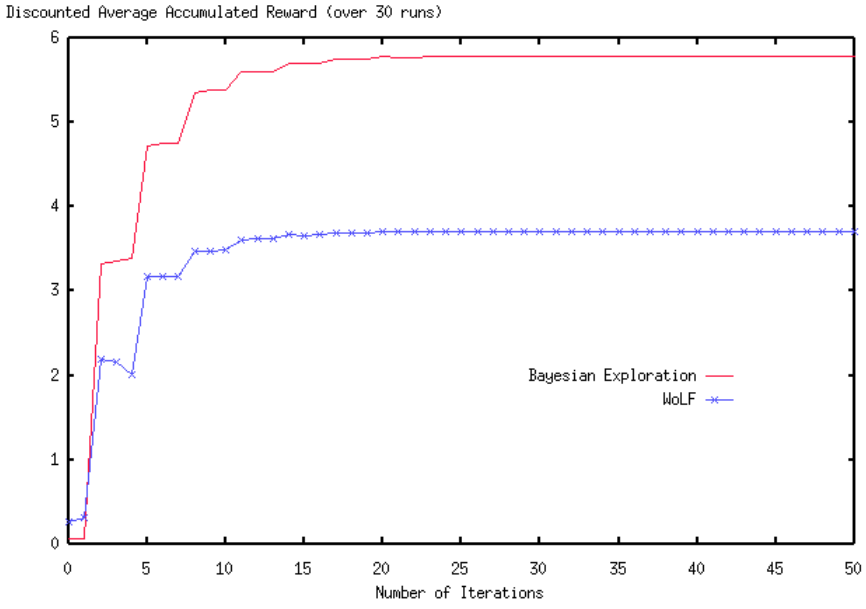
Opt in or out Domain

“Low” Noise





$\gamma = 0.99$

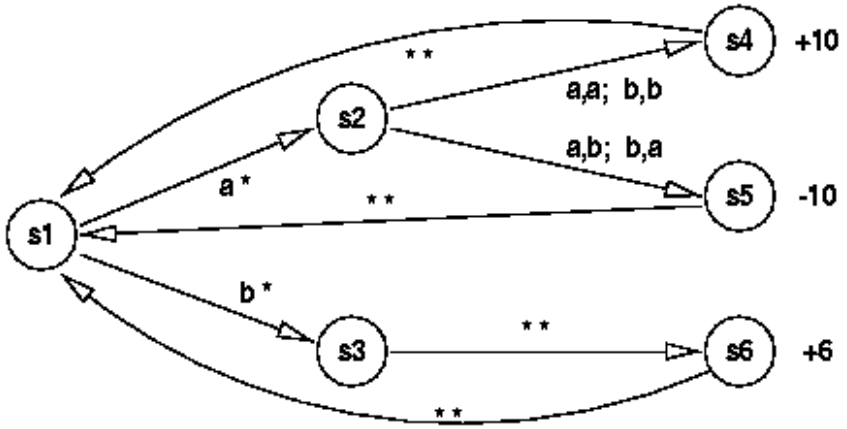


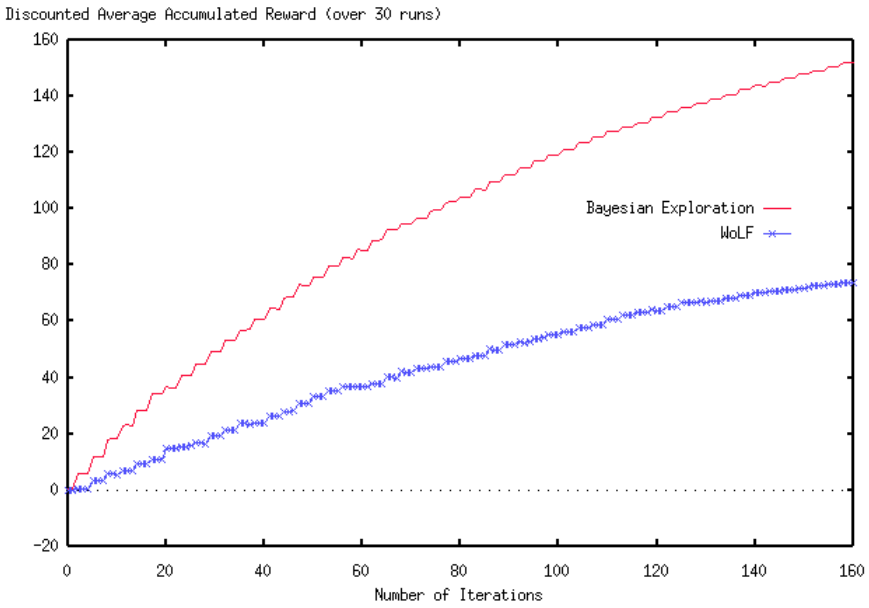
$\gamma = 0.75$

“Opt in or out” Results II

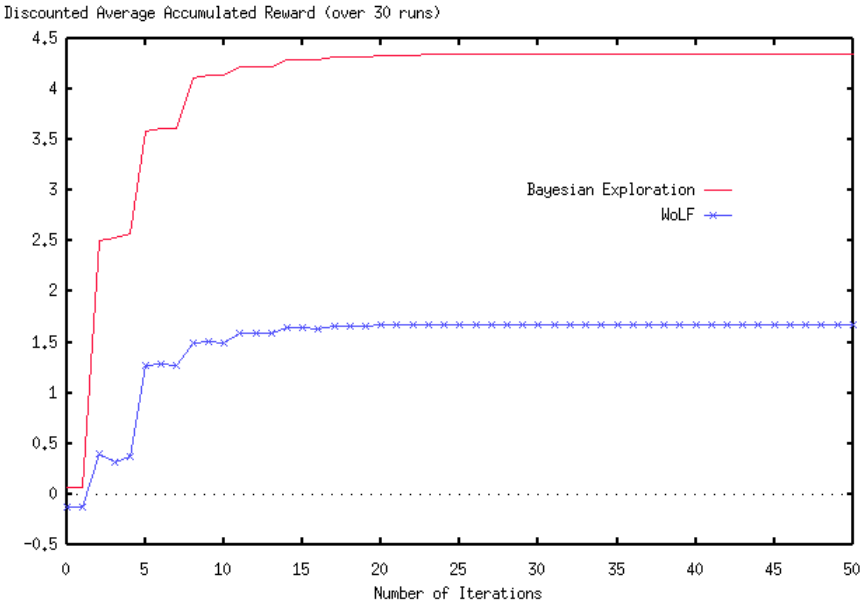
Opt in or out
Domain

“Medium” Noise



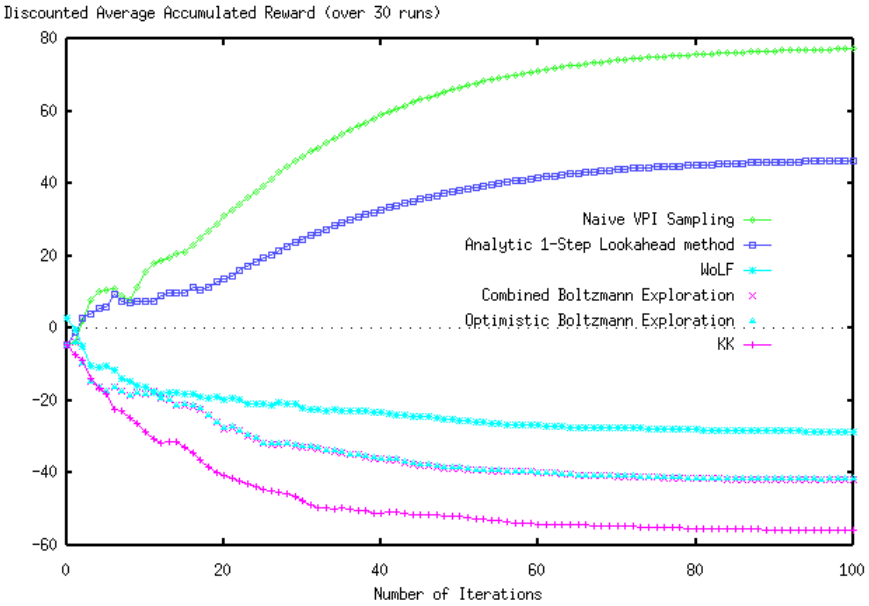


$\gamma = 0.99$

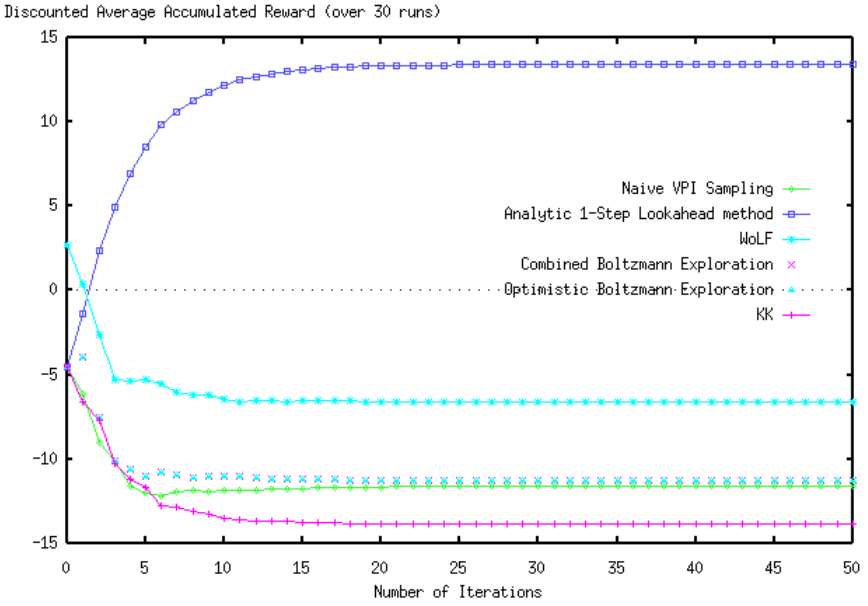


$\gamma = 0.75$

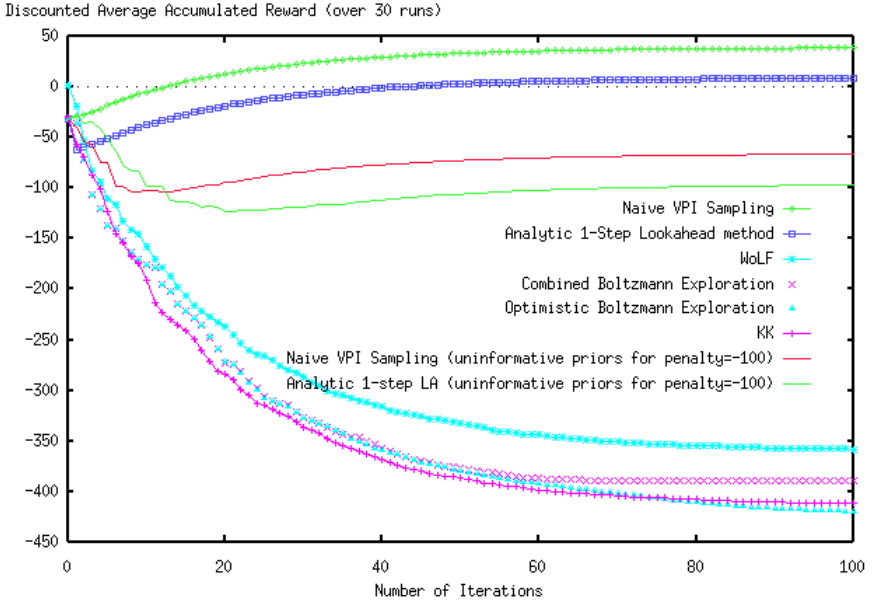
Penalty Game Results



Uninformative Priors, k=-20, $\gamma = 0.95$



Uninformative Priors, k=-20, $\gamma = 0.75$



Informative Priors, k=-100, $\gamma = 0.95$

Concluding Remarks

✓ Bayesian agents explicitly reason about their uncertainty regarding the domain and their opponents' strategies

✓ We provided a formulation of optimal exploration under this model

✓ We developed several computational approximations for Bayesian exploration in MARL.

➢ Bayesian exploration agents don't *necessarily* converge to optimal equilibria; they weigh exploration benefits against exploration costs…

- But in many cases, they do converge to optimal equilibria

➢ Generally, they perform better than other approaches wrt. discounted reward

➢ The model is flexible

- Can use various priors; opponent models; discount/horizon

Future Work

- Framework is general, but experiments involved only identical interest games
  - Need to apply framework to more general problems

- Use more sophisticated opponent models than fictitious play beliefs

- More work on computational approximations to estimating VPI or solving the belief state MDP is required

- Develop computationally tractable means of representing and reasoning with distributions over strategy models

Thank you!

Any questions?

Coordination in Multiagent Reinforcement Learning: A Bayesian Approach

Georgios Chalkiadakis
gehalk@cs.toronto.edu
Craig Boutilier
cebly@cs.toronto.edu

Computational Approximations: Naïve VPI Sampling ("BVPI") I

- This method tries to estimate the (myopic) value of obtaining perfect information about Q(a,s); *does not* perform 1-step LA in belief space

- $gain_{s,a}(q) = \begin{cases} EV(a2, s) - q, \text{ if } a=a1 \text{ and } q < EV(a2,s) \\ q - EV(a1, s), \text{ if } a \neq a1 \text{ and } q > EV(a1,s) \\ 0 \qquad\qquad \text{otherwise} \end{cases}$

Computational Approximations: Naïve VPI Sampling ("BVPI") II

- A finite set of k models are sampled from P_M
- Each sampled j MDP is solved, w.r.t. P_S, giving optimal $Q^j(a_i,s)$ for each a_i in that MDP, and average $EV(a_i,s)$ over all k MDPs
- For each a_i, compute the gain w.r.t. each $Q^j(a_i,s)$. Define $EVPI(a_i,s)$ to be the average over all k MDPs of $gain_{s,a^i}(Q^j(a_i,s))$
- Execute the action that maximizes

$$EV(a_i,s) + EVPI(a_i,s)$$

Computational Approximations II

- Problems… A*R*S successor belief states; direct evaluation of the integral over all models m is impossible

➔ *sampling*: some number of models can be sampled; the MDPs solved; Q-values estimated by averaging over the results

MDPs – Connection with RL

- RL: the agent-environment interaction can be modeled as an MDP
- MDP: <S, A, R, Pr>; R(s,r), Pr(s,a,s')
- MDPs: R, Pr are *known*
- RL: an MDP can be viewed as *a complete specification of the RL environment, that satisfies the Markov property*
  - » *Markov Property: current state is independent of previous states or actions*

MDPs-RL continued

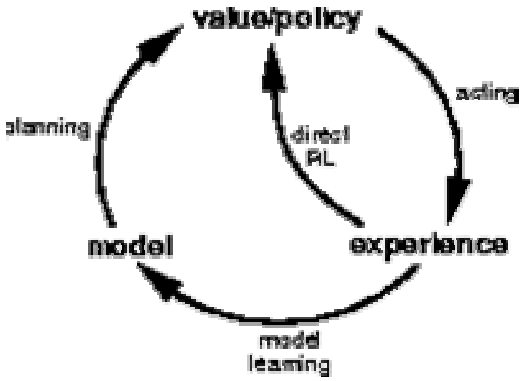
- MDPs infinite-horizon model of optimal behavior:
  Goal: Construct a policy π:S→A such that

  $E_\pi [\Sigma_{t=0\ldots\infty}\gamma^t R^t \mid S^0=s]$ is maximized

  V*(s): optimal value at s denotes the long term desirability of s, and can be computed, e.g., with value/policy iteration
- RL : *same goal* **but…**
  - *In the RL case, R and Pr are* ***not known…***

  So, in the RL case, the agent has to learn a policy based on her interactions with the environment.
  This can be done via ***direct (model-free)*** *methods, or via* ***model-based*** *methods…*

Model-based RL



Model not known, but can be *learned*…

- *Model-based* algorithms use a model of the environment to update the value function. In case the model is not given a priori (RL) it has to be estimated:

$$\langle s,a,r,t \rangle \rightarrow \text{update } \hat{R} \text{ and } \hat{Pr}$$

to maintain an estimated $\langle S, A, \hat{R}, \hat{Pr} \rangle$ MDP

By learning a model, the agent makes fuller use of experiences. Also, costly repetition of steps in the environment can be avoided.

Bayesian Model-Based RL

[Dearden et al., 1999: "Model-based Bayesian exploration"]

- Prior density *P* over (transition dynamics) *D* and (reward distributions) *R* is assumed; *P* is updated with each <s,a,r,t>; *Action selection* using *P(D,R|H)*
- *P* is factored over *R* and *D*
- *P(D)* and *P(R)* are the products of independent local densities (e.g., $P(D^{s,a})$ & $P(R^{s,a})$) for each distribution Pr(s,a,t) or Pr(s,a,r)
- Each $P(D^{s,a})$, $P(R^{s,a})$ is a Dirichlet; Dirichlet priors are conjugate to the multinomial distribution → Dirichlet posteriors

(Dirichlet distributions use some prior counts (hyper-parameters) for the possible outcomes, and update those based on observed experience to come up with a prediction for each outcome)

$$P(D^{s,a}|H^{s,a}) = a \Pr(H^{s,a}|D^{s,a}) P(D^{s,a})$$

$$P(R^{s,a}|H^{s,a}) = a \Pr(H^{s,a}|R^{s,a}) P(R^{s,a})$$

The agent uses these posteriors to decide on an appropriate action.

Stochastic Games

- A stochastic game is a tuple

  $<S, N, A_1,...,A_n, p_T, r_1,...,r_n>$
- Analogies with MDPs are apparent, but actions are *joint* ;
- Goal: maximization of the sum of expected discounted rewards – but now other agents are present too…
- *Repeated* games are a special case of stochastic games having only one state. A repeated game is made up from repetitions of a single strategic (normal form/ matrix) game…

Some notation: σ, σ_i, σ_{-i}, $BR(\sigma_{-i})$

Q-values, Boltzmann, OB, CB Exploration

"Q-values" can be calculated with value iteration:

$$Q(a,s) = E_{Pr(r|\,a,s)}[r|\,a,s] + \gamma \sum_{s'} V(s')$$

$$V(s) = \max_a Q(a,s)$$

- Boltzmann Exploration

  Action a is chosen with probability: $\dfrac{e^{Q(s,a)/T}}{\sum_{a'} e^{Q(s,a')/T}}$

Optimistic Boltzmann: instead of $Q(a_i)$, uses $MaxQ(a_i)$

Combined Boltzmann: instead of $Q(a_i)$, uses:

$$C(a_i) = \rho\, MaxQ(a_i) + (1-\rho)\, EV(a_i)$$

($EV(a_i)$ is the expected Q-value of a_i given fictitious play beliefs about the opponent's strategy)

[Claus & Boutilier 98]

Identical Interest Games: Some previous approaches

- Claus & Boutilier had brought those issues regarding coordination to light…
- *JALs*: Q-learning of Q-values of joint actions + use of fictitious play → calculation of *expected Q-values*; exploration is biased by the expected Q-values (Combined Boltzmann Exploration); convergence to equilibrium, but *not necessarily* an optimal equilibrium.
  - » The introduction of the penalty game was meant to show that, really, maybe sometimes it doesn't worth it to converge to optimal…

Computational Approximations: Analytic One-Step Look-ahead Method ("BOL")

- When there is only one state, we <u>can</u> compute the **one-step LA** expected value of perfoming an action at b, analytically:

$1StepLAVal(a_i, b) =$

$= \sum_{a_{-i}} Pr(a_{-i} | b) \sum_r Pr_{Dirichlet<a_i, a_{-i}>}(r) \{r+\gamma\ ExpVal(b')\} =$

$= \sum_{a_{-i}} Pr(a_{-i} | b) \sum_r Pr_{Dirichlet<a_i, a_{-i}>}(r) \{r+\gamma/(1-\gamma)\ MaxEV_b,\}$

because $ExpVal(b') = (1/(1-\gamma))\ MaxEV_{b'}$,

$MaxEV_b$, being the value of the optimal individual action at b'

Exploration vs. Exploitation "in a multiagent guise"

- Coordination requires exploration in parts of the policy space that are unrewarding.
- Coordination to an optimal equilibrium should be weighted against the possible costs.
- The strategies of others should be taken into consideration; is what we know so far enough or not?

Identical Interest Games: Some previous approaches II

- What others do:
  - Goal: Maximize the “discounted accumulated rewards”
  - But actually: seek/force convergence to optimal equilibrium
    - These are not actually compatible: they are compatible only if the agents actually start playing the optimal equilibrium early enough, **but in reality exploration will undoubtedly lead to costs in the process.**

Identical Interest Games: Some previous approaches III

- *The "Optimistic Assumption"*: Each agent assumes that others will act optimally, i.e. the chosen individual actions can be combined in an optimal action vector

  …*If so*, *therefore*, acting greedily in respect with the estimated Q-values can lead to the optimal equilibrium

- [Lauer & Riedmiller 2000] embody the optimistic assumption in the Q-values over *individual actions*, along with a mechanism to resolve ties between equilibria (The policy is updated only if Q-values are improved.)
- [Wang & Sandholm 2002], similarly, use *biased* Adaptive Play (Fictitious play with random samples; actions are chosen if have been recently played **and** they are contained into a set of optimal joint actions…) to force (and prove) convergence to optimal equilibria.
- [Kapetanakis & Kudenko 2002] use a heuristic exploration method that counts the frequency of achieving maximum reward so far in order to select an optimal *individual action*.

A *Bayesian* View of MARL

Note: Our approach is generic; but the results we have so far are well-suited to make our points regarding multiagent coordination in an RL environment.

The agents maintain probabilistic beliefs over the space of models and the space of opponent strategies to account for the effects of actions on:

- Knowledge/uncertainty of underlying model
- Knowledge/uncertainty of others' strategies
- Expected immediate reward
- Expected future behavior of opponents

Theoretical Underpinnings I

- Belief state: $b=<P_M, P_S, s, h>$ over current MDP models and opponent strategies models
- Updated belief state:

  $b'=b(<s, a, r, t>)=<P'_M, P'_S, t, h'>$
- Densities obtained by Bayes rule
  - $P'_M(m)= zPr(t,r \mid a,m)P_M(m)$
  - $P'_S(\sigma_{-i}) = zPr(a_{-i} \mid s,h, \sigma_{-i})P_S(\sigma_{-i})$

Theoretical Underpinnings II

- Assumptions: parameter independence, Dirichlet priors ← conjugate for the multinomial distributions we wish to learn, use of simple fictitious play models…

In brief…

- Coordination of agents activities an important focus of multiagent learning (and RL)
  - (Identical interest) stochastic games provide a useful model for studying such problems
- Much emphasis in MARL research is placed on ensuring that MARL algorithms eventually converge to desirable equilibria.

✓ We propose a Bayesian model for optimal exploration in MARL problems

Exploration costs are weighed against expected exploration benefits.

Reasoning about how one's actions will influence the behavior of others is required.