Prior-Free Exploration Bonus for and beyond Near Bayes-Optimal Behavior

Kenji Kawaguchi
BWBP Artificial Intelligence Laboratory, Japan
kawaguchi.kenji9@gmail.com

Hiroshi Sato
National Defense Academy of Japan, Japan
hsato@nda.ac.jp

Abstract

We study Bayesian reinforcement learning (RL) as a solution of the exploration-exploitation dilemma. As full Bayesian planning is intractable except for special cases, previous work has proposed several approximation methods. However, these were often computationally expensive or limited to Dirichlet priors. In this paper, we propose a new algorithm that is fast and of polynomial time for near Bayesian optimal policy with any prior distributions that are not greatly misspecified. Perhaps even more interestingly, the proposed algorithm can naturally avoid being misled by incorrect beliefs, while effectively utilizing useful parts of prior information. It can work well even when an utterly misspecified prior is assigned. In that case, the algorithm will follow PAC-MDP behavior instead, if an existing PAC-MDP algorithm does so. The proposed algorithm naturally outperformed other algorithms compared with it on a standard benchmark problem.

1 Introduction

One of the chief challenges in reinforcement learning (RL) [Sutton and Barto, 1998] is the exploration-exploitation trade-off; in RL, an agent needs to explore the world in order to gain new knowledge while it must exploit its current knowledge to earn rewards. One elegant solution for this dilemma is Bayesian RL, in which the agent considers the exploitation of possible future knowledge. Because of this consideration, the agent naturally explores in order to exploit its potential future knowledge. However, in general, such a Bayesian planning is intractable and requires approximation, for example, with the Monte-Carlo method [Wang et al., 2005; Asmuth and Littman, 2011; Browne et al., 2012].

A computationally cheaper way to approximate Bayesian planning is to modify a reward function (or a transition function) to force an agent to explicitly explore. It was shown that a simple algorithm can achieve near Bayesian optimal behavior in polynomial time in this way [Kolter and Ng, 2009]. However, the algorithm works only with an independent Dirichlet prior and a known reward function.

Along this line, several researchers developed more generally applicable algorithms. Yet, some have achieved only PAC-MDP behavior [Sorg et al., 2010] and only in very limited cases [Sorg et al., 2010; Araya-López et al., 2012]. To the best of our knowledge, those algorithms have assured polynomial sample complexity for desired behaviors only with Dirichlet priors.

In this paper, we propose a novel algorithm that can achieve near Bayesian optimal policy in polynomial time for any priors that are not greatly misspecified. In addition, we will discuss and demonstrate that our proposed algorithm not only is more widely applicable, but also performs better than previous algorithms in many cases.

2 Method

A Markov Decision Process (MDP) [Puterman, 1994] is a tuple $\{S, A, R, P, \gamma\}$ where S is a set of states, A is a set of actions, and P is a transition probability function that maps a state and an action to a probability of an agent transitioning to a new state. Also, R is a reward function that defines rewards for each state-action pair, and γ is a discount factor.

2.1 Bayesian Reinforcement Learning

In a situation where the reward and transition probability functions are unknown, the agent needs to estimate them from its experience. A straightforward way to do this is to use Maximum Likelihood Estimation (MLE) based on the frequentist approach. However, an agent with mere MLE can fall into a sub-optimal policy, because the agent may not explore a state-action pair that was once incorrectly judged to have a low value [Strens, 2000]. This is due to the fact that the agent with MLE does not consider its estimation's uncertainty while planning [Meuleau and Bourgine, 1999].

One elegant solution is the Bayesian approach, which explicitly represents epistemic uncertainty by introducing the probability distribution over models b as the agent's belief. In full Bayesian planning, the agent recognizes the transition of its belief b as well as that of the state s defined in MDP. Trying to maximize expected rewards while planning, the agent plans in accordance with the following means:

(a) $R(b,s,a) := \int R(s,a) b[R(s,a)] dR(s,a)$ and,

(b) $P(s' \mid b,s,a) := \int P(s' \mid s,a) b[P(s' \mid s,a)] dP(s' \mid s,a)$.

Concretely, the agent plans by using the following value function [Duff, 2002]:

$$V^*(b,s) = \max_a R(b,s,a) + \gamma \sum_{s'} P(s' \mid b,s,a) V^*(b',s') \quad (1)$$

where s' is the potential next state transitioned to from the current state s, and b' is the possible belief updated to when a reward and a transition are observed with the current belief b.

As can be seen, the Bayesian optimal agent with the value function in equation (1) can naturally account for potential incorrectness in its current belief by considering possible future beliefs b'. Hence, the agent can effectively explore while exploiting the current belief. However, as the number of possible belief states is typically very large, full Bayesian planning with equation (1) is intractable in most cases. Therefore, some approximation is required.

2.2 Exploitation plus Exploration Bonus

Because the intractability of Bayesian planning comes from the consideration of possible future beliefs, we can solve the problem simply by disregarding it. That is, we can use

$$\bar{V}^*(b,s) = \max_a R(b,s,a) + \gamma \sum_{s'} P(s' \mid b,s,a) \bar{V}^*(b,s') .$$

However, we are now facing exactly the same problem that we had with MLE, which was that the agent does not choose to explore. Indeed, one can use any strategy that has been used for MLE to force the agent to explore. In particular, we will use the reward bonus function R' and let the agent use the following internal value function while planning:

$$\tilde{V}^*(b,s) = \max_a \tilde{R}(b,s,a) + \gamma \sum_{s'} P(s' \mid b,s,a) \tilde{V}^*(b,s') , \quad (2)$$

where $\tilde{R}$ is defined to be the reward function R plus the reward bonus function R'.

The MLE version of an internal value function similar to equation (2) was used by Model Based Interval Estimation with Exploration Bonus (MBIE-EB) [Strehl et al., 2008]. MBIE-EB can ensure that the agent follows PAC-MDP behavior [Brafman and Tennenholtz, 2003; Strehl, 2007], but it does not utilize prior information. In order to make use of prior information, equation (2), which is an approximation of Bayesian planning, was used with the bonus being the posterior variance [Sorg et al., 2010]. For the sake of clarity, we refer to this algorithm as epistemic variance bonus (EVB). Although EVB can utilize information in prior belief efficiently, it aims to achieve PAC-MDP behavior similar to that of MBIE-EB. This is a notable disadvantage of EVB, because PAC-MDP algorithms show over-exploration despite prior knowledge and are not optimal in that sense.

Bayesian Exploration Bonus (BEB) [Kolter and Ng, 2009] is an existing algorithm in this line that can ensure that the agent will follow a near Bayesian optimal policy. However, it is only applicable to the special case of an independent Dirichlet prior and a known reward function.

2.3 Prior-Free Exploration Encouragement

We propose a new algorithm, called Prior Free Exploration Encouragement (PFEE). In order to guarantee high performance for different priors, PFEE uses a reward bonus, the exact form of which does not depend on the shapes of priors. To do so, instead of focusing on posterior distribution (which is what related work has done), we pay attention to the history of the estimated means $R(b)$ and $P(b)$. Concretely, along with equation (2), PFEE uses the following bonus:

$$R' = \beta_R \left(\frac{1}{n+1} + \sigma^2_{R(b,s,a)} \right) + \beta_P \left(\frac{1}{n+1} + \sum_{s'} \sigma^2_{P(s'|b,s,a)} \right) (3)$$

where β is the adjustable parameter, n is the number of the most recent estimated means $R(b)$ or $P(b)$, and σ^2 represents the variance of the n number of $R(b)$ or $P(b)$. The number n can be arbitrarily decreased to neglect the old information. Thus, the bonus R' tends to decrease as n is set to be larger, but the σ^2 term penalizes it if the n recent estimated means differ from each other. In other words, equation (3) motivates the agent to explore the state-action pairs, for which the sequence of the estimated means has not yet converged.

Notice that both n and σ^2 are defined in terms of *the means estimated with sequential beliefs*, while similar variables in previous work were characterized by *observations* or *posterior distributions*. Thus, n is not the number of observations used by MBIE-EB, nor is it counts as implied in BEB. For example, the initial value of the variable similar to n in MBIE-EB is 0, and in BEB it varies in accordance with the size of the Dirichlet prior. In contrast, a natural choice of n at the beginning of computation is 1, regardless of prior. This is because a designer initially assigns a prior to an agent, which usually makes the agent have '1' estimation value in its history. Also, the σ^2 term is not the posterior variance used by EVB. For instance, the initial value of the variance in EVB varies based on the prior's shape, and it should not be 0 unless the agent is in a known MDP. However, the variance term of PFEE is initially 0 if n is equal to 1 as stated above.

The reason why we separated the term $1/(n+1)$ in equation (3) is to restrict the number of parameters β. Thus, we believe that its performance can be improved, if an additional parameter is added. However, in this paper, we use only equations (2) and (3) to maintain comparability with existing algorithms. For the same reason, we initialize n with 1 and set n to be equal to the number of all $R(b)$ or $P(b)$ in the agent's history: n = 1 + (the number of $R(b)$ and $P(b)$ updated).

3 Theoretical Results

This section provides a proof that an agent following PFEE will have a near Bayesian optimal policy in polynomial time for priors that are not greatly misspecified. Also, we discuss how PFEE behaves when priors are greatly misspecified. To prove the former, we first show that the agent always uses an internal value function higher than the Bayesian optimal value function, and then indicate that the internal value function will get close to the Bayesian value function as the agent gains experience. We will prove the first point with the following four Lemmas. Proofs of Lemma 1 through 4 can be found in appendix A.

Lemma 1. *Let $\bar{R}(b_n, s, a)$ be an average of n mean rewards estimated with n beliefs for a state-action pair. Then,*

$$\left| \bar{R}(b_n, s, a) - \bar{R}(b_{n+1}, s, a) \right| \le \frac{1}{n+1} .$$

Note that in order to maintain simple representations, we made an assumption that will not affect the essential meaning of the final result: $R(b,s,a)$ = [0,1]. For the same reason, we also define H to be the time horizon such that γ^H » 0.

Lemma 2. *Let $\sigma_{R(b_n,s,a)}$ be the standard deviation of n mean rewards estimated based on n beliefs. Then,*

$$\left|\sigma^2_{R(b_n,s,a)} - \sigma^2_{R(b_{n+1},s,a)}\right| \le \frac{1}{n+1} .$$

Lemma 3. *Let $R(b_n,s,a)$ denote a mean reward estimated with the n-th belief for a state-action pair. Let $\lambda = 1/\eta\epsilon$ where η is any positive real number less than 1 and ϵ* represents a small quantity. *Then,*

$$\Pr\left(\left|R(b_n,s,a) - \overline{R}(b_n,s,a)\right| < \lambda\sigma^2_{R(b_n,s,a)} + \epsilon\right) > 1-\eta .$$

As illustrated above, we deal with distributions of samples (for estimated means and variances), as well as with the probability distribution of a full population (for prior beliefs). However, when we talk about the distribution of samples, we do not focus on the underlying full population, from which the sample is made. Instead, we treat sampled points as the population of interest.

Lemma 4. *Let $\overline{P}(s' \mid b_n,s,a)$ and $\sigma_{P(s'|b_n,s,a)}$ be an average and a standard deviation of n mean transition probabilities estimated with n beliefs. Then with these, the bounds stated by Lemma 1 through 3 hold for the corresponding transition probability's variables as well.*

Now, we are ready to provide the first important technical result.

Lemma 5. *Let $V^*(b,s)$ and $\tilde{V}^{\mathcal{A}}(b,s)$ denote the optimal Bayesian value function and the value function used by PFEE (defined in equation (2) and (3)). Then, with the parameter β at least 2Hλ and with probability at least 1-4H|S||A|η,*

$$\tilde{V}^{\mathcal{A}}(b,s) \ge V^*(b,s) - 4\epsilon .$$

Proof. Let b_n be the n-th belief where n is the number of estimation values considered in the variance term of the bonus in PFEE. With probability at least 1-2$H|S||A|\eta$,

$$\begin{aligned} V^*(b_n,s) &= \max_a R(b_n,s,a) + \gamma\sum_{s'} P(s' \mid b_n,s,a) V^*(b_{n+1},s') \\ &\le \max_a \overline{R}(b_n,s,a) + \gamma\sum_{s'} \overline{P}(s' \mid b_n,s,a) \overline{V}_1^*(b_{n+1},s') + \\ &\qquad \lambda(\sigma^2_{R(b_n,s,a)} + \sum_{s'}\sigma^2_{P(b_n,s,a)}) + 2\epsilon = \overline{V}_1^*(b_n,s) \end{aligned}$$

where $\overline{V}_1^*(b_{n+1},s')$ is defined by the third line of the equation. In the equation above, the first line describes the definition. The second line follows Lemmas 3, 4 and the final step in the proof of Lemma 3 of [Sorg et al., 2010]. It holds true with probability at least 1-2η per belief updating for each state-action pair.

Because $\overline{V}_1^*(b_n,s)$ includes belief updating, what we need to do next is to exclude it as the following:

$$\begin{aligned} &\overline{V}_1^*(b_n,s) \\ &\le \max_a \overline{R}(b_n,s,a) + \gamma\sum_{s'} \overline{P}(s' \mid b_n,s,a) \overline{V}_2^*(b_n,s') + 2\epsilon + \\ &\qquad \lambda(\sigma^2_{R(b_n,s,a)} + \sum_{s'}\sigma^2_{P(b_n,s,a)} + \frac{2H}{n+1}) + \frac{2H}{n+1} = \overline{V}_2^*(b_n,s) . \end{aligned}$$

This inequality can be established by applying Lemmas 1, 2, and 4 for H Bayesian belief updates in the planning phase. Then, with the above definition of $\overline{V}_2^*(b_n,s)$,

$$\begin{aligned} &\overline{V}_2^*(b_n,s) \\ &\le \max_a R(b_n,s,a) + \gamma\sum_{s'} P(s' \mid b_n,s,a) \tilde{V}_1^*(b_n,s') + 4\epsilon + \\ &\qquad \lambda(2\sigma^2_{R(b_n,s,a)} + 2\sum_{s'}\sigma^2_{P(b_n,s,a)} + \frac{2H}{n+1}) + \frac{2H}{n+1} = \tilde{V}_1^*(b_n,s) \\ &\le \max_a R(b_n,s,a) + \gamma\sum_{s'} P(s' \mid b_n,s,a) \tilde{V}_1^*(b_n,s') + 4\epsilon + \\ &\qquad \lambda\left(2\sigma^2_{R(b_n,s,a)} + \frac{2H}{n+1}\right) + \lambda\left(2\sum_{s'}\sigma^2_{P(b_n,s,a)} + \frac{2H}{n+1}\right) \\ &\le \max_a \tilde{R}(b_n,s,a) + \gamma\sum_{s'} P(s' \mid b_n,s,a) \tilde{V}^{\mathcal{A}}(b_n,s) + 4\epsilon \\ &= \tilde{V}^{\mathcal{A}}(b_n,s) + 4\epsilon \end{aligned}$$

with probability at least 1-2 $H|S||A|\eta$. The first inequality is shown by Lemma 3 with probability at least 1-2η per belief updating for each state-action pair. The second inequality is true because λ is not less than 1. Similarly, the third inequality holds, because H is at least 1. Summarizing the above, with probability at least 1-4$H|S||A|\eta$,

$$V^*(b_n,s) \le \overline{V}_1^*(b_n,s) \le \overline{V}_2^*(b_n,s) \le \tilde{V}^{\mathcal{A}}(b_n,s) + 4\epsilon \qquad □$$

Another key insight behind PFEE is that an agent should follow Bayesian optimal policy only when the agent's belief is reliable. We will define misleading and non-misleading prior distributions based on whether or not a reliable belief can be obtained in a large number of time steps. As it is natural to say that the agent's belief is reliable when the estimated values' variance gets close to 0, we will define the types of priors as the following.

Definition 1. *Let σ^2 denote variances of estimation values regarding both rewards and transition probability. A prior distribution is said to be* ***misleading*** *if $\sigma^2 > \epsilon/H^2\lambda$ when $n=H^2\lambda/\epsilon$. Otherwise, if $\sigma^2 \le \epsilon/H^2\lambda$ when $n=H^2\lambda/\epsilon$, the prior distribution is said to be* ***non-misleading****.*

Finally, we can show one of the main results in this section.

Theorem 1. *Let $V^{\mathcal{A}}(b,s)$ denote the value function described in equation (1) with a policy of PFEE rather than with a Bayesian optimal policy. Suppose that the agent updates its belief only until $n=H^2\lambda/\epsilon$. Let η be equal to $\delta/4|S||A|H$. Then by using a non-misleading prior, PFEE will follow a policy that is 13ϵ-close to the Bayesian optimal policy in polynomial time with probability at least $1-2\delta$:*

$$V^{\mathcal{A}}(b,s) \ge V^*(b,s) - 13\epsilon$$

Proof. Let K be a subset of states, on each of which the agent has $n=H^2\lambda/\epsilon$ estimated mean values. Accordingly, define

$\tilde{V}_K^{\mathcal{A}}(b,s)$ as the value function that is equal to the bonus-excluded value function $V^{\mathcal{A}}(b,s)$ on K and equal to the bonus-included value function $\tilde{V}^{\mathcal{A}}(b,s)$ elsewhere. In addition, let A_K be the event where $\tilde{V}_K^{\mathcal{A}}(b,s)$ becomes equal to $\tilde{V}^{\mathcal{A}}(b,s)$. We first show the relationship of $V^{\mathcal{A}}(b,s)$ and $\tilde{V}_K^{\mathcal{A}}(b,s)$. We can limit the upper bound of PFEE's bonus by the maximum reward value H to maintain *optimism* described in Lemma 5. Thus, by following Lemma 5 of [Kolter and Ng, 2009], for the time horizon H,

$$V^{\mathcal{A}}(b,s) \geq \tilde{V}_K^{\mathcal{A}}(b,s) - H^2 P(A_K)\,. \tag{4}$$

Secondly, we describe the bound for $\tilde{V}_K^{\mathcal{A}}(b,s)$ and $\tilde{V}^{\mathcal{A}}(b,s)$, which will be the bonus term with $n{=}H^2\lambda/\epsilon$ for H time steps,

$$\left|\tilde{V}_K^{\mathcal{A}}(b,s) - \tilde{V}^{\mathcal{A}}(b,s)\right| \leq 4\epsilon + 4\epsilon\,. \tag{5}$$

Finally, when $P(A_K) \leq \epsilon/H^2$, with probability at least 1-2δ,

$$\begin{aligned} V^{\mathcal{A}}(b,s) &\geq \tilde{V}_K^{\mathcal{A}}(b,s) - H^2 P(A_K) \\ &\geq \tilde{V}_K^{\mathcal{A}}(b,s) - \epsilon \\ &\geq \tilde{V}^{\mathcal{A}}(b,s) - 9\epsilon \\ &\geq V^{*}(b,s) - 13\epsilon\,. \end{aligned}$$

The first line uses equation (4), the second line follows the assumption of $P(A_K) \leq \epsilon/H^2$, the third line uses equation (5), and the fourth line follows Lemma 5. Thus, PFEE follows a policy 13ϵ-close to Bayesian optimal policy when $P(A_K) \leq \epsilon/H^2$. Otherwise, when $P(A_K) > \epsilon/H^2$, by Hoeffding's inequality, n will be greater than or equal to $n = H^2\lambda/\epsilon$ at least after

$$O\left(\frac{n|S||A|}{P(A_K)} \ln\frac{|S||A|}{\delta}\right) = O\left(\frac{|S||A|H^4\lambda}{\epsilon^2} \ln\frac{|S||A|}{\delta}\right)$$

time steps with probability at least $1-\delta$. Then, after the time steps, by a strategy similar to the above,

$$V^{\mathcal{A}}(b,s) \geq \tilde{V}^{\mathcal{A}}(b,s) - 8\epsilon > V^{*}(b,s) - 13\epsilon$$

with probability at least 1-2δ. □

Now that we have discussed the behavior of PFEE only for non-misleading priors, we will consider its general behavior for all types of priors.

Lemma 6. *Define B_σ and B_{EVB} to be the bonus' variance term of PFEE and the posterior variance bonus used in EVB respectively. Let $V(s)$ be the objective (true) value function of an MDP. Then for any prior,*

$$\begin{aligned} V^{\mathcal{A}}(b,s) &\geq V^{*}(b,s) - 9\epsilon - B_\sigma \\ &\geq V^{*}(s) - 9\epsilon - B_\sigma - B_{EVB} \end{aligned}$$

with high probability and in polynomial time.

Proof. The first line follows theorem 1, and the second line follows the proof of Lemma 3 in [Sorg et al., 2010]. □

Lemma 6 well represents PFEE's performance. Notice that B_σ gets close to 0 when the prior is non-misleading, but B_{EVB} cannot be 0 unless the posterior distribution has converged on one point of value. Thus, for any non-misleading priors, PFEE can use the value function close to the Bayesian optimal function $V^{*}(b,s)$, while EVB cannot. On the other hand, when the prior is misleading (the first line of the equation with $B_\sigma{\neq}0$), PFEE forces an agent to keep exploring. This is a preferable property of PFEE, because it means that PFEE takes advantage of prior information only when it is useful.

Definition 2. *Define the term* ***completely misleading prior*** *such that with that prior, B_σ is almost always higher than B_{EVB}, and B_σ gets close to 0 only after B_{EVB} goes to 0.*

Corollary 1. *For completely misleading priors, PFEE will follow PAC-MDP behavior if EVB does so.*

Corollary 1 is implied by Lemma 6. If B_σ gets close to 0 when B_{EVB} goes to 0, it is already the time when PFEE's value function becomes close to the true value function. This happens when EVB achieves PAC-MDP behavior.

Moreover, Lemma 6 implies that if the prior is misleading but not completely misleading, PFEE will have a greediness that is better balanced than PAC-MDP and a near Bayesian optimal strategy. This is because an agent with PFEE continues to explore if the prior is misleading (less greedy than misleading belief-oriented Bayes optimal behavior), but stops doing so and follows near Bayesian optimal policy when its belief becomes reliable (when B_σ gets close to 0) (greedier than over-exploring PAC-MDP).

4 Experimental Results

In this section, we present the practical performance of PFEE and other existing algorithms in the 5-state chain world, which is a standard benchmark problem in the literature. In the chain world, there are 5 states, S_1 through S_5, and the agent initially exists in S_1. In all states, the agent can choose two actions, going to S_{i+1} from S_i (where i+1←5 if i = 5), or returning to S_1. But, with probability 0.2, the agent "slips" and performs the opposite action as intended. Rewards are 0.2 for returning to S_1, 1.0 for going to S_5 from S_5, and 0 otherwise. Although optimal policy is to choose to move toward S_5 in all states, this setting encourages non-exploring agents to settle on S_1. For more information on the chain problem, see [Strens, 2000]. To make our results comparable with previously published results, we report the algorithms' performances by showing total rewards in the first 1000 steps.

Table 1 Performance with Uniform Prior

Size of prior	-	0.001	1	3|S|
MBIE-EB	336.5±0.1	-	-	-
EVB	-	268.2±0.3	345.4±0.1	297.2±0.1
BEB	-	344.5±0.1	346.6±0.1	201.7±0.1
PFEE	-	343.3±0.1	346.2±0.1	305.5±0.1

Table 1 shows the algorithms' average performances in 10^5 runs, along with the standard errors, for different sizes of uniform prior (i.e., full prior used in [Poupart, 2006]). Since true transition probabilities are not uniform, the large size of

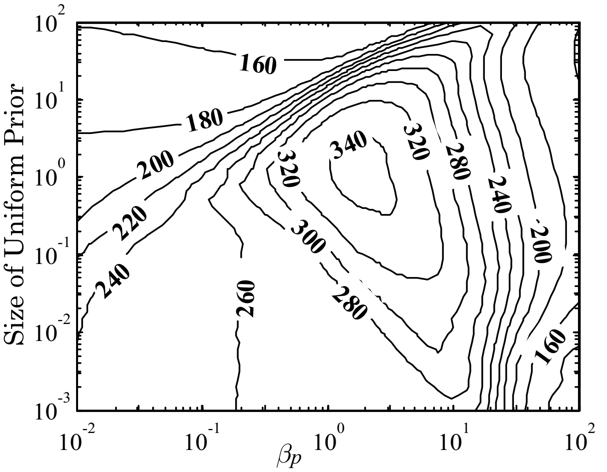

Figure 1. EVB's Performance with Uniform Prior for Different Degrees of Prior and Parameter Value

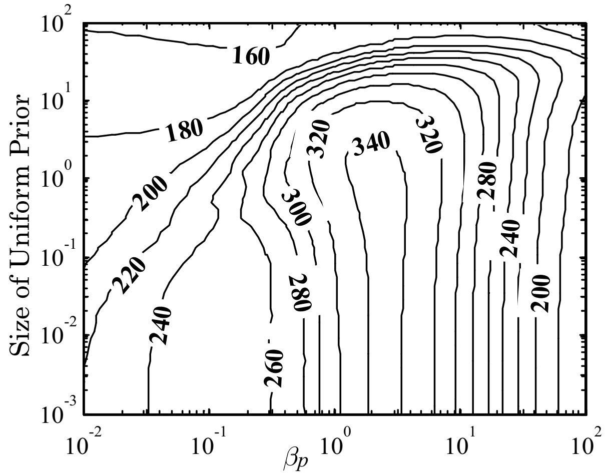

Figure 2. PFEE's Performance with Uniform Prior for Different Degrees of Prior and Parameter Value

uniform prior (3|S|) is likely a (not completely but) misleading prior, and others (0.001 and 1) are expected to be non-misleading. The total rewards of more than 300 are underlined. The parameter β_p of each algorithm is optimized in the same way that previous work adopted, and the optimum β_p was estimated to be 2.5 for MBIE-EB and BEB (indeed, this is the same value reported in [Kolter and Ng, 2009]), 2.0 for EVB, and 2.2 for PFEE. Table 1 illustrates several points. First, BEB can outperform PAC-MDP (EVB and MBIE-EB in this case) when the prior is not greatly misspecified (i.e. when the prior size is 0.001 or 1). This is because the bonus in BEB decreases at the rate $O(1/c)$, where c denotes the number of counts in a prior. This is faster than PAC-MDP's rate $O(\sqrt{1/c})$, and hence an agent following BEB is greedier than an agent following PAC-MDP algorithms. Second, EVB outperformed BEB when the prior was greatly misspecified (i.e. when the prior size is 3|S|=15). The reason for this is the same as for the case above: an agent following BEB is greedier with its belief, which in this case is incorrect to a large degree. Finally, while PFEE performed almost as well as BEB when the prior was not greatly misspecified, it outperformed other algorithms when the prior was greatly misspecified (i.e. when the size of uniform prior was 3|S|). This is what we inferred in the previous section: PFEE will follow Bayesian optimal policy in polynomial time for a non-misleading prior, and can perform better than PAC-MDP and Bayesian optimal behavior for a (not completely but) misleading prior. More concretely for this case, the first term of PFEE's bonus, $1/(n+1)$, decreases at the same rate as BEB's bonus, and thereby PFEE is similarly as greedy as BEB when its bonus' variance term is negligible. This is when the prior is non-misleading. Otherwise, when the prior is misleading, an agent following PFEE keeps exploring effectively, unlike BEB, because of the variance term of its bonus. Thus, PFEE outperformed PAC-MDP and at the same time worked better than BEB in a wider variety of cases.

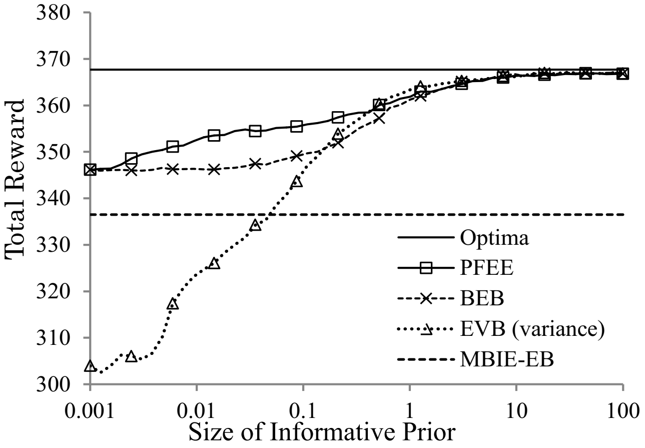

Figure 3. Performance with Informative Prior

In order to illustrate the versatility of PFEE more directly, we show the performances of PFEE and EVB for various sizes of uniform prior and values of parameter β_p in Figures 1 and 2. The 10^5 runs' results are averaged in the figures. PFEE achieved total rewards of more than 340 when the prior size is 10^{-3} to $10^{0.5}$, while EVB did so only for $10^{-0.5}$ to $10^{0.5}$.

The fact that PFEE can handle misspecified priors better than existing algorithms does not mean that PFEE has an inferior ability to utilize *informative priors*. Figure 3 shows performances of each algorithm with optimized parameter β_p for different sizes of *informative prior*. Here, the *informative prior* is defined as: 0.001 + (the true transition probabilities) ×(prior's sizes indicated on the *x*-axis of Figure 3). A similar way to construct *informative priors* was used in [Poupart, 2006]. The results shown are total rewards averaged over 10^4 runs for each prior size, and prior sizes c_j are set such that $c_{j+1} = c_j + c_j \times 0.25$ from $c = 0.001$ to $c \leq 100$ (i.e. the number of the markers in the figure are much less than those of actual data points). As can be seen, PFEE again outperformed existing algorithms. In fact, this result makes sense intuitively, as PFEE's bonus (variance term) can indirectly exploit the information built in the *informative priors*, while BEB's cannot. BEB's bonus utilizes only the amount of information (i.e., the number of counts). EVB's bonus can use that information more directly, but an agent following EVB over-explores when the amount of information is not very large.

5 Conclusion

In this paper, we introduced a new algorithm called PFEE, with which an agent will follow near Bayesian optimal policy in polynomial time for any non-misleading prior (with appropriate parameter values). We derived the algorithm by paying attention to the sequences of the estimated mean values. Regardless of prior, some of the sequences' properties hold true and we utilized this fact to limit its sample complexity for different types of priors.

Perhaps even more importantly, we showed that there are preferable consequences of focusing on the sequences of means generated in the Bayesian setting. Concretely, for completely misleading priors, PFEE will follow PAC-MDP in polynomial time at least when the previous algorithm (EVB) does so. Also, for misleading but not completely misleading priors, PFEE will explore until the belief becomes reliable and then follow Bayesian optimal policy. Thus, PFEE not only utilizes prior information efficiently if it is useful, but also avoids being misled by incorrect beliefs and keeps exploring to make sure to gain a reliable belief.

We demonstrated the above points in the standard chain problem. PFEE outperformed EVB and worked as well as BEB, and in many cases surpassed BEB. PFEE worked better than BEB when the prior is misspecified or informative.

It should be noted that in this paper, we set the number of estimated mean values n to be 1 + (the number of the means estimated) as the simplest configuration. Thus, as future work, one may improve PFEE's performance by adjusting that number based on some criteria. Also, a performance similar to PFEE would be expected for a future algorithm based on PAC-Bayes bounds [Fard and Pineau, 2010], and thus it will be interesting to compare these when such an algorithm becomes available.

A Technical Proofs

This appendix presents the proofs of Lemma 1 through 4 in that order. To make representations concise, we omit the arguments for state-action pairs.

Proof. (of Lemma 1) Let $\bar{R}(b_n)$ be an average of n mean reward values estimated with n beliefs. Then,

$$\begin{aligned}
&\left|\bar{R}(b_n) - \bar{R}(b_{n+1})\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} R(b_i) - \frac{1}{n+1}\sum_{i=1}^{n+1} R(b_i)\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} R(b_i) - \frac{1}{n+1}\left(\sum_{i=1}^{n} R(b_i) + R(b_{n+1})\right)\right| \\
&= \left|\frac{1}{n(n+1)}\sum_{i=1}^{n} R(b_i) - \frac{1}{n+1} R(b_{n+1})\right| \\
&= \frac{1}{n+1}\left|\frac{1}{n}\sum_{i=1}^{n} R(b_i) - R(b_{n+1})\right| \le \frac{1}{n+1}
\end{aligned}$$

The first equality is followed by the definition and the rest can be shown by algebraic manipulations. The last inequality is based on the assumption that $R(b) = [0,1]$. □

Proof. (of Lemma 2) Let $\sigma_{R(b_n)}$ be a standard deviation of n mean rewards estimated with n beliefs. Then,

$$\begin{aligned}
&\left|\sigma^2_{R(b_n)} - \sigma^2_{R(b_{n+1})}\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} R^2(b_i) - \left(\frac{1}{n}\sum_{i=1}^{n} R(b_i)\right)^2 - \frac{1}{n+1}\sum_{i=1}^{n+1} R^2(b_i) + \left(\frac{1}{n+1}\sum_{i=1}^{n+1} R(b_i)\right)^2\right| \\
&= \frac{1}{n+1}\left|\frac{1}{n}\sum_{i=1}^{n} R^2(b_i) - \frac{n}{(n+1)} R^2(b_{n+1}) - \right. \\
&\qquad \left. \frac{2n+1}{n^2(n+1)}\left(\sum_{i=1}^{n} R(b_i)\right)^2 + \frac{2R(b_{n+1})}{(n+1)}\sum_{i=1}^{n} R(b_i)\right| \le \frac{1}{n+1}
\end{aligned}$$

The first equality is followed by the definition, and the second equality can be shown by algebraic manipulations. The last inequality can be derived with the assumption of $R(b) = [0,1]$. Notice that in the absolute value in the third line, there are only two underlying variables $\Sigma R(b)$ and $R(b_{n+1})$ that can affect the absolute value. When these two variables have a huge difference, the left-hand side of the equation gets close to $1/(n+1)$. □

Proof. (of Lemma 3) By using Chebyshev's inequality,

$$\Pr\left(\left|R(b_n) - \bar{R}(b_n)\right| < \lambda\sigma^2_{R(b_n)}\right) > 1 - \frac{1}{\lambda^2\sigma^2} \tag{A1}$$

and

$$\Pr\left(\left|R(b_n) - \bar{R}(b_n)\right| < \epsilon\right) \ge 1 - \frac{\sigma^2}{\epsilon^2}\,. \tag{A2}$$

Note that if we had cared about X, from which the sample $R(b_n)$ would have been weakly exchangeably drawn (this assumption itself is likely unrealistic), then we would have had the following (see [Kabán, 2011]):

$$\Pr\left(\left|X_i - \bar{R}(b_n)\right| < \lambda\sigma^2_{R(b_n)}\right) > 1 - \frac{1}{\lambda^2\sigma^2} - \frac{1}{n}\,.$$

But, this is not the case. Hence, we use equations (A1) and (A2). When $\sigma^2 \ge \epsilon^2\eta$, based on equation (A1),

$$\begin{aligned}
\Pr\left(\left|R(b_n) - \bar{R}(b_n)\right| < \epsilon \le \lambda\sigma^2_{R(b_n)}\right) &> 1 - \frac{1}{\lambda^2\sigma^2} \\
&\ge 1 - \eta\,.
\end{aligned}$$

On the other hand, when $\sigma^2 < \epsilon^2\eta$, with equation (A2),

$$\begin{aligned}
\Pr\left(\left|R(b_n) - \bar{R}(b_n)\right| < \lambda\sigma^2_{R(b_n)} < \epsilon\right) &\ge 1 - \frac{\sigma^2}{\epsilon^2} \\
&> 1 - \eta\,.
\end{aligned}$$

Therefore,

$$\Pr\left(\left|R(b_n) - \bar{R}(b_n)\right| < \lambda\sigma^2_{R(b_n)} + \epsilon\right) > 1 - \eta\,.$$

□

Proof Sketch. (of Lemma 4) One can show the same results of Lemmas 1 to 3 for transition probabilities by using this appendix's equations with $R(b)$ substituted by $\Sigma_{s'}P(s'|b)$ and by noticing two points: the commutative law of addition holds for $\Sigma_{s'}$ and Σ_i, and $\Sigma_{s'}P(s'|b) = 1$ (in the usual case) or $\Sigma_{s'}P(s'|b) = 0$ (this may not make sense in the setting of MDP, but it is not prohibited for an agent to plan in this way). □

References

[Araya-López et al., 2012] Mauricio Araya-López, Vincent Thomas, and Olivier Buffet. Near-Optimal BRL using Optimistic Local Transitions. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[Asmuth and Littman, 2011] John Asmuth and Michael L. Littman. Approaching Bayes-optimality using Monte-Carlo tree search. In *Proceedings of the 21st International Conference on Automated Planning and Scheduling*, 2011.

[Brafman and Tennenholtz, 2003] Ronen I. Brafman and Moshe Tennenholtz. R-MAX – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(2):213–231, 2003.

[Browne et al., 2012] Cameron Browne, Edward Powley, Daniel Whitehouse, Simon Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012.

[Duff, 2002] Mog Duff. *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts, Amherst, 2002.

[Fard and Pineau, 2010] Mahdi Milani Fard and Joelle Pineau. PAC-Bayesian model selection for reinforcement learning. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1624-1632, 2010.

[Kabán, 2011] Ata Kabán. Non-parametric detection of meaningless distances in high dimensional data. *Statistics and Computing*, 22(2):375-385. 2011

[Kolter and Ng, 2009] J. Zico Kolter and Andrew Y. Ng. Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th International Conference on Machine Learning*, pages 513-520, 2009.

[Meuleau and Bourgine, 1999] Nicolas Meuleau and Paul Bourgine. Exploration of multi-state environments: local measures and back-propagation of uncertainty. *Machine Learning*, 35(2):117-154, 1999.

[Poupart, 2006] Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of 23rd International Conference in Machine Learning*, 2006.

[Puterman, 1994] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, John Wiley & Sons, 1994.

[Sorg et al., 2010] Jonathan Sorg, Satinder Singh, and Richard L. Lewis. Variance-based rewards for approximate Bayesian reinforcement learning. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

[Strehl, 2007] Alexander L. Strehl. *Probably approximately correct (PAC) exploration in reinforcement learning*. PhD thesis, Rutgers University, New Brunswick, 2007.

[Strehl et al., 2008] Alexander L. Strehl and Michael L. Littman. An analysis of model-based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309-1331, 2008.

[Strens, 2000] Malcolm Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the 7th International Conference in Machine Learning*, 2000.

[Sutton and Barto, 1998] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, 1998.

[Wang et al., 2005] Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.