Variational methods for Reinforcement Learning

Thomas Furmston **David Barber**
Computer Science Department, University College London, London WC1E 6BT, UK.

Abstract

We consider reinforcement learning as solving a Markov decision process with unknown transition distribution. Based on interaction with the environment, an estimate of the transition matrix is obtained from which the optimal decision policy is formed. The classical maximum likelihood point estimate of the transition model does not reflect the uncertainty in the estimate of the transition model and the resulting policies may consequently lack a sufficient degree of exploration. We consider a Bayesian alternative that maintains a distribution over the transition so that the resulting policy takes into account the limited experience of the environment. The resulting algorithm is formally intractable and we discuss two approximate solution methods, Variational Bayes and Expectation Propagation.

1 Introduction

Reinforcement Learning (RL) is the problem of learning to act optimally through interaction and simulation in an unknown environment (Sutton and Barto, 1998) and may be applied to sequential decision problems where the underlying dynamics of the environment is unknown, for example helicopter control (Abbeel et al., 2007), the cart-pole problem (Rasmussen and Deisenroth, 2008) and elevator scheduling (Crites and Barto, 1995). We assume a model-based approach for which we need to estimate the parameters of the transition model based on limited interaction with the environment. A classical approach to learning an environment model is to use a point estimator, such as the maximum likelihood estimator. However, these may result in myopic policies since only the known observed transitions are assumed possible. As an alternative, we describe a Bayesian approach in which a prior distribution is placed over the environment model and updated as data from the environment is received. This environment distribution maintains the possibility of transitions to parts of the space that have not yet been observed but nevertheless may prove rewarding. The optimal policy is then obtained by integrating over all possible environment models. To deal with the difficulties of carrying out this integral we discuss two approximate methods, Variational Bayes (VB) (see for example (Beal and Ghahramani, 2003)) and Expectation Propagation (EP) (Wainwright and Jordan, 2008; Minka, 2001). For simplicity of exposition, we assume throughout that the reward model is known, but that the transition model needs to be learned from experience. Extending the approach to an unknown reward model is essentially straightforward.

2 Variational MDPs

An MDP can be described by an initial state distribution $p_1(s_1)$, transition distributions $p(s_{t+1}|s_t, a_t)$, and a reward function $r_t(s_t, a_t)$, where the state and action at time t are denoted by s_t and a_t respectively. For a discount factor γ the reward is defined as $r_t(s_t, a_t) = \gamma^{t-1} r(s_t, a_t)$ for a stationary reward $r(s_t, a_t)$. We assume a stationary policy, π, defined as a set of conditional distributions over the action space[1], $\pi_{a,s} = p(a_t = a|s_t = s, \pi)$. The total expected reward of the MDP (the policy utility) is

$$U(\pi) = \sum_{t=1}^{H} \sum_{s_t, a_t} r_t(s_t, a_t) p(s_t, a_t|\pi) \qquad (1)$$

where H is the horizon, which can be either finite or infinite, and $p(s_t, a_t|\pi)$ is the marginal of the joint state-

Appearing in Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

[1] More generally, one may consider policies which depend on the belief, $\pi_{a,s,\mathcal{D}} = p(a|s, p(\theta|\mathcal{D}), \pi)$, similar to the encoding of RL as a POMDP(Duff, 2002), though we leave this case for future study.

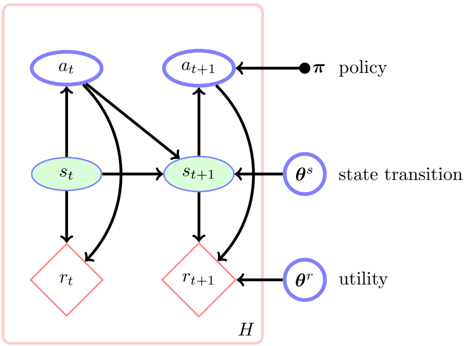

Figure 1: RL represented as a model-based MDP transition and policy learning problem. Rewards depend on the current and past state and the past action, $r_t(s_t, a_t)$. The policy $p(a_t|s_t, \pi)$ determines the decision and the environment is modeled by the transition $p(s_{t+1}|s_t, a_t)$. Based on a history of actions, states and reward, the task is maximize the expected summed rewards with respect to the policy $\boldsymbol{\pi}$. In the MDP setup, the state transition and utilities are known; in our RL setup we have a distribution over these quantities.

action trajectory distribution

$$p(s_{1:H}, a_{1:H}|\pi) = p(a_H|s_H, \pi)p_1(s_1) \times \prod_{t=1}^{H-1} p(s_{t+1}|s_t, a_t)p(a_t|s_t, \pi). \quad (2)$$

In this paper we consider the episodic case, so that the horizon is finite. Graphically we can represent this using an influence diagram, figure 1. Given a transition model $p(s_{t+1}|s_t, a_t)$, the MDP learning problem is to find a policy π that maximizes (1). By expressing the utility (1) as the likelihood function of an appropriately constructed mixture model the MDP can be solved using techniques from probabilistic inference, such as EM (Toussaint et al., 2006) or MCMC (Hoffman et al., 2008). We follow a construction equivalent to (Toussaint et al., 2006) but which has the advantage of not requiring auxiliary variables, see *e.g.* (Dayan and Hinton, 1997; Kober and Peters, 2009; Furmston and Barber, 2009). Without loss of generality, we assume the reward is non-negative and define the reward weighted path distribution

$$\hat{p}(s_{1:t}, a_{1:t}, t|\pi) = \frac{r_t(s_t, a_t)p(s_{1:t}, a_{1:t}|\pi)}{U(\pi)} \quad (3)$$

This distribution is properly normalised, as can be seen from (1) and (2). We now define a variational distribution $q(s_{1:t}, a_{1:t}, t)$, and take the Kullback-Leibler divergence between the q-distribution and (3). Since

$$\mathrm{KL}(q(s_{1:t}, a_{1:t}, t)||\hat{p}(s_{1:t}, a_{1:t}, t|\pi)) \geq 0 \quad (4)$$

we obtain a lower bound on the log utility

$$\log U(\pi) \geq H(q(s_{1:t}, a_{1:t}, t)) + \langle \log \hat{p}(s_{1:t}, a_{1:t}, t|\pi)\rangle_q \quad (5)$$

where $\langle\cdot\rangle_q$ denotes the average *w.r.t.* $q(s_{1:t}, a_{1:t}, t)$ and $H(\cdot)$ is the entropy function. An EM algorithm can be obtained from the bound in (5) by iterative coordinate-wise maximisation:

E-step For fixed π^{old} find the best q that maximises the *r.h.s.* of (5). For no constraint on q, this gives $q = \hat{p}(s_{1:t}, a_{1:t}, t|\pi^{\text{old}})$.

M-step For fixed q find the best π that maximises the *r.h.s.* of (5). This is equivalent to maximising the 'energy' $\langle \log \tilde{p}(s_{1:t}, a_{1:t}, t|\pi)\rangle_q$ *w.r.t.* π.

Maximisation of the energy term *w.r.t.* π, under the constraint that the policy is a distribution, gives

$$\pi^{new}_{a,s} \propto \sum_{t=1}^{H}\sum_{\tau=1}^{t} q(s_\tau = s, a_\tau = a, t) \quad (6)$$

For this M-step the required marginals of the q-distribution can be calculated in linear time using message passing since the distribution is chain structured (Wainwright and Jordan, 2008). The EM algorithm is run until the policy converges to a (possibly local) optima.

3 Variational Reinforcement Learning

In the RL problem we assume the transition distributions $\boldsymbol{\theta}$ formed from $\theta^{s'}_{s,a} = p(s'|s, a)$ are unknown and need to be estimated on the basis of interaction with the environment. These interactions are observed transitions $\mathcal{D} = \{(s_n, a_n) \to s_{n+1}, n = 1, \ldots, N\}$. A classical approach it is to use a point estimate of the transition model, such as the maximum likelihood (ML) estimator. However, for small amounts of observed transitions, these estimators harshly assume that unobserved transitions will simply never occur. Such an over-confident estimate can adversely affect the overall policy solution and result in myopic policies that are unaware of potentially beneficial state-action pairs. Whilst this over-confidence can be ameliorated by adding pseudo-counts, this still does not reflect the uncertainty in the estimate of the transition.

We propose an alternative Bayesian solution that maintains a distribution over transitions. The posterior of $\boldsymbol{\theta}$ is formed from Bayes' rule

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \quad (7)$$

As $\boldsymbol{\theta}$ is a set of independent categorical distributions a natural conjugate prior $p(\boldsymbol{\theta})$ is the product of independent Dirichlet distributions, *i.e.*

$$p(\boldsymbol{\theta}) \sim \prod_{s,a} \mathrm{Dir}(\theta^{\cdot}_{s,a}|\alpha^{\cdot}_{s,a}) \tag{8}$$

where $\boldsymbol{\alpha}$ are hyper-parameters. This gives a posterior

$$p(\boldsymbol{\theta}|\mathcal{D}) = \prod_{s,a} \mathrm{Dir}(\theta^{\cdot}_{s,a}|c^{\cdot}_{s,a} + \alpha^{\cdot}_{s,a}) \tag{9}$$

where c is the count of observed transitions:

$$c^{s'}_{s,a} = \sum_{n=1}^{N} \mathbb{I}\left[s_n = s, a_n = a, s_{n+1} = s'\right] \tag{10}$$

The task now is to find the policy that maximizes the expected utility given the environmental data

$$U(\pi|\mathcal{D}) = \int U(\pi|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \tag{11}$$

where $U(\pi|\boldsymbol{\theta})$ is given by (1) with transitions $\boldsymbol{\theta}$.

Our aim is to form an EM style approach to learning π. Assuming the reward is non-negative we construct a probability distribution for which the normalization constant is equal to (11). Consider the following unnormalised distribution defined over state-action paths and times $t = 1, ..., H$,

$$\tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi) = r(s_t, a_t)p(s_{1:t}, a_{1:t}|\boldsymbol{\theta}, \pi) \tag{12}$$

where $p(s_{1:t}, a_{1:t}|\boldsymbol{\theta}, \pi)$ is the marginal of (2) given the transitions $\boldsymbol{\theta}$. Using (12) we now define a joint distribution over state-action paths, times and transitions

$$\hat{p}(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta}|\pi, \mathcal{D}) = \frac{\tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi)p(\boldsymbol{\theta}|\mathcal{D})}{U(\pi|\mathcal{D})} \tag{13}$$

This distribution is properly normalised, which can be verified through use of (1) and (11). The Kullback-Leibler divergence between a variational distribution $q(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta})$, and (13) gives the bound

$$\mathrm{KL}(q(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta})||\hat{p}(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta}|\pi, \mathcal{D})) \geq 0 \tag{14}$$

from which we obtain

$$\begin{aligned}\log U(\pi|\mathcal{D}) \geq H(q(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta})) + \left\langle \log p(\boldsymbol{\theta}|\mathcal{D})\right\rangle_q \\ + \left\langle \log \tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi)\right\rangle_q\end{aligned} \tag{15}$$

where $\langle\cdot\rangle_q$ denotes the average *w.r.t.* $q(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta})$. An EM algorithm for optimising the bound with respect to π is:

E-step For fixed π^{old} find the best q that maximises the *r.h.s.* of (15). For no constraint on q, this gives $q = \hat{p}(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta}|\pi^{\mathrm{old}}, \mathcal{D})$.

M-step For fixed q find the best π that maximises the *r.h.s.* of (15). This is equivalent to maximising the 'energy' $\left\langle \log \tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi)\right\rangle_q$ *w.r.t.* π.

To perform the M-step we need the maximum of $\left\langle \log \tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi)\right\rangle_q$ *w.r.t.* π. As the policy is independent of the transitions this maximisation gives updates of the form

$$\pi^{new}_{a,s} \propto \sum_{t=1}^{H}\sum_{\tau=1}^{t} q(s_\tau = s, a_\tau = a, t) \tag{16}$$

Calculating the policy update is now a matter of calculating the marginals of the q-distribution from the previous E-step. If no functional restriction is placed on the q-distribution then it will take the form of (13), where π will equal the policy of the previous M-step. However, examining the form of (13), the exact state-action marginals of this distribution are computationally intractable. This can be understood by first carrying out the integral over $\boldsymbol{\theta}$, which has the effect of coupling together all time slices of the path distribution $\hat{p}(s_{1:t}, a_{1:t}, t)$.

In the following we discuss two approaches to dealing with this intractability. The first, Variational Bayes, restricts the functional form of the q-distribution in the E-step such that the updates in the M-step become tractable. The second approximates the marginals of the q-distribution directly using Expectation Propagation.

4 Variational Bayes

To ensure computational tractability, a suitable restriction on the functional form of the q-distribution is to make the factorised approximation:

$$q(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta}) = q(s_{1:t}, a_{1:t}, t)q(\boldsymbol{\theta}). \tag{17}$$

This approximation maintains the lower bound in (15) which now takes the form

$$\begin{aligned}\log U(\pi|\mathcal{D}) \geq H(q_{\boldsymbol{x}}) + H(q_{\boldsymbol{\theta}}) + \left\langle \log p(\boldsymbol{\theta}|\mathcal{D})\right\rangle_{q_{\boldsymbol{\theta}}} \\ + \left\langle \log \tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi)\right\rangle_{q_{\boldsymbol{\theta}} q_{\boldsymbol{x}}}\end{aligned} \tag{18}$$

Where we have used the notation $q_{\boldsymbol{\theta}} \equiv q(\boldsymbol{\theta})$, and $q_{\boldsymbol{x}} \equiv q(s_{1:t}, a_{1:t}, t)$. The variational Bayes procedure now iteratively maximizes (18) with respect to the distributions $q_{\boldsymbol{x}}$ and $q_{\boldsymbol{\theta}}$. Taking the functional derivative of (18) with respect to $q_{\boldsymbol{x}}$ and $q_{\boldsymbol{\theta}}$, whilst holding the other fixed, gives the following update equations:

$$q(s_{1:t}, a_{1:t}, t) \propto e^{\left\langle \log \tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi)\right\rangle_{q_{\boldsymbol{\theta}}}} \tag{19}$$

$$q(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \mathcal{D})e^{\left\langle \log \tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi)\right\rangle_{q_{\boldsymbol{x}}}} \tag{20}$$

Algorithm 1 VB EM Algorithm

```
Input: policy $\pi$, reward $r$, prior $\boldsymbol{\alpha}$ and transition counts $c$.
repeat
  For fixed policy $\pi$
  repeat
    Calculate the $q$-marginals (21) and (23).
  until Convergence of the marginals.
  Update the policy according to (16).
until Convergence of the policy.
```

Expansion of the $\log \tilde{p}(s_{1:t}, a_{1:t}, t|\boldsymbol{\theta}, \pi)$ term in (19) shows that $q(s_{1:t}, a_{1:t}, t)$ is proportional to

$$r(s_t, a_t)\pi_{a_t,s_t} p_1(s_1) \prod_{\tau=1}^{t-1} e^{\left\langle \log \theta_{s_{\tau+1},s_\tau,a_\tau} \right\rangle_{q_{\boldsymbol{\theta}}}} \pi_{a_\tau,s_\tau}. \quad (21)$$

This is the same form as the original MDP (1,2) with the transitions $\boldsymbol{\theta}$ replaced with unnormalised transitions

$$\tilde{\theta}(s', s, a) \equiv e^{\left\langle \log \theta_{s',s,a} \right\rangle_{q_{\boldsymbol{\theta}}}}. \quad (22)$$

The averages of $\log \theta$ in the exponent can be computed using standard digamma functions. Given $q_{\boldsymbol{\theta}}$, the marginals $q(s_\tau, a_\tau, t)$ can be then calculated using message passing on the corresponding factor graph (Kschischang et al., 2001).

A similar calculation for the transition parameters gives the update

$$q(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \mathcal{D}) e^{\sum_{t=1}^{H} \sum_{\tau=1}^{t-1} \left\langle \log \theta_{s_{\tau+1},s_\tau,a_\tau} \right\rangle_{q_{\boldsymbol{x}}}}$$

The summation of the states and actions in the exponent means that we may write

$$q(\boldsymbol{\theta}) = \prod_{s,a} \mathrm{Dir}\left(\theta^{\cdot}_{s,a} | \alpha^{\cdot}_{s,a} + c^{\cdot}_{s,a} + \tilde{r}^{\cdot}_{s,a}\right) \quad (23)$$

where

$$\tilde{r}^{s'}_{s,a} = \sum_t \sum_\tau q(s_{\tau+1} = s', s_\tau = s, a_\tau = a) \quad (24)$$

Equation (23) has an intuitive interpretation: for each triple (s', s, a) we have the prior $\alpha^{s'}_{s,a}$ term and the observed counts $c^{s'}_{s,a}$ which deal with the posterior of the transitions. The term $\tilde{r}^{s'}_{s,a}$ encodes an approximate expected reward obtained from starting in state s, taking action a, entering state s' and then following π afterwards. The posterior $q(\boldsymbol{\theta})$ is therefore a standard Dirichlet posterior on transitions but biased towards transitions that are likely to lead to higher expected reward. Under the approximation (17) the E-step consists of calculating the distributions (21) and (23). As these distributions are coupled we need to iterate them until convergence.

The form of the M-step is calculated by maximising the bound (18) with respect to π. This leads to the same updates as (16) except the q-distribution now takes the form of (21). A summary of VB-EM is given in algorithm (1).

4.1 Hierarchical Variational Bayes

So far we have assumed that the hyper-parameters, $\boldsymbol{\alpha}$, are fixed. However the quality of the policy learned can be strongly dependent on $\boldsymbol{\alpha}$. If the components of $\boldsymbol{\alpha}$ are set too low any initial data points will dominate the transition posterior and the probability of unobserved transitions will be small. On the other hand if $\boldsymbol{\alpha}$ is set too high an excessively large amount of data points will be required to dilute the prior effect on the posterior. To overcome this problem we can extend the model by placing a prior distribution over $\boldsymbol{\alpha}$ and then update the posterior as data from the environment is received. This extension is straightforward under the variational approximation $q_{\boldsymbol{x}} q_{\boldsymbol{\theta}} q_{\boldsymbol{\alpha}}$. In our experiments we use the hyper-parameter distribution independently for each component of $\boldsymbol{\alpha}$:

$$p(\alpha) \propto e^{-20(\alpha-1)^2}, \qquad \alpha \geq 0$$

which has the effect of retaining significant posterior variance in the transition model, damping overly greedy exploitation.

5 Expectation Propagation

In order to implement the Variational Reinforcement Learning approach of §3 we require the marginals of the intractable distribution $q = \hat{p}(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta}|\pi^{\mathrm{old}}, \mathcal{D})$. As an alternative to the variational Bayes factorised approach we here consider an approximate message passing (AMP) approach that approximates the required marginals directly.

The graphical structure of $q(s_{1:t}, a_{1:t}, \boldsymbol{\theta}, t)$ is loopy but sparse, so that a sum-product algorithm may provide reasonable approximate marginals, see figure 2. The messages for the factor graph version of the sum-product algorithm take the following form.

$$\mu_{x \to f}(x) = \prod_{h \in n(x) \setminus \{f\}} \mu_{h \to x}(x) \quad (25)$$

$$\mu_{f \to x} = \sum_{\sim\{x\}} f(X) \prod_{y \in n(f) \setminus \{x\}} \mu_{y \to f}(y) \quad (26)$$

where $\sum_{\sim\{x\}}$ means the sum over all variables except x, $n(\cdot)$ is the set of neighbouring nodes and X are the

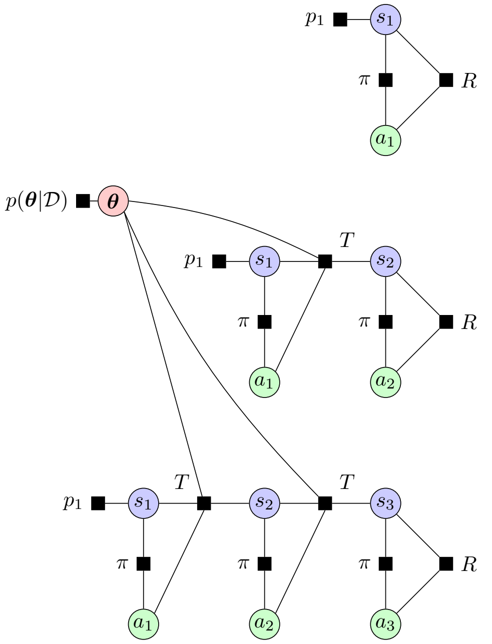

Figure 2: A factor graph representation of $q(s_{1:t}, a_{1:t}, t, \boldsymbol{\theta})$ for transition factors T, reward factors R and policy factor π, for a $H = 3$ horizon. The square nodes represent the various factors (functions) of the distribution and the circle nodes represent the variables. The initial time has no transition. The t^{th} chain is the t^{th} row of this diagram for fixed θ.

variables of the factor f. At convergence the singleton marginals are approximated by

$$p(x) = \prod_{f \in F_x} \mu_{f \to x}(x) \tag{27}$$

where F_x means the set of functions in the factor graph that depend on x. As can be seen from (26) and (25) all the messages that involve the factors p_1, π, and R are trivial, requiring only summations of discrete functions. Also, as the factor node $p(\boldsymbol{\theta}|\mathcal{D})$ is a leaf node this message is also trivial. However, the messages between $\boldsymbol{\theta}$ and the transition factors T are intractable. To see this we examine a message from T to an action node a[2]

$$\mu_{T \to a}(\mathsf{a}) = \sum_{\mathsf{s},\mathsf{s}'} \mu_{s \to T}(\mathsf{s}) \mu_{s \to T}(\mathsf{s}') \int d\boldsymbol{\theta} \mu_{\boldsymbol{\theta} \to T}(\boldsymbol{\theta}) \theta^{\mathsf{s}'}_{\mathsf{a},\mathsf{s}}. \tag{28}$$

In order for (28) to be tractable we need $\mu_{\boldsymbol{\theta} \to T}(\boldsymbol{\theta})$ to be the product of independent Dirichlet's. However, using (25) we have that $\mu_{\boldsymbol{\theta} \to T}(\boldsymbol{\theta})$ takes the form

$$\mu_{\boldsymbol{\theta} \to T}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathcal{D}) \prod_{T' \neq T} \mu_{T' \to \boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{29}$$

where $\mu_{T' \to \boldsymbol{\theta}}(\boldsymbol{\theta})$ is given by

$$\mu_{T' \to \boldsymbol{\theta}}(\boldsymbol{\theta}) = \sum_{\mathsf{s}',\mathsf{a},\mathsf{s}} \mu_{a \to T'}(\mathsf{a}) \mu_{s \to T'}(\mathsf{s}) \mu_{s' \to T'}(\mathsf{s}') \theta^{\mathsf{s}'}_{\mathsf{s},\mathsf{a}}. \tag{30}$$

From (29) and (30), $\mu_{\boldsymbol{\theta} \to T}(\boldsymbol{\theta})$ is a mixture of Dirichlet's where the number of mixtures is exponential in the planning horizon H. This makes messages such as (28) computationally intractable. Following the general approach outlined in (Minka, 2001) to make a tractable approximate implantation we therefore project the messages $\mu_{T \to \boldsymbol{\theta}}(\boldsymbol{\theta})$ to a product of independent Dirichlet's by moment matching. Given the projection $\tilde{q}(\boldsymbol{\theta})$ we use (26) and (27) to obtain the approximate message

$$\tilde{\mu}_{T \to \boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\tilde{q}(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D}) \prod_{T' \neq T} \tilde{\mu}_{T' \to \boldsymbol{\theta}}(\boldsymbol{\theta})}. \tag{31}$$

Given a message initialisation and a message passing schedule $\mathcal{S}$, the AMP algorithm can be summarized as in algorithm (2). For our experiments we used the schedule $\mathcal{S}$ outlined in algorithm (3).

Algorithm 2 AMP EM Algorithm

```
Input: policy $\pi$, reward $r$, prior $\boldsymbol{\alpha}$, transition counts $c$ and message-passing schedule $\mathcal{S}$.
repeat
  For fixed policy $\pi$
  repeat
    Perform message-passing according to $\mathcal{S}$ using EP to approximate messages $\mu_{T' \to \boldsymbol{\theta}}(\boldsymbol{\theta})$.
  until Convergence of the messages.
  Update the policy according to (16).
until Convergence of the policy.
```

6 Experiments

6.1 Incorporation of uncertainty

The first experiment is designed to demonstrate that our objective function indeed incorporates uncertainty in the knowledge of the environment into the policy optimisation process. The experiment is performed on a problem small enough that for short horizons the objective function (11) and the EM update (16) can be calculated exactly. This allows for characteristics of the objective function to be gleaned without the complicating issue of approximations.

[2] We have dropped the time dependence on the factors and the variables to ease the notation.

Algorithm 3 AMP message-passing Schedule

```
repeat
  for $t = 1$ to $H$ do
    Perform message-passing along the $t^{th}$ chain, $q(s_1, a_1, ..., s_t, a_t)$, figure 2, holding all the messages $\mu_{\boldsymbol{\theta} \to T}(\boldsymbol{\theta})$ fixed.
  end for
  repeat
    for each $\mu_{T \to \boldsymbol{\theta}}(\boldsymbol{\theta})$ do
      Perform Expectation-Propagation to obtain $q(\boldsymbol{\theta})$, then use (31) to update $\mu_{T \to \boldsymbol{\theta}}(\boldsymbol{\theta})$.
    end for
  until Convergence of all the messages $\mu_{T \to \boldsymbol{\theta}}(\boldsymbol{\theta})$.
until Convergence of the $q$-distribution.
```

The experiment was performed on a toy two-state problem, with the transition and reward matrices given in figure 3. The horizon was set to $H = 5$ and the initial state is 1. The aim of the experiment is to compare the average total expected utility of the policies obtained from the Bayesian and point-based objective functions. The average is taken over the true transition model, θ_{true}, and we compare these averages for increasing numbers of observed transitions, N. We set the distribution over the true transition model to be uniform. Writing the quantities of interest down algebraically we have for the Bayesian objective function

$$\mathbb{E}_{p(\theta_{\text{true}})}[\mathbb{E}_{p(\mathcal{D}|\theta_{\text{true}},N)}[U(\hat{\pi}^{\mathcal{D}}|\theta_{\text{true}})]] \\ = \int d\theta_{\text{true}} d\mathcal{D} U(\hat{\pi}^{\mathcal{D}}|\theta_{\text{true}})) p(\mathcal{D}|\theta_{\text{true}}, N) p(\theta_{\text{true}}) \tag{32}$$

where $\hat{\pi}^{\mathcal{D}}$ is the optimal policy of the Bayesian objective function. For the ML objective function we have

$$\mathbb{E}_{p(\theta_{\text{true}})}[\mathbb{E}_{p(\hat{\pi}^{\text{ML}}|\theta_{\text{true}},N)}[U(\hat{\pi}^{\text{ML}}|\theta_{\text{true}})]] \\ = \int d\theta_{\text{true}} d\hat{\pi}^{\text{ML}} U(\hat{\pi}^{\text{ML}}|\theta_{\text{true}}) p(\hat{\pi}^{\text{ML}}|\theta_{\text{true}}, N) p(\theta_{\text{true}}) \tag{33}$$

where similarly $\hat{\pi}^{\text{ML}}$ is the optimal policy of the ML objective function.

As we can calculate the objective function $U(\pi|\mathcal{D})$ exactly, we can also calculate (32) for reasonable values of N. It remains to calculate (33), where the difficult term is the probability distribution over the optimal policy, which we now detail.

$$T_i = \begin{bmatrix} \theta_i & 1 - \theta_i \\ 1 - \theta_i & \theta_i \end{bmatrix}, \qquad R = \begin{bmatrix} 4 & 10 \\ 1 & 1 \end{bmatrix}$$

Figure 3: The transition and reward matrices for the two-state toy problem. T_i represents the transition matrix from state s_i, where the columns correspond to actions and the rows correspond to the next state. The reward matrix R is defined so that the actions run along the rows and the states run along the columns.

The settings of the reward matrix and the horizon are such that, given (θ_1, θ_2) are known, the optimal action in state s_2 is a_1 for all values of θ_2. This means that when the transition dynamics are known the optimal policy can be given by a single parameter, $\hat{\pi}_{s_1,a_1}$. In the experiment we set $\theta_1 = \theta_2 = \theta$, so that $\hat{\pi}_{s_1,a_1} = 1$ when $\theta < \hat{\theta}$, and $\hat{\pi}_{s_1,a_1} = 0$ otherwise, where $\hat{\theta} = 0.7021$. The fact that we know the point, $\hat{\theta}$, at which the optimal policy of the MDP changes means that we can form a distribution of $\hat{\pi}^{\text{ML}}_{s_1,a_1}$. Given the sample size and the true value of the transition parameter we have the distribution

$$p(\hat{\pi}^{\text{ML}}_{s_1,a_1} = 1|N, \theta_{\text{true}}) = \sum_{\{n \le N | n/N < \hat{\theta}\}} B_{N,\theta_{\text{true}}}(n)$$

where $B_{N,\theta_{\text{true}}}$ is the density function of the Binomial distribution with parameters $(N, \theta_{\text{true}})$. Having obtained the distribution over the optimal policy it is now possible to calculate (33).

We calculated (32) and (33) for increasing values of the N, the results of which are shown in figure 4. It can be observed that the Bayesian objective function consistently outperforms the point-based objective function. We expect a more dramatic difference in larger problems for which the amount of uncertainty in the transition parameters is greater.

It should be noted that while the point-based objective function will always produce a deterministic policy the Bayesian objective function can produce a stochastic policy. This naturally incorporates an explorative type behaviour into the policy that will lead to a reduction in the uncertainty in the environment.

6.2 The chain problem

We compare the EM RL algorithms on the standard 'chain' benchmark RL problem (Dearden et al., 1998) which has 5 states each having 2 possible actions, as shown in figure 5. The initial state is 1 and every action is flipped with 'slip' probability $p_{\text{slip}} = 0.2$, making the environment stochastic. The optimal policy is to travel down the chain towards state 5, which is achieved by always selecting action 'a'.

In the experiments the total 1000 time-steps are split into 10 episodes each of 100 time-steps. During each episode the policy and transition model are fixed, and the transitions and rewards from the RL environment are collated. At the end of each episode the policy and transition model are updated. All policies are ini-

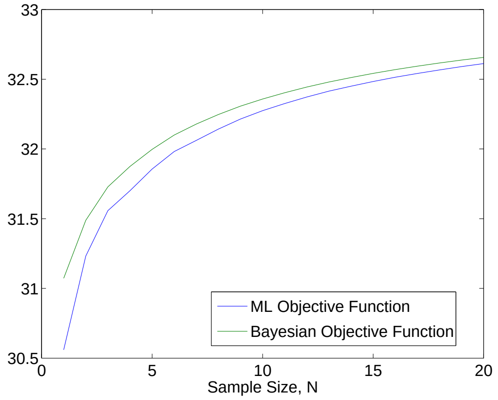

Figure 4: The average total expected reward of the policies obtained from the Bayesian objective function, $U(\pi|\mathcal{D})$, and the maximum likelihood objective function, $U(\pi|\boldsymbol{\theta}_{\mathrm{ML}})$. The sample size is plotted against the average total expected reward.

tialised randomly from a uniform distribution. For the methods based on a fixed hyper-parameters $\boldsymbol{\alpha}$, we set each component of $\boldsymbol{\alpha}$ to 1.

Convergence of all MDP solvers was determined when the L_1 norm of the policy between successive iterations is less than 0.01. The methods we compared are described below.

ML EM The mean $\boldsymbol{\theta}$ is computed from the Dirichlet posterior $p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\alpha})$. This is used as a point-based estimate of the transition model in the MDP EM algorithm of §2.

SEM At the end of each episode we obtained an approximation to the optimal policy using sampling. We first draw samples $\boldsymbol{\theta}_i, i = 1, \ldots, I$ from the posterior $p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\alpha})$. For each sample $\boldsymbol{\theta}_i$ we then compute the exact conditional marginals $\hat{p}(s_\tau, a_\tau, t|\boldsymbol{\theta}_i)$ by message passing on the chain. Averaging over the samples gives the Stochastic EM update

$$\pi^{new}_{s,a} \propto \sum_{i=1}^{I} \sum_{t=1}^{H} \sum_{\tau=1}^{t} \hat{p}(s_\tau, a_\tau, t|\boldsymbol{\theta}_i)$$

In the experiments we set I so that this method has roughly the same runtime as the AMP EM algorithm.

VB EM At the end of each episode the approach described in §4 is used. The hyper-parameter $\boldsymbol{\alpha}$ is fixed throughout to 1.

AMP EM At the end of each episode, the approach described in §5 is run, which approximates the marginal statistics required for EM learning using Expectation Propagation.

HVB-EM As for VB-EM but extended to the hyper-parameter distribution, as described in §4.1.

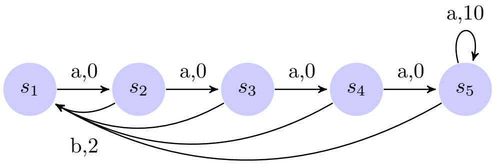

Figure 5: The single-chain problem state-action transitions with rewards $r(s_t, a_t)$. The initial state is state 1. There are two actions a, b, with each action being flipped with probability 0.2.

The results, averaged over a 100 experiments, are shown in figure 6. The AMP and stochastic EM algorithms consistently outperform the ML EM algorithm. This is in agreement with our previous results and suggests that both of these algorithms are able to make reasonable approximations to the true marginals of the q-distribution. Despite the encouraging initial performance of the variational Bayes algorithms, the ML EM algorithm eventually performs better than both the fixed hyper-parameter and hierarchical VB variants. This suggests that the factorised approximation inherent in the VB leads to difficulties. One potential issue is that under the factorisation assumptions, the unnormalised transitions (22) have the form

$$\tilde{\theta}^{s'}_{s,a} = \frac{e^{\Psi(\alpha^{s'}_{s,a})}}{e^{\Psi(\sum_{s'} \alpha^{s'}_{s,a})}} \tag{34}$$

where Ψ represents the digamma function. For $\sum_{s'} \tilde{\theta}^{s'}_{s,a} < 1$ the contributions of the first time points in the unnormalised distribution (21) exponentially dominate. As a result there is a bias towards the initial time-steps, forcing both of the variational Bayes algorithms to focus on only locally optimal policies. Finally we note that the prior on the hyper-parameters, $\boldsymbol{\alpha}$, was beneficial to the variational Bayes algorithm. This is unsurprising since it maintains posterior variance. We would expect a similar improvement in performance for a hierarchical Expectation Propagation approach.

In the variational Bayes algorithm the q-distributions had to be iterated around 15 times on average. The approximate message passing algorithm had to repeat the message passing schedule around 10 times on average, where the Expectation Propagation section of the schedule had to be repeated around 2 times for convergence. Under the current implementation the

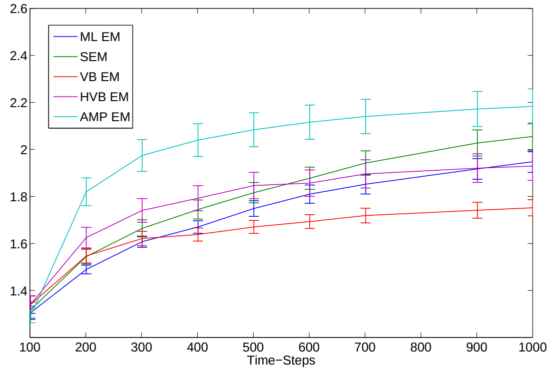

Figure 6: Results from the chain problem in figure 5 with average reward $\frac{1}{t}\sum_{\tau}^{t} r_\tau$ plotted against time t. The plot shows the results for approximate message passing (light blue), hierarchical variational Bayes (purple), variational Bayes (red), stochastic EM (green) and and the EM algorithm of §2 using the maximum likelihood estimator (dark blue). The results represent performance averaged over 100 runs of the experiment.

variational Bayes algorithm is able to perform an EM step in approximately 0.15 seconds, while the approximate message passing algorithm takes approximately 5 seconds.

7 Conclusions

Framing Markov Decision Problems as inference in a related graphical model has been recently introduced and has the potential advantage that methods in approximate inference can be exploited to help overcome difficulties associated with classical MDP solvers in large-scale problems. In this work, we performed some groundwork theory that extends these techniques to the case of reinforcement learning in which the parameters of the MDP are unknown and need to be learned from experience. An exact implementation of such a Bayesian formulation of RL is formally intractable and we considered two approximate solutions, one based on variational Bayes, and the other on Expectation Propagation, our initial findings suggesting that the latter approach is to be generally preferred.

References

P. Abbeel, A. Coates, M. Quigley, and A. Ng. An Application of Reinforcement Learning to Aerobatic Helicopter Flight. *NIPS*, 19:1–8, 2007.

M. J. Beal and Z. Ghahramani. The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. In *Bayesian Statistics*, volume 7, pages 453–464. Oxford University Press, 2003.

R. Crites and A. Barto. Improving Elevator Performance Using Reinforcement Learning. *NIPS*, 8: 1017–1023, 1995.

P. Dayan and G. E. Hinton. Using Expectation-Maximization for Reinforcement Learning. *Neural Computation*, 9:271–278, 1997.

R. Dearden, N. Friedman, and S. Russell. Bayesian Q learning. *AAAI*, 15:761–768, 1998.

M. Duff. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes.* PhD thesis, University of Massachusetts Amherst, 2002.

T. Furmston and D. Barber. Solving deterministic policy (PO)MPDs using Expectation-Maximisation and Antifreeze. *European Conference on Machine Learning (ECML)*, 1:50–65, 2009. Workshop on Learning and data Mining for Robotics.

M. Hoffman, A. Doucet, N. de Freitas, and A. Jasra. Trans-dimensional MCMC for Bayesian Policy Learning. *NIPS*, 20:665–672, 2008.

J. Kober and J. Peters. Policy search for motor primitives in robotics. *NIPS*, 21:849–856, 2009.

F. R. Kschischang, B. J. Frey, and H-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47:498–519, 2001.

T. P. Minka. Expectation Propagation for approximate Bayesian inference. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

C. Rasmussen and M. Deisenroth. Probabilistic inference for fast learning in control. In S. Girgin, M. Loth, R. Munos, P. Preux, and D. Ryabko, editors, *Recent Advances in Reinforcement Learning*, pages 229–242, 2008.

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

M. Toussaint, S. Harmeling, and A. Storkey. Probabilistic inference for solving (PO)MDPs. Research Report EDI-INF-RR-0934, University of Edinburgh, School of Informatics, 2006.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.