Solving MDP's using EM and Antifreeze

Thomas Furmston & David Barber

Department of Computer Science
University College London
London WC1E 6BT, UK

September 7, 2009

Outline

Markov Decision Processes

A Markov decision process is made of the following components

- n states, $\{x_1, ..., x_n\}$, m actions, $\{a_1, ..., a_m\}$, and T discrete time points.
- Transition probabilities $T_{x_{t+1},x_t,a_t} = p(x_{t+1}|x_t, a_t)$
- The policy $\pi_{a_t,x_t} = p(a_t|x_t)$
- The utility/reward function $u(x_t, a_t)$
- The discount factor $\gamma \in [0, 1]$

The system is assumed to be Markovian such that a state-action trajectory can be described by the distribution

$$p(x_{1:t}, a_{1:t}|\pi) = p(x_1)p(a_1|x_1, \pi) \prod_{\tau=1}^{t-1} p(x_{\tau+1}|x_\tau, a_\tau)p(a_\tau|x_\tau, \pi) \quad (1)$$

Markov Decision Processes

For a given policy π the total expected utility over an horizon T is given by the equation

$$u(\pi) = \sum_{t=1}^{T} \sum_{x_t, a_t} \gamma^t u(x_t, a_t) p(x_t, a_t | \pi) \tag{2}$$

- The problem is now to find the policy π^* that maximises (2).
- At present we are only solving the MDP, that is the transitions and rewards are assumed known, but the model can be extended to the RL problem.

Approaches for Policy Optimisation

- Many of the classical solution methods to MDPs, such as policy iteration, are based on fixed point solutions to the Bellman Equation.
- Policy iteration is provably convergent in time polynomial in the n, m, is intuitively easy to understand and easy to implement.
- However, it doesn't take advantage of any structure in the MDP nor can any uncertainty in the parameters be factored into the model.

Our Aims

- To make use of techniques from probabilistic inference, which will allow us to model structure and uncertainty into the domain.
- To derive an EM algorithm for solving MDPs that is equivelent to [1] but simpler and more general in its derivation.
- To extend to work of [1] to deal with deterministic policies in a rigourous way, allowing faster convergence.
- To deal with the exploration issues that arise from deterministic policies.

Our Results

- We have introduced a simplified probabilistic modelling treatment of MDPs, similar to [2].
- The algorithm for deterministic policies is an original algorithm, different from policy/value and classical algorithms.
- We have shown how to do deal with exploration problems that occur with deterministic policies and domains.

EM Construction

The construction of the EM algorithm will run as follows

- Firstly we will construct a lower bound on the quantity of interest, in this case the logarithm of the total expected utility.
- Once this bound has been established we will iterate between evaluating it (E-step) and maximising it (M-step).

Constructing the Lower Bound

To construct the lower bound on equation (2) we assume that the utility is non-negative, which allows us to form the distribution

$$\hat{p}(x_{1:t}, a_{1:t}, t|\pi) = \frac{1}{u(\pi)}\gamma^t u(x_t, a_t)p(x_{1:t}, a_{1:t}|\pi) \tag{3}$$

Introducing the auxiliary distribution $q(x_{1:t}, a_{1:t}, t)$ and taking the Kullback-Leibler divergence between q and $\hat{p}$ gives

$$KL(q||\hat{p}) = H(q) - \langle \log u_t(x_t, a_t) \rangle_q - \langle \log p(x_{1:t}, a_{1:t}|\pi) \rangle_q + \log u(\pi) \tag{4}$$

where $\langle \cdot \rangle_q$ denotes the average w.r.t. the q-distribution, and H is the entropy function. We now use the fact that the Kullback-Leibler divergence is non-negative $\forall$ distributions $\hat{p}$, q, to obtain the bound

$$\log u(\pi) \geq -H(q) + \langle \log u(x_t, a_t) \rangle_q + \langle \log p(x_{1:t}, a_{1:t}|\pi) \rangle_q \tag{5}$$

M-Step

In the M step we are interested in maximising the distribution w.r.t π, so separating out the policy terms from equation (5) we obtain the energy term

$$E(\pi) = \sum_{t=1}^{T} \sum_{\tau=1}^{t} \sum_{x_\tau, a_\tau} q(x_\tau, a_\tau, t) \log \pi_{a_\tau, x_\tau} \tag{6}$$

where we use a tabular policy $\pi_{a_\tau, x_\tau} \equiv p(a_\tau | x_\tau, \pi)$. For a stationary policy the resulting M-step may be written as

$$\pi_{a,x} \propto \sum_{t=1}^{T} \sum_{\tau=1}^{t} q(x_\tau, a_\tau, t) \tag{7}$$

If we set $T = \infty$ in equation (7) and perform some simple manipulations then one recovers the previous result of [1].

E-Step

- If we place no functional restriction on the q-distribution then the maximum occurs when $KL(q||\hat{p}) = 0$, that is when $q(x_{1:t}, a_{1:t}, t) = \hat{p}(x_{1:t}, a_{1:t}, t|\pi^{old})$.
- From the form of the M-steps we see that the E-step requires the calculation of the marginals $q(x_\tau, a_\tau, t)$.
- Straightforward since, as a graphical model, q is simply a chain distribution for which marginal inference can be achieved in linear time $O(T)$ via standard message-passing techniques [2, 1].

Maze Problem

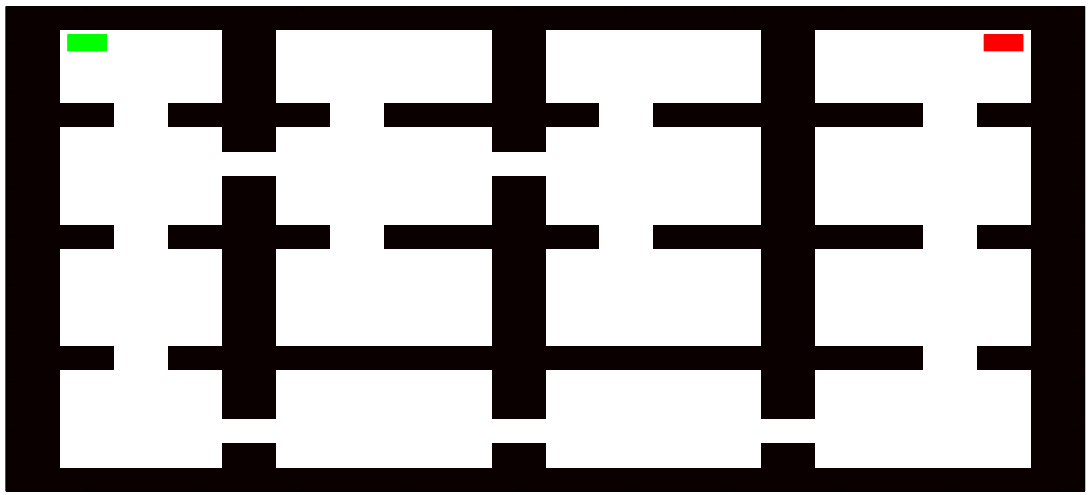

Figure: Maze considered in the MDP experiments. The walls are black, with initial state in the top left corner (green), the goal state in the top right corner (red) and the rest of the maze in white. There are in total 240 states. The agent has four actions available; up, down, right and left. If the agent moves into a wall it remains in its current state. The environment is stochastic with any action resulting in any of the other actions being performed with probability 0.05.

Maze Results

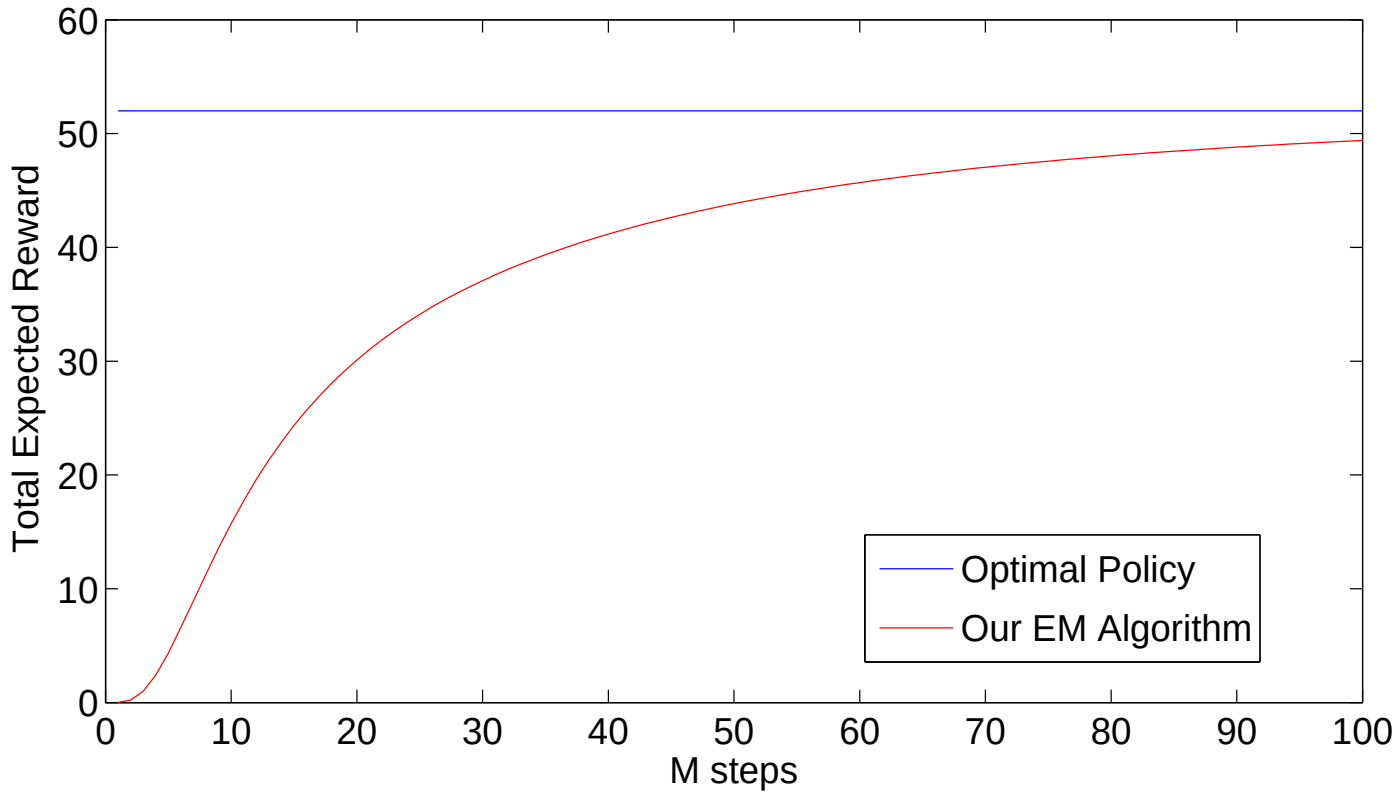

Figure: The results of the EM algorithm applied to the maze problem described previously. The algorithm can be seen to be montonically increasing to total expected reward toward the optimal.

Cut-off Time

- If T is assumed finite and known then our framework can be readily implemented.
- However, for T infinite one has to select a point at which to terminate the summation in equation (7). Although the framework doesn't currently yield a formal method, we can use the time marginal $\hat{p}(t|\pi)$ to gain an indication of a suitable cut-off point.
- To demonstrate the cut-off affect, we considered a simple maze navigation problem where the agent has to learn to traverse a maze from an initial state to a goal state. We considered the maze in figure 1 with the discount factor set to $\gamma = 0.95$ and the horizon set to $T = \infty$.

Cut-off Time - Experiment

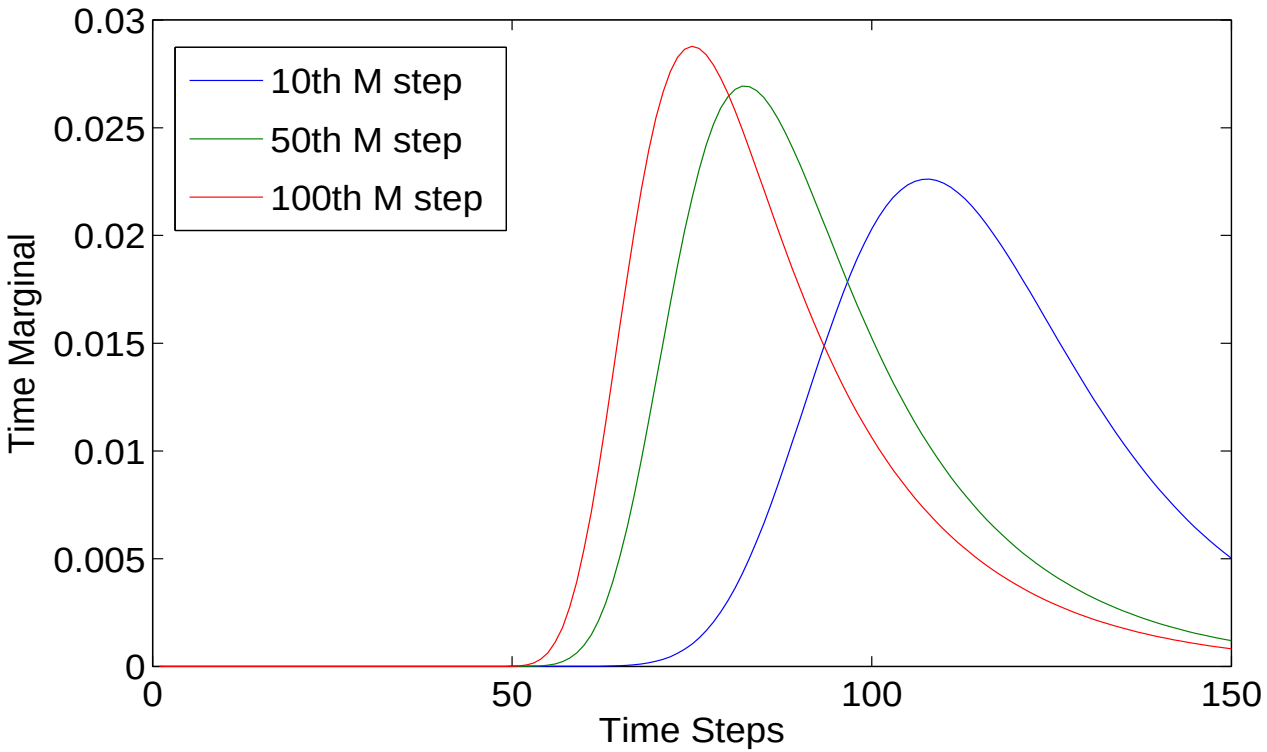

Figure: The time marginal $\hat{p}(t|\pi)$ for the maze in figure 1 at the 10^{th}, 50^{th} and the 100^{th} M-steps. After 50 M-steps a horizon $T \approx 150$ suffices, whereas in earlier M-steps a larger horizon is needed to account for the inferior policy.

EM Algorithm for Deterministic Policies

- If we look at the results in figure 2 we see that the algorithm converges montonically to the optimal policy.
- However, the stochastic nature of the policy seems to slow down the convergence rate.
- To improve the rate of convergence we restrict the algorithm to the space of deterministic policies.
- The reasons for this are that the optimal policy of an MDP is deterministic, and that similar algorithms show good convergence.

EM Algorithm for Deterministic Policies

- With this in mind we now restrict the policy space to deterministic policies, that is $\pi_{a,x} = \delta(a, a^*(x))$, and run through the same derivation as previously.
- The energy term now becomes

$$E(a^*) = \sum_{t=1}^{T} \sum_{\tau=1}^{t-1} \sum_{x_{\tau+1}, x_\tau} q(x_{\tau+1}, x_\tau, t) \log p(x_{\tau+1} | x_\tau, a^*(x_\tau)) + \sum_{t=1}^{T} \sum_{x_t} q(x_t, t) \log u(x_t, a^*(x_t)) \quad (8)$$

EM Algorithm for Deterministic Policies

For each state x we now determine the action a that maximises the energy, equation (8). This corresponds to finding for each state x the action that maximises

$$\sum_{x'} \left\{ \sum_{t=1}^{T} \sum_{\tau=1}^{t-1} q(x_{\tau+1} = x', x_\tau = x, t) \right\} \log p(x'|x, a) + \left\{ \sum_{t=1}^{T} q(x_t = x, t) \right\} \log u(x, a) \quad (9)$$

The E-step for the q-distribution is as before, except that we also require the two-time marginals $q(x_{\tau+1} = x', x_\tau = x, t)$.

EM Algorithm for Deterministic Policies - Experiment

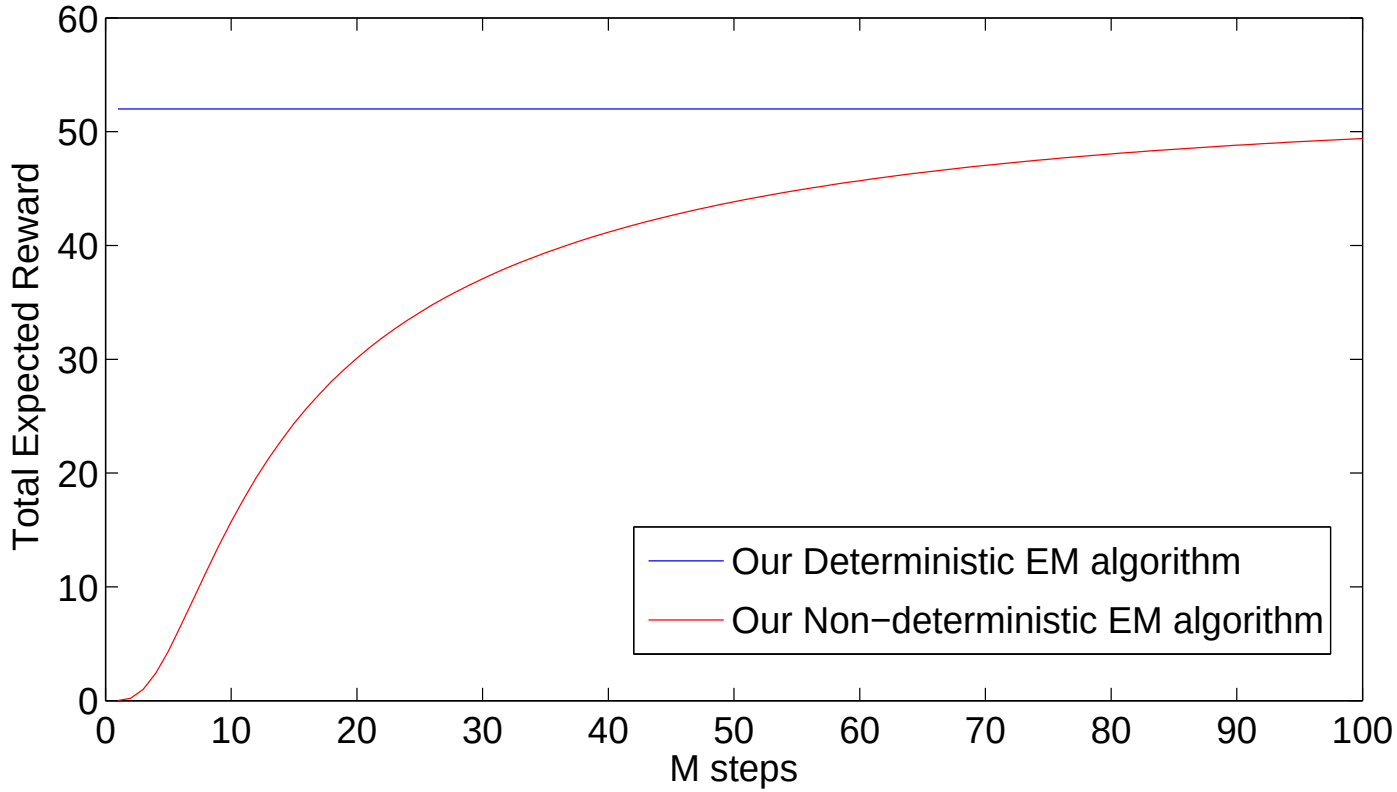

Figure: Our deterministic policy EM algorithm compared with the non-deterministic EM algorithm on the maze problem. The discount factor was set to $\gamma = 1$ and the horizon was set to $T = 100$. The algorithms each performed 100 M-steps and were initialised with the same uniform policy, the total expected utility is as given in equation (2).

EM Freezing

- The M-step updates in the EM algorithm characteristically 'freeze', in a deterministic or near-deterministic observation distribution, leading to extremely small increases in the log-likelihood.
- This problem also occurs in our EM approach when the transitions and the policy are both deterministic.
- To counter this problem it is possible to add noise to the environment, rendering it non-deterministic, and then solve the MDP in this new environment.

EM Freezing

- Explicitly, for each state we define the new transition $p_\epsilon(x'|x,a)$ as a convex combination of the transition with a distribution

$$p_\epsilon(x'|x,a) = (1-\epsilon)p(x'|x,a) + \epsilon\Gamma_x(x') \tag{10}$$

  where $\epsilon \in [0,1)$ and $\Gamma_x(x')$ is an arbitrary probability distribution and then solve the MDP $\langle \mathcal{X}, \mathcal{A}, U, p_\epsilon \rangle$.
- The idea behind this is encourage 'exploration' during the E-step and therefore enable the algorithm to escape local minima, similar to ϵ-greedy policies used in various Monte Carlo solution methods to MDPs [3].

EM Freezing - Experiment

- To illustrate the validity of the 'antifreeze' method in an MDP setting we consider the simple maze problem in figure 5.
- Since the environment is deterministic, the normal deterministic EM algorithm would perform trivial updates on this problem and freeze.
- In the experiments we use an antifreeze distribution $\Gamma_x(x')$ to be uniform for all states that satisfy the condition $\beta_T(x) = 0$. The transitions of the remaining states where left unchanged.
- In the experiment we set ϵ was set to 0.35.

EM Freezing - Maze

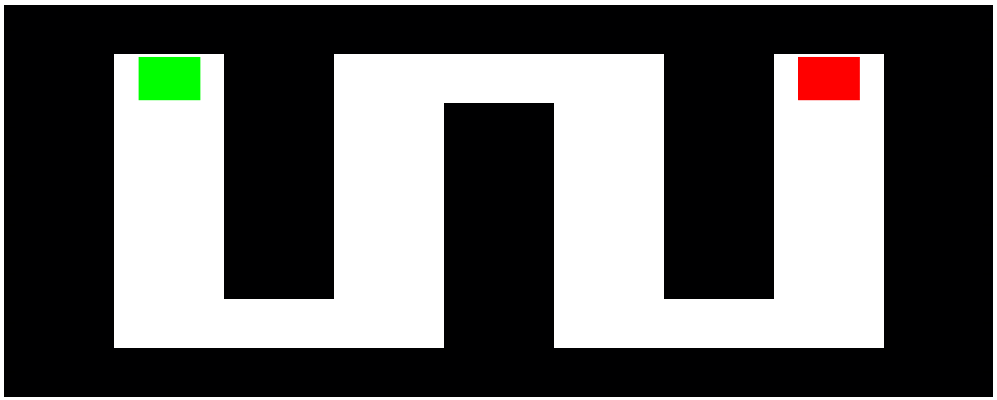

Figure: Maze considered in the deterministic transition and deterministic policy experiments. The colour scheme is the same as previously, while the discount factor was set to $\gamma = 1$ and the horizon set to $T = 40$. There are a total of 27 states. The transitions are deterministic and the policy is initialised in the worst possible way, taking an action that moves in the opposite direction of the optimal policy, *e.g.* when the agent is in the initial state it will move upwards instead of downwards.

EM Freezing - Results

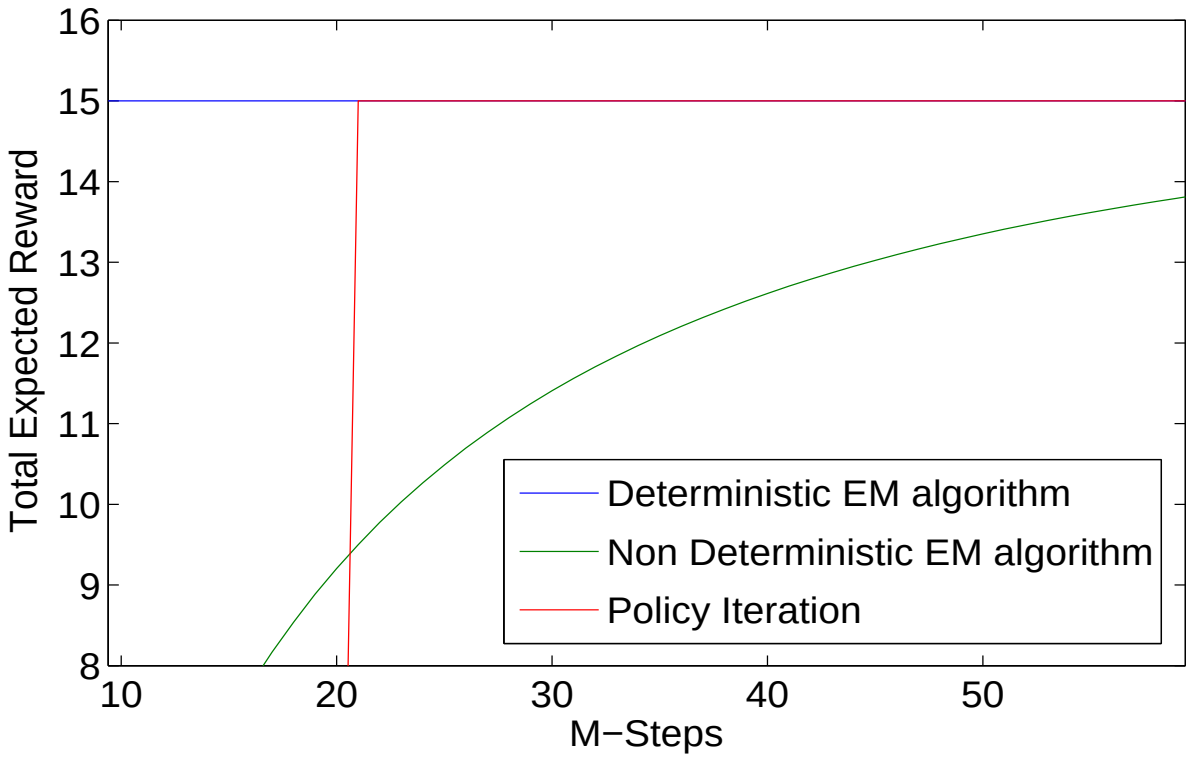

Figure: The results of the EM freezing experiment performed on our Deterministic EM algorithm with noise added to the enviroment, the non-deterministic EM algorithm and policy iteration.

Summary

- We have introduced a simple probabilistic approach to solving MDPs that is similar to [1] but simpler and more general in its derivation.
- We have extended our model to rigourously handle deterministic policies which has shown faster convergence rates.
- We have presented a novel way to introduce "exploration" into to the E-step making the process more robust to local optima.

Future

In the future we plan to extend the current framework in the following directions

- Applying the framework to structured state and action spaces.
- Use structured q approximations to deal with massive state spaces.
- Deal with the RL problem where uncertainty in the rewards and transitions is factored into the model.

 Probabilistic inference for solving (PO)MDPs - Toussaint, M. and Harmeling, S. and Storkey, A. - University of Edinburgh, School of Informatics - Research Report, 2006.

 Graphical Models, Exponential Families, and Variational Inference - Wainwright, M. J. and Jordan, M. I. - Foundations and Trends in Machine Learning, 2008.

 Reinforcement Learning: An Introduction - Sutton, R. S. and Barto, A. G. - MIT Press, 1998.