



# Bayesian optimistic Kullback–Leibler exploration

Kanghoon Lee<sup>1</sup> · Geon-Hyeong Kim<sup>1</sup>  · Pedro Ortega<sup>2</sup> · Daniel D. Lee<sup>3</sup> · Kee-Eung Kim<sup>1</sup>

Received: 1 April 2018 / Accepted: 28 September 2018 / Published online: 19 December 2018  
© The Author(s) 2018

## Abstract

We consider a Bayesian approach to model-based reinforcement learning, where the agent uses a distribution of environment models to find the action that optimally trades off exploration and exploitation. Unfortunately, it is intractable to find the Bayes-optimal solution to the problem except for restricted cases. In this paper, we present BOKLE, a simple algorithm that uses Kullback–Leibler divergence to constrain the set of plausible models for guiding the exploration. We provide a formal analysis that this algorithm is near Bayes-optimal with high probability. We also show an asymptotic relation between the solution pursued by BOKLE and a well-known algorithm called Bayesian exploration bonus. Finally, we show experimental results that clearly demonstrate the exploration efficiency of the algorithm.

**Keywords** Model-based Bayesian reinforcement learning · Bayes-adaptive Markov decision process · PAC-BAMDP

---

Editors: Masashi Sugiyama and Yung-Kyun Noh.

---

✉ Kee-Eung Kim  
kekim@cs.kaist.ac.kr

Kanghoon Lee  
khlee@ai.kaist.ac.kr

Geon-Hyeong Kim  
ghkim@ai.kaist.ac.kr

Pedro Ortega  
pedroortega@google.com

Daniel D. Lee  
ddl46@cornell.edu

<sup>1</sup> School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

<sup>2</sup> Google UK, DeepMind 6th Floor, Six Pancras Square, Kings Cross, London N1C 4AG, UK

<sup>3</sup> Cornell Tech, 2 West Loop Road, New York, NY 10044, USA

## 1 Introduction

Reinforcement learning (RL) agents face the fundamental problem of maximizing long-term rewards while actively exploring an unknown environment, commonly referred to as *exploration versus exploitation trade-off*. Model-based Bayesian reinforcement learning (BRL) is a principled framework for computing the optimal trade-off from the Bayesian perspective by maintaining a posterior distribution over the model of the unknown environment and computing Bayes-optimal policy (Duff 2002; Poupart et al. 2006; Ross et al. 2007). Unfortunately, it is intractable to exactly compute Bayes-optimal policies except for very restricted cases.

Among a large and growing body of literature on model-based BRL, we focus on algorithms with formal guarantees, particularly PAC-BAMDP (Kolter and Ng 2009; Araya-López et al. 2012). These algorithms are followed by rigorous analyses showing that they are able to perform nearly as well as the Bayes-optimal policy after executing a polynomial number of time steps. They are variants of PAC-MDP algorithms (Kearns and Singh 2002; Strehl and Littman 2008; Asmuth et al. 2009), which guarantee near-optimal performance with respect to the optimal policy of an unknown ground-truth model, to the BRL setting. As such, they balance exploration and exploitation by adopting *optimism in the face of uncertainty principle* as in many PAC-MDP algorithms using additional reward bonus for state-action pairs that are less executed than others (Kolter and Ng 2009), assuming optimistic transitions to states with higher values (Araya-López et al. 2012), or using posterior samples of models (Asmuth 2013).

In this paper, we propose a PAC-BAMDP algorithm based on optimistic transitions with an information-theoretic bound, which we name Bayesian optimistic Kullback–Leibler exploration (BOKLE). Specifically, BOKLE computes policies by constructing an optimistic MDP model in the neighborhood of posterior mean of transition probabilities, defined in terms of Kullback–Leibler (KL) divergence. We provide an analysis showing that BOKLE is near Bayes-optimal with high probability, i.e. PAC-BAMDP. In addition, we show that BOKLE asymptotically reduces to a well-known PAC-BAMDP algorithm, namely Bayesian exploration bonus (BEB) (Kolter and Ng 2009), with a reward bonus equivalent to that of UCB-V (Audibert et al. 2009), which strengthen our understanding of how optimistic transitions and reward bonuses relate to each other. Finally, although our contribution is mainly in the formal analysis of the algorithm, we provide experimental results on well-known model-based BRL domains and show that BOKLE performs better than some representative PAC-BAMDP algorithms in the literature.

We remark that perhaps the most relevant work in the literature is KL-UCRL (Filippi et al. 2010), where the transition probabilities are optimistically chosen in the neighborhood of empirical transition (multinomial) probabilities, also defined in terms of KL divergence. KL-UCRL is also shown to perform nearly optimally under a different type of formal analysis, namely regret. Compared to KL-UCRL, BOKLE can be seen as extending the neighborhood to be defined over Dirichlets. In addition, perhaps not surprisingly, we lose the connection to existing PAC-BAMDP algorithms if we make BOKLE optimal under Bayesian regret. We provide details on this issue in the next section after we review some necessary background.

## 2 Background

A Markov decision process (MDP) is a common environment model for RL, defined by a tuple  $\langle S, A, P, R \rangle$ , where  $S$  is a finite set of states,  $A$  is a finite set of actions,  $P = \{p_{sa} \in$

$\Delta_{S|s \in S, a \in A}$  is the transition distribution, i.e.  $p_{sas'} = \Pr(s'|s, a)$ , and  $R(s, a) \in [0, R_{\max}]$  is the reward function. A (stationary) policy  $\pi : S \rightarrow A$  specifies the action to be executed in each state. For a fixed time horizon  $H$ , the value function of a given policy  $\pi$  is defined as  $V_H^\pi(s) = \mathbb{E}[\sum_{t=0}^{H-1} R(s_t, \pi(s_t)) | s_0 = s]$ , where  $s_t$  is the state at time step  $t$ . The optimal policy  $\pi_H^*$  is typically obtained by computing optimal value function  $V_H^*$  that satisfies Bellman optimality equation  $V_H^*(s) = \max_{a \in A} [R(s, a) + \sum_{s' \in S} p_{sas'} V_{H-1}^*(s')]$  using classical dynamic programming methods (Puterman 2005).

In this paper, we consider model-based BRL where the underlying environment is modeled as an MDP with unknown transition distribution  $P = \{\mathbf{p}_{sa}\}$ . Following the Bayes-adaptive MDP (BAMDP) formulation with discrete states (Duff 2002), we represent  $\mathbf{p}_{sa}$ 's as multinomial parameters and maintain the posterior over these parameters (i.e. belief  $b$ ) using the flat Dirichlet-multinomial (FDM) distribution (Kolter and Ng 2009; Araya-López et al. 2012). Formally, given Dirichlet parameters  $\alpha_{sa}$  for each state-action pair, which consist of both initial prior parameters  $\alpha_{sa}^0$  and execution counts  $\mathbf{n}_{sa}$ , the prior over the transition distribution  $\mathbf{p}_{sa}$  is given by

$$\text{Dir}(\mathbf{p}_{sa}; \alpha_{sa}) = \frac{1}{B(\alpha_{sa})} \prod_{s'} p_{sas'}^{\alpha_{sas'} - 1} \tag{1}$$

where  $B(\alpha_{sa}) = \prod_{s'} \Gamma(\alpha_{sas'}) / \Gamma(\sum_{s'} \alpha_{sas'})$  is the normalizing constant,  $\Gamma$  is the gamma function. The FDM assumes independent transition distributions among state-action pairs so that

$$b(P) = \prod_{s,a} \text{Dir}(\mathbf{p}_{sa}; \alpha_{sa}).$$

Upon observing a transition tuple  $(s, a, s')$ , this prior belief is updated by

$$b_a^{ss'}(P) = \eta p_{sas'} \prod_{\hat{s}, \hat{a}} \text{Dir}(\mathbf{p}_{\hat{s}\hat{a}}; \alpha_{\hat{s}\hat{a}}) = \prod_{\hat{s}, \hat{a}} \text{Dir}(\mathbf{p}_{\hat{s}\hat{a}}; \alpha_{\hat{s}\hat{a}} + \delta_{\hat{s}, \hat{a}, \hat{s}'}(s, a, s'))$$

where  $\delta_{\hat{s}, \hat{a}, \hat{s}'}(s, a, s')$  is the Kronecker delta function that yields 1 if  $(s, a, s') = (\hat{s}, \hat{a}, \hat{s}')$  and 0 otherwise,  $\eta$  is the normalizing factor. This is equivalent to incrementing the single Dirichlet parameter corresponding to the observed transition:  $\alpha_{sas'} \leftarrow \alpha_{sas'} + 1$ . Thus, the belief is equivalently represented by its Dirichlet parameters,  $b = \{\alpha_{sa} | s \in S, a \in A\}$ , and this results in  $\alpha_{sa} = \alpha_{sa}^0 + \mathbf{n}_{sa}$  where  $\alpha_{sa}^0$  is the initial Dirichlet parameters and  $\mathbf{n}_{sa}$  is the execution counts.

The BAMDP formulates the task of computing Bayes-optimal policy as a stochastic planning problem. Specifically, the BAMDP augments environment state  $s$  with current belief  $b$ , which essentially captures the uncertainty in the transition distribution as part of the state space. Then, the optimal value function of the BAMDP should satisfy Bellman optimality equation

$$V_H^*(s, b) = \max_a [R(s, a) + \sum_{s'} \mathbb{E}[p_{sas'} | b] V_{H-1}^*(s', b_a^{ss'})]$$

where  $\mathbb{E}[p_{sas'} | b] = \alpha_{sas'} / \sum_{s''} \alpha_{sas''}$ . Unfortunately, it is intractable to find the solution except for restricted cases primarily because the number of beliefs grows exponentially in  $H$ .

Before we present our algorithm, we briefly review some of the most relevant work in the literature on RL. Since Rmax (Brafman and Tenenholz 2002) and  $E^3$  (Kearns and Singh 1998), a growing body of research has been devoted to algorithms that can be shown to achieve near-optimal performance with high probability, i.e. probably approximately correct (PAC).

Depending on whether the learning target is the optimal policy or the Bayes-optimal policy, these algorithms are classified as either PAC-MDP or PAC-BAMDP. They commonly construct and solve optimistic MDP models of the environment by defining confidence regions of transition distributions centered at empirical distributions, or by adding larger bonuses to rewards of state-action pairs that are less executed than others.

Model-based interval estimation (MBIE) (Strehl and Littman 2005) is a PAC-MDP algorithm that uses confidence regions of transition distributions captured by the 1-norm distance of  $O(1/\sqrt{n_{sa}})$ , where  $n_{sa}$  is the execution count of action  $a$  in state  $s$ . Bayesian optimistic local transition (BOLT) (Araya-López et al. 2012) is a PAC-BAMDP algorithm that uses confidence regions of transition distributions captured by the 1-norm distance of  $O(1/n_{sa})$ , although not explicitly mentioned in the work. MBIE-EB (Strehl and Littman 2008) is a simpler version that uses additive rewards of  $O(1/\sqrt{n_{sa}})$ . On the other hand, BEB (Kolter and Ng 2009) is a PAC-BAMDP algorithm that uses additive rewards of  $O(1/n_{sa})$ . These results imply that we can significantly reduce the degree of exploration in PAC-BAMDP compared to PAC-MDP, which is natural: the learning target is the Bayes-optimal policy (which we know but hard to compute) rather than the optimal policy of the environment (which we don't know).

On the other hand, UCRL2 (Jaksch et al. 2010) uses the 1-norm distance bound of  $O(1/\sqrt{n_{sa}})$  for confidence regions of transition distributions, and is shown to produce near optimal policy under the notion of regret, a formal analysis framework alternative to PAC-MDP. KL-UCRL (Filippi et al. 2010) uses the KL bound of  $O(1/n_{sa})$  to achieve the same regret, while exhibiting a better performance in experiments. This empirical advantage is due to the continuous change in optimistic transition models being constructed with the KL bound. Now, it would be interesting to question ourselves whether we can reduce the degree of exploration if we switch to *Bayesian* regret, as was the case with PAC-BAMDP. Unfortunately, there is some evidence to the contrary. In Bayes-UCB (Kaufmann et al. 2012), it was shown that the Bayesian bandit algorithm requires the same degree of exploration as KL-UCB (Garivier and Cappé 2011). In PSRL (Osband et al. 2013), the formal analysis uses the same set of plausible models as in UCRL2. Hence, we strongly believe that we cannot reduce the degree of exploration under the Bayesian regret criterion.

These results motivate us to investigate a PAC-BAMDP algorithm that uses optimistic transition models with the KL bound of  $O(1/n_{sa}^2)$ , which is the main result of this paper.

### 3 Bayesian optimistic KL exploration

In order to characterize confidence regions of transition distributions defined by KL bound, we first define  $C_{\alpha_{sa}}$  for each state-action pair  $s, a$ , which specifies the KL divergence threshold from the posterior mean  $\mathbf{q}_{sa}$ . Here,  $C_{\alpha_{sa}}$  is the parameter of the algorithm proportional to  $O(1/n_{sa}^2)$ , which will be discussed later. Algorithm 1 presents our algorithm, Bayesian Optimistic KL Exploration (BOKLE), that precisely uses this idea for computing optimistic value functions. For each state-action pair  $s, a$ , the optimistic Bellman backup in BOKLE essentially seeks the solution to the following convex optimization problem

$$\max_{\mathbf{p}} \sum_{s'} p_{s'} \tilde{V}(s') \quad \text{subject to} \quad \begin{aligned} D_{KL}(\mathbf{q}_{sa} \parallel \mathbf{p}) &\leq C_{\alpha_{sa}} \\ \sum_{s'} p_{s'} &= 1 \\ p_{s'} &\geq 0, \forall s' \in S \end{aligned} \quad (2)$$

**Algorithm 1** BOKLE

---

**Input:**  $s_0$  : initial state  
 $\{\alpha_{sa}^0\}$ : Dirichlet parameters of initial belief for each state-action pair  $s, a$

- 1:  $(s, b) \leftarrow (s_0, \{\alpha_{sa}^0\})$
- 2: **for**  $t = 1, 2, \dots, T$  **do**  
     % mean of the belief
- 3:  $\forall (s, a, s') \in S \times A \times S : q_{sas'} = \frac{\alpha_{sas'}}{\sum_{s''} \alpha_{sas''}}$
- 4: **for**  $h = 1, 2, \dots, H$  **do**  
     % optimistic backup within KL-neighborhood
- 5:  $\forall (s, a) \in S \times A : \tilde{Q}_h(s, b, a) = R(s, a) + \max_{\mathbf{p}: D_{KL}(\mathbf{q}_{sa} \parallel \mathbf{p}) \leq C_{\alpha_{sa}}} \sum_{s'} P_{s'} \tilde{V}_{h-1}(s', b)$
- 6:  $\forall s \in S : \tilde{V}_h(s, b) = \max_{a \in A} \tilde{Q}_h(s, b, a)$
- 7: **end for**
- 8: Execute action  $a^* = \operatorname{argmax}_{a \in A} \tilde{Q}_h(s, b, a)$
- 9: Observe new state  $s'$  and update the belief and the state  $b \leftarrow b_{a^*}^{s'}$ ,  $s \leftarrow s'$
- 10: **end for**

---

recursively using  $\tilde{V}$  from the previous step, which can be solved in polynomial time by the barrier method (Boyd and Vandenberghe 2004) (Details are available in the ‘‘Appendix A’’).

**4 PAC-BAMDP analysis**

BOKLE algorithm described in Algorithm 1 obtains the optimistic value function over a KL bound of  $O(1/n_{sa}^2)$ . This exploration bound is much tighter than that of KL-UCRL (Filippi et al. 2010),  $O(1/n_{sa})$ , since BOKLE seeks Bayes-optimal actions whereas KL-UCRL seeks the ground-truth actions. Similarly, the Pinsker inequality implies that the exploration bound can be much tighter than the 1-norm bound of  $O(1/\sqrt{n_{sa}})$  in MBIE (Strehl and Littman 2005) and UCRL2 (Jaksch et al. 2010). In this section, we provide a PAC-BAMDP analysis of BOKLE algorithm even though it optimizes over asymptotically much tighter bound than others. Also, we show the sample complexity bound  $O(\frac{|S||A|H^4 R_{\max}^2}{\epsilon^2} \log \frac{|S||A|}{\delta})$  in Theorem 1, the same complexity bound in BOLT (Araya-L3pez et al. 2012), which is the main result of our analysis.

Before we embark on providing the main theorem, we define KL bound parameter  $C_{\alpha_{sa}}$  in Eq. (2).

**Definition 1** Given a Dirichlet distribution with parameter  $\alpha_{sa}$ , let  $\mathbf{q}_{sa}$  be the mean of the posterior distribution. Then,  $C_{\alpha_{sa}}$  is the maximum KL divergence

$$C_{\alpha_{sa}} = \max_{\substack{h=1, \dots, H, \\ \hat{s} \text{ s.t. } n_{sa\hat{s}} \neq 0}} D_{KL}(\mathbf{q}_{sa} \parallel \mathbf{p}^{h, \hat{s}})$$

where  $\mathbf{p}^{h, \hat{s}}$  is the mean of the Dirichlet distribution with parameter  $\alpha'_{sa} = \alpha_{sa} + h\mathbf{e}_{\hat{s}}$  where  $\mathbf{e}_{\hat{s}}$  is the standard base, i.e.  $\alpha'_{sa}$  can be ‘‘reached’’ from  $\alpha_{sa}$  in  $h$  steps by applying  $h$  Bayesian updates from  $\alpha_{sa}$  so that  $\alpha'_{sas'} = \alpha_{sas'}$  for all  $s' \neq \hat{s}$  except  $\alpha'_{sa\hat{s}} = \alpha_{sa\hat{s}} + h$ .

We note that, in Definition 1,  $h$  artificial pieces of evidence are only applied to the state  $\hat{s}$ , which is the state observed at least once by action  $a$  in state  $s$  while the agent is learning. Therefore,  $\alpha_{sa\hat{s}}$  asymptotically increases at a ratio of  $p_{sa\hat{s}} n_{sa}$ , which results in  $C_{\alpha_{sa}}$  diminishing at a ratio of  $O(1/n_{sa}^2)$  if we regard the true underlying transition probability  $p_{sa\hat{s}}$  as a domain-specific constant.

**Proposition 1**  $C_{\alpha_{sa}}$  defined in Definition 1 diminishes at a ratio of  $O(1/n_{sa}^2)$  and is upper bounded by  $H^2 / \min_{\hat{s} \text{ s.t. } n_{sa\hat{s}} \neq 0} \alpha_{sa} \alpha_{sa\hat{s}}$  where  $\alpha_{sa} = \sum_{s'} \alpha_{sas'}$  is the sum of Dirichlet parameters.

We provide the proof of Proposition 1 in the ‘‘Appendix C.1’’. We now present the main theorem stating that BOKLE is PAC-BAMDP.

**Theorem 1** Let  $A_t$  be the policy followed by BOKLE at time step  $t$  using  $H$  as the horizon for computing value functions with KL bound parameter  $C_{\alpha_{sa}}$  defined in Definition 1, and let  $s_t$  and  $b_t$  be the state and the belief (the parameter of the FDM posterior) at that time. Then, with probability at least  $1 - \delta$ , the Bayesian evaluation of  $A_t$  is  $\epsilon$ -close to the optimal Bayesian evaluation

$$\mathbb{V}_H^{A_t}(s_t, b_t) \geq \mathbb{V}_H^*(s_t, b_t) - \epsilon$$

for all but

$$O\left(\frac{|S||A|H^4 R_{max}^2}{\epsilon^2} \log \frac{|S||A|}{\delta}\right)$$

time steps. In this equation, the definition of Bayes value function  $\mathbb{V}_H^\pi(s, b)$  is

$$\mathbb{V}_H^\pi(s, b) = \sum_a R(s, a) + \sum_{s'} \mathbb{E}[p_{sas'}|b] \mathbb{V}_{H-1}^\pi(s', b_a^{s'}) \tag{3}$$

Our proof of Theorem 1 is based on showing three essential properties of being PAC-BAMDP: *bounded optimism*, *induced inequality*, and *mixed bound*. We provide the proofs of three properties in the ‘‘Appendix C’’ by following the steps analogous to the analyses of BEB (Kolter and Ng 2009) and BOLT (Araya-López et al. 2012).

**Lemma 1** (Bounded Optimism) Let  $s_t$  and  $b_t$  be the state and the belief at time step  $t$ . Then,  $\tilde{V}_H(s_t, b_t)$ , computed by BOKLE with  $C_{\alpha_{sa}}$  defined in Definition 1, is lower bounded by

$$\tilde{V}_H(s_t, b_t) \geq \mathbb{V}_H^*(s_t, b_t) - \frac{H^2 V_{max}}{\alpha_{s_t} + H}$$

where  $\mathbb{V}_H^*(s_t, b_t)$  is the  $H$ -horizon Bayes-optimal value,  $V_{max}$  is the upper bound on the  $H$ -horizon value function, and  $\alpha_{s_t} = \min_a \alpha_{s_t a}$ .

Compared to the optimism lemma that appears in all PAC-BAMDP analysis (Kolter and Ng 2009; Araya-López et al. 2012), this lemma is much more general, because we allow  $\mathbb{V}_H^*(s_t, b_t)$  to be less than  $\tilde{V}_H(s_t, b_t)$  by at most  $O(1/n_{sa})$ . In the proof of the main theorem, we show that this weaker condition is still sufficient to establish that the algorithm is PAC-BAMDP.

The second lemma states that, if we evaluate a policy  $\pi$  on two different rewards and transition distributions,  $R, \mathbf{p}$  and  $\hat{R}, \hat{\mathbf{p}}$ , where  $R(s, a) = \hat{R}(s, a)$  and  $\mathbf{p}_{sa} = \hat{\mathbf{p}}_{sa}$  on a set  $K$  of ‘‘known’’ state-action pairs (Brafman and Tennenholtz 2002), the two value functions will be similar given that the probability of escaping from  $K$  is small. This is a slight modification of the induced inequality lemma used in PAC-MDP analysis, essentially the same lemma in BEB (Kolter and Ng 2009) and BOLT (Araya-López et al. 2012). The known set  $K$  is defined by

$$K = \left\{ (s, a) \mid \alpha_{sa} = \sum_{s'} \alpha_{sas'} \geq m \right\}$$

where  $m$  is a threshold parameter that represents state-action pair with enough evidence. This definition will be used frequently in the rest of this section. We will later derive an appropriate value of  $m$  that results in the PAC-BAMDP bound in Theorem 1.

**Lemma 2** (Induced Inequality) *Let  $\mathbb{V}_h^\pi(s, b)$  be the Bayesian evaluation of a policy  $\pi$  defined by Eq. (3), and  $a$  be the action selected by the policy at  $(s, b)$ . We define the mixed value function by*

$$\hat{\mathbb{V}}_{h+1}^\pi(s, b) = \begin{cases} R(s, a) + \sum_{s'} E[p_{sas'}|b] \hat{\mathbb{V}}_h^\pi(s', b') & \text{if } (s, a) \in K \\ \hat{R}(s, a) + \sum_{s'} \hat{p}_{sas'} \hat{\mathbb{V}}_h^\pi(s', b') & \text{if } (s, a) \notin K \end{cases}$$

for the known set  $K$ , where  $\hat{p}_{sas'}$  is a transition probability that can be different from the expected transition probability  $E[p_{sas'}|b]$  and  $b'$  is the updated belief of  $b$  by observing state transition  $(s, a, s')$ . Let  $A_K$  be the event that a state-action pair not in  $K$  is visited when starting from state  $s$  and following policy  $\pi$  for  $H$  steps. Then,

$$\mathbb{V}_H^\pi(s, \alpha) \geq \hat{\mathbb{V}}_H^\pi(s, \alpha) - V_{\max} \Pr(A_K)$$

where  $V_{\max}$  is the upper bound on the  $H$ -horizon value function and  $\Pr(A_K)$  is the probability of event  $A_K$ .

The last lemma bounds the difference between the value function computed by BOKLE and the mixed value function, where the reward and transition distribution  $\hat{R}, \hat{\mathbf{p}}$  are set to those used by BOKLE. Note that  $\hat{R} = R$  in our case, since BOKLE only modifies transition distribution.

**Lemma 3** (BOKLE Mixed Bound) *Let the known set  $K = \{(s, a) | \alpha_{sa} = \sum_{s'} \alpha_{sas'} \geq m\}$ . Then, the difference between the value obtained by BOKLE,  $\tilde{\mathbb{V}}_H$ , and the mixed value of BOKLE’s policy  $\mathcal{A}_t$  with BOKLE’s transition probabilities  $\hat{\mathbf{p}}_{sa}$  for  $K$ ,  $\hat{\mathbb{V}}_H^{\mathcal{A}_t}$ , is bounded by*

$$\tilde{\mathbb{V}}_H(s_t, b_t) - \hat{\mathbb{V}}_H^{\mathcal{A}_t}(s_t, b_t) \leq \frac{(\sqrt{2}/p^{\min} + 1)H^2V_{\max}}{m}$$

where  $p^{\min} = \min_{s,a,s'} p_{sas'}$  is the minimum non-zero transition probability of each action  $a$  in each state  $s$  on the true underlying environment, which is a domain-specific constant.

Finally, we provide the proof of Theorem 1 using the three lemmas.

**Proof**

$$\begin{aligned} & \mathbb{V}_H^{\mathcal{A}_t}(s_t, b_t) \\ & \geq \hat{\mathbb{V}}_H^\pi(s_t, b_t) - V_{\max} \Pr(A_K) \\ & \geq \tilde{\mathbb{V}}_H(s_t, b_t) - \frac{(\sqrt{2}/p^{\min} + 1)H^2V_{\max}}{m} - V_{\max} \Pr(A_K) \\ & \geq \mathbb{V}_H^*(s_t, b_t) - \frac{(\sqrt{2}/p^{\min} + 1)H^2V_{\max}}{m} - \frac{H^2V_{\max}}{m + H} - V_{\max} \Pr(A_K) \\ & \geq \mathbb{V}_H^*(s_t, b_t) - \frac{\epsilon}{2} - V_{\max} \Pr(A_K) \end{aligned} \tag{4}$$

by applying Lemma 2 (induced inequality) in the first inequality and noticing that  $\mathcal{A}_t$  equals  $\tilde{\pi}$  unless  $A_K$  occurs, Lemma 3 (mixed bound) in the second inequality, Lemma 1 (bounded optimism) in the third inequality. We obtain the last line if we set

$$m = \frac{(2\sqrt{2}/p^{\min} + 4)H^2 V_{\max}}{\epsilon}.$$

This particular value is set to satisfy  $\frac{(\sqrt{2}/p^{\min}+1)H^2 V_{\max}}{m} < \frac{\epsilon}{4}$  and  $\frac{H^2 V_{\max}}{m+H} < \frac{\epsilon}{4}$ , which can be easily checked.

If  $\Pr(A_K) \leq \frac{\epsilon}{2V_{\max}}$ , from Eq. (4), we obtain  $\mathbb{V}_H^{A_t}(s_t, \alpha_t) \geq \mathbb{V}_H^*(s_t, \alpha_t) - \epsilon$ . If  $\Pr(A_K) > \frac{\epsilon}{2V_{\max}}$ , using the Hoeffding and union bounds, with probability at least  $(1 - \delta)$ ,  $A_K$  will occur no more than  $O(\frac{|S||A|m}{\Pr(A_K)} \log \frac{|S||A|}{\delta}) = O(\frac{|S||A|H^4 R_{\max}^2}{\epsilon^2} \log \frac{|S||A|}{\delta})$  time steps since  $V_{\max} \leq HR_{\max}$  (It can be easily checked using the fact that  $\Pr(A_K) > \frac{\epsilon}{2V_{\max}} \geq \frac{\epsilon}{2HR_{\max}}$  with the  $m$  as described above).  $\square$

As we mentioned before, this sample complexity bound  $O(\frac{|S||A|H^4 R_{\max}^2}{\epsilon^2} \log \frac{|S||A|}{\delta})$  is the same as the bound of BOLT (Araya-López et al. 2012)  $O(\frac{|S||A|H^2}{\epsilon^2(1-\gamma)^2} \log \frac{|S||A|}{\delta})$  and better than the bound of BEB (Kolter and Ng 2009)  $O(\frac{|S||A|H^6}{\epsilon^2} \log \frac{|S||A|}{\delta})$  if we reconcile the differences in the problem settings (in BOKLE:  $V_{\max} = HR_{\max}$ , in BEB:  $V_{\max} = H$ , and in BOLT:  $V_{\max} = 1/(1 - \gamma)$ ).

### 5 Relating to BEB

In this section, we discuss how BOKLE relates to BEB (Kolter and Ng 2009). The first few steps of our analysis share some similarities with KL-UCRL (Filippi et al. 2010), but we go further to derive asymptotic approximate solutions in order to make the connection. For the asymptotic analysis, from now on, we will consider confidence regions of transition distributions centered at the posterior mode rather than the mean since both asymptotically converge the same value after a large number of observations.

The mode of the transition distribution in Eq. (1) is  $\mathbf{r}$  given by

$$r_s = \frac{\alpha_s - 1}{\sum_{s'} \alpha_{s'} - |S|},$$

where we dropped the state-action subscript for brevity. If we define the overall concentration parameter  $N = \sum_s \alpha_s - |S|$ , then we can rewrite the belief as  $\text{Dir}(\mathbf{p}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_s p_s^{Nr_s}$ , and its log density as

$$\log \text{Dir}(\mathbf{p}; \boldsymbol{\alpha}) = \sum_s Nr_s \log p_s - \log B(\boldsymbol{\alpha}).$$

Then, the difference of log densities between  $\mathbf{p}$  and the mode  $\mathbf{r}$  becomes

$$\log \text{Dir}(\mathbf{p}; \boldsymbol{\alpha}) - \log \text{Dir}(\mathbf{r}; \boldsymbol{\alpha}) = -ND_{KL}(\mathbf{r}||\mathbf{p})$$

Thus, we can see that isocontours of the Dirichlet density function  $\text{Dir}(\mathbf{p}; \boldsymbol{\alpha}) = \epsilon$  are equivalent to the uniform KL divergence from the mode, i.e.  $D_{KL}(\mathbf{r}||\mathbf{p}) = \epsilon'/N$  with an appropriately chosen  $\epsilon'$ . This shows why KL bound neighborhood is a better idea than the 1-norm neighborhood: the former can be seen as conditioning directly on the density.

Explicitly representing the non-negativity constraint of probabilities, the Lagrangian  $\mathcal{L}$  of the problem in Eq. (2) can be written with the multipliers  $\nu, \mu_s \geq 0$  and  $\lambda$  as

$$\mathcal{L} = \sum_s p_s V(s) - \nu \left( \sum_s r_s \log \frac{r_s}{p_s} - C_{\boldsymbol{\alpha}} \right) - \lambda \left( \sum_s p_s - 1 \right) + \sum_s \mu_s p_s,$$



which has the analytical solution

$$p_s^* = \left[ \frac{\nu}{\lambda - \mu_s - V(s)} \right] r_s = \begin{cases} 0 & \text{if } r_s = 0 \\ \left[ \frac{\nu}{\lambda - V(s)} \right] r_s & \text{if } r_s \neq 0 \end{cases},$$

where the multiplier  $\mu_s = 0$  when if  $r_s \neq 0$ . This is because KL divergence is well-defined only when  $p_s = 0 \Rightarrow r_s = 0$ , and  $\mu_s p_s = 0$  while  $r_s \neq 0$ . Thus,  $\mu_s$  was omitted in the earlier formulation.

We focus on the case  $r_s \neq 0$ , where the solution can be rewritten as

$$p_s^* = \left[ 1 - \frac{1}{\nu} (V(s) - \lambda') \right]^{-1} r_s,$$

with constant  $\lambda'$  is determined by the condition  $\sum_s p_s = 1$ . In the regime  $\nu \gg 1$  (i.e.  $C_\alpha \approx 0$ ), we can approximate this solution by the first-order Taylor expansion:

$$p_s^* \approx \left[ 1 + \frac{1}{\nu} (V(s) - \lambda') \right] r_s = \left[ 1 + \frac{1}{\nu} (V(s) - E_{\mathbf{r}}[V]) \right] r_s \tag{5}$$

where  $E_{\mathbf{r}}[V] = \sum_s r_s V(s)$ .

Then, the KL divergence can be approximated by the second-order Taylor expansion (The proof is available in the ‘‘Appendix D’’):

$$C_\alpha = \frac{1}{2\nu^2} \text{Var}_{\mathbf{r}}[V]$$

where  $\text{Var}_{\mathbf{r}}[V] = E_{\mathbf{r}}[(V - E_{\mathbf{r}}[V])^2]$ . Thus,  $\nu$  can be approximated as

$$\nu \approx \sqrt{\frac{\text{Var}_{\mathbf{r}}[V]}{2C_\alpha}}.$$

Using this  $\nu$  in Eq. (5), we obtain

$$p_s^* \approx \left[ 1 + \sqrt{\frac{2C_\alpha}{\text{Var}_{\mathbf{r}}[V]}} (V(s) - E_{\mathbf{r}}[V]) \right] r_s \quad \text{and}$$

$$\sum_s p_s^* V(s) \approx \sum_s r_s V(s) + \sqrt{2C_\alpha \text{Var}_{\mathbf{r}}[V]}.$$

We can now derive an approximation to the dynamic programming update performed in BOKLE:

$$\begin{aligned} \tilde{V}_h(s, b) &= \max_{a \in A} \left[ R(s, a) + \sum_{s'} p_{sa}^*(s') \tilde{V}_{h-1}(s', b) \right] \\ &\approx \max_{a \in A} \left[ R(s, a) + \sqrt{2C_{\alpha_{sa}} \text{Var}_{\mathbf{r}_{sa}}[V]} + \sum_{s'} r_{sas'} \tilde{V}_{h-1}(s', b) \right]. \end{aligned}$$

which is comparable to the value function computed in BEB (Kolter and Ng 2009):

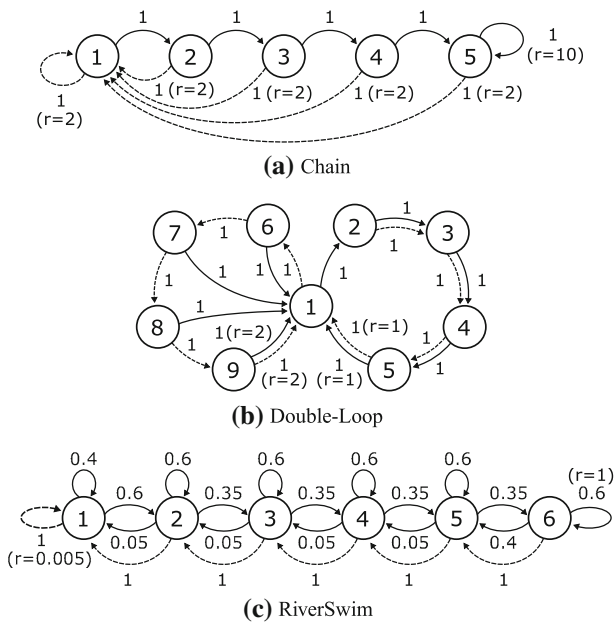
$$V_h^{\text{BEB}}(s, b) = \max_{a \in A} \left[ R(s, a) + \frac{\beta^{\text{BEB}}}{1 + \sum_{s''} \alpha_{sas''}} + \sum_{s'} E[p_{sas'} | b] V_{h-1}^{\text{BEB}}(s', b) \right]$$

for some constant  $\beta^{\text{BEB}}$ . This highly suggests that the additive reward  $\sqrt{C\alpha_{sa}\text{Var}_{r_{sa}}[V]}$  corresponds to BEB exploration bonus  $\frac{\beta^{\text{BEB}}}{1+\sum_{s'}\alpha_{sa's'}}$ , ignoring the mean-mode difference in the transition model. As we have discussed in the previous section,  $C\alpha_{sa} = O(1/n_{sa}^2)$ , which is consistent with BEB exploration bonus  $O(1/n_{sa})$ . In addition, BOKLE scales the additive reward by  $\sqrt{\text{Var}_{r_{sa}}[V]}$ , which incentivizes the agent to explore actions with a higher variance in values, a similar but different formulation compared to Variance-Based Reward Bonus (VBRB) (Sorg et al. 2010). Interestingly, adding the square-root of the empirical variance coincides with the exploration bonus in UCB-V (Audibert et al. 2009), which is a variance-aware upper confidence bound (UCB) algorithm in bandits.

### 6 Experiments

Although our contribution is mainly in the formal analysis of BOKLE, we present simulation results on three BRL domains. We emphasize that the experiments are intended as a preliminary demonstration of how the different exploration strategies compare to each other, and *not* as a rigorous evaluation on real-world problems.

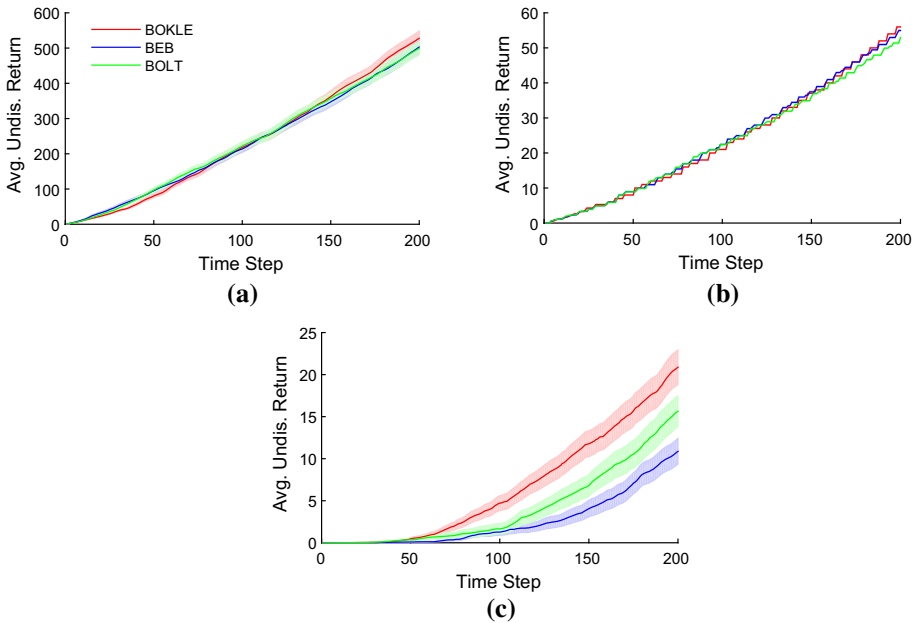
**Chain** (Strens 2000) consists of 5 states and 2 actions as shown in Fig. 1a. The agent starts in state 1 and for each time step can either move on to the next state (action  $a$ , solid edges) or reset to state 1 (action  $b$ , dotted edges). The transition distributions make the agent perform the other action with a “slip” probability of 0.2. The agent receives a large reward of 10 by executing action  $a$  in the rightmost state 5 or a small reward of 2 by executing action  $b$  in any state. **Double-Loop** (Dearden et al. 1998) consists of 9 states and 2 deterministic



**Fig. 1** Three benchmark domains: **a** chain (top), **b** double-loop (middle), and **c** RiverSwim (bottom). The solid (resp. dotted) arrows indicate transition probabilities and rewards for action  $a$  (resp.  $b$ ), but only non-zero rewards are represented together with transition probabilities

**Table 1** Average returns and their standard errors in chain, double-loop, and RiverSwim from 50 runs of 1000 time steps

Algorithm	Chain	Double-loop	RiverSwim
BOKLE	3470.8 (± 44.32)	391.1 (± 0.28)	236.29 (± 3.07)
BEB	3344.04 (± 42.02)	374.9 (± 0.15)	211.8 (± 3.64)
BOLT	3231.64 (± 36.15)	370.4 (± 0.08)	215.1 (± 4.05)



**Fig. 2** Average return versus time step in **a** chain (top-left), **b** double-loop (top-right), **c** RiverSwim (bottom) for three PAC-BAMDP algorithms: BOKLE, BEB, and BOLT. The shaded region represents the standard error

actions as shown in Fig. 1b. It has two loops with a shared (starting) state 1, and the agent has to execute action *b* (dotted edges) to complete the loop with a higher reward of 2, instead of the easier loop with a lower reward of 1. **RiverSwim** (Filippi et al. 2010; Strehl and Littman 2008) consists of 6 states and 2 actions as shown in Fig. 1c. The agent starts in state 1, and can swim either to the left (action *b*, dotted edges) or the right (action *a*, solid edges). The agent has to swim all the way to state 6 to receive a reward of 1, which requires swimming against the current of the river. Swimming to the right has a success probability of 0.35, and a small probability 0.05 of drifting to the left. Swimming to the left always succeeds, but receives a much smaller reward of 0.005 in state 1.

Table 1 compares the returns collected from three PAC-BAMDP algorithms averaged over 50 runs of 1000 timesteps: BOKLE (our algorithm in Algorithm 1), BEB (Kolter and Ng 2009), and BOLT (Araya-López et al. 2012). To handle the sparsity of the transition distributions better, BOKLE used confidence regions centered at the posterior mode. In all experiments, we used the discount factor  $\gamma = 0.95$  for computing internal value functions.

For each domain, we varied the algorithm parameters as follows: for BOKLE,  $C_\alpha = \epsilon/N^2$  where  $\epsilon \in \{0.1, 0.25, 0.5, 1, 5, 10, 25, 50\}$ ; for BEB,  $\beta \in \{0.1, 1, 5, 10, 25, 50, 100, 150\}$ ; for BOLT,  $\eta \in \{0.1, 1, 5, 10, 25, 50, 100, 150\}$ , and selected the best parameter setting for each domain.

In Fig. 2, we show the cumulative returns versus time steps on the onset of each simulation. It is evident from the figure that the learning performance of BOKLE is better than those of BEB and BOLT. These results reflect our discussions on the advantage of KL bound exploration in the previous section.

It is noteworthy that BOKLE performs better than BOLT in the experiments, even though their sample complexity bounds are the same. This result is supported by the discussion in Filippi et al. (2010) on the comparison between KL-UCRL (Filippi et al. 2010) and UCRL2 (Jaksch et al. 2010): For constructing the optimistic transition model, KL-UCRL uses KL divergence bound of  $O(1/n_{sa})$  whereas UCRL2 uses 1-norm distance bound of  $O(1/\sqrt{n_{sa}})$ . Although the formal bounds of these two algorithms are the same, KL-UCRL performs better than UCRL2 in the experiments. This is due to the desirable properties of the neighborhood models under KL divergence, being continuous with respect to the estimated value and robust with respect to unlikely transitions. This insight carries on to BOKLE versus BOLT, since BOKLE uses KL divergence bound of  $O(1/n_{sa}^2)$  whereas BOLT uses 1-norm distance bound of  $O(1/n_{sa})$ .

## 7 Conclusion

In this paper, we introduced Bayesian optimistic Kullback–Leibler exploration (BOKLE), a model-based Bayesian reinforcement learning algorithm that uses KL divergence in constructing the optimistic posterior model of the environment for Bayesian exploration. We provided a formal analysis showing that the algorithm is PAC-BAMDP, meaning that the algorithm is near Bayes-optimal with high probability.

As we have discussed in previous sections, using KL divergence is a natural measure of bounding the credible region of multinomial transition models when constructing optimistic models for exploration. It directly yields the log ratio of the posterior density to the mode, which results in smooth isocontours in the probability simplex. In addition, we showed that the optimistic model constrained by KL divergence can be quantitatively related to other algorithms that use an additive reward approach for exploration (Kolter and Ng 2009; Sorg et al. 2010; Audibert et al. 2009). We presented simulation results on a number of standard BRL domains, highlighting the advantage of using KL exploration.

A number of promising directions for future work include extending the approach to other families of priors and continuous state/action spaces, as well as their formal analyses. In particular, we believe that BOKLE can be extended to the continuous case, similar to UCCRL (Ortner and Ryabko 2012), and it would be an important direction for our future work.

**Acknowledgements** This work was supported by the ICT R&D program of MSIT/IITP (No. 2017-0-01778, Development of Explainable Human-level Deep Machine Learning Inference Framework) and was conducted at High-Speed Vehicle Research Center of KAIST with the support of the Defense Acquisition Program Administration and the Agency for Defense Development under Contract UD170018CD.

## Appendix A: Polynomial time optimization

We show that the optimization problem

$$\begin{aligned} &\text{maximize} && \sum_s p_s \tilde{V}(s) \\ &\text{subject to} && D_{KL}(\mathbf{q} \parallel \mathbf{p}) \leq C_\alpha \\ &&& p_s \geq 0, \forall s \in S \\ &&& \sum_s p_s = 1 \end{aligned}$$

can be solved in polynomial time by the barrier method.

First of all, we show that the above problem is a convex optimization problem since only the convex optimization problems can be applied to the barrier method. Since the objective function and simplex constraints are linear, it is obviously convex. Moreover, the KL bound  $\{\mathbf{p} \mid D_{KL}(\mathbf{q} \parallel \mathbf{p}) \leq C_\alpha\}$  is a convex set by the property of KL divergence. Thus, the problem is a kind of convex optimization problem.

The proof in Chapter 11 of Boyd and Vandenberghe (2004) guarantees that a convex optimization problem with certain assumptions takes a polynomial number of Newton steps. Thus, if the problem satisfies these assumptions, the result of the proof can be directly applied. The assumptions are as follows:

- $-t \sum_s p_s \tilde{V}(s) + \phi(\mathbf{p})$  is closed and self-concordant for all  $t \geq t^{(0)}$ .
- The sublevel sets of the original optimization problem are bounded.

In the first assumption,  $\phi(\mathbf{p}) = -\log(C_\alpha - D_{KL}(\mathbf{q} \parallel \mathbf{p})) - \sum_s \log p_s$  and  $t^{(0)} > 0$ . Now, we will show that the assumptions hold.

Let  $\text{dom}\phi = \{\mathbf{p} \mid \kappa \geq 0, p_s \geq 0, \forall s \in S\}$  be the domain of  $\phi$ . Then  $-t \sum_s p_s \tilde{V}(s) + \phi(\mathbf{p})$  is closed since it is a continuous function and  $\text{dom}\phi$  is compact.

Let  $\kappa = C_\alpha - D_{KL}(\mathbf{q} \parallel \mathbf{p})$  and  $\eta_s = \kappa/q_s$ . Then,

$$\begin{aligned} \frac{\partial^2 \phi}{\partial p_s^2} &= \left(\frac{q_s/p_s}{\kappa}\right)^2 (\eta_s^2 + \eta_s + 1), \\ \frac{\partial^3 \phi}{\partial p_s^3} &= -\left(\frac{q_s/p_s}{\kappa}\right)^3 (2\eta_s^3 + 2\eta_s^2 + 3\eta_s + 2). \end{aligned}$$

From  $\kappa > 0$  and  $\eta_s > 0$ , we obtain

$$\begin{aligned} \frac{\partial^2 \phi}{\partial p_s^2} &= \left(\frac{q_s/p_s}{\kappa}\right)^2 (\eta_s^2 + \eta_s + 1) \geq 0, \\ 4 \left(\frac{\partial^2 \phi}{\partial p_s^2}\right)^3 - \left(\frac{\partial^3 \phi}{\partial p_s^3}\right)^2 &= \left(\frac{q_s/p_s}{\kappa}\right)^6 (4\eta_s^5 + 8\eta_s^4 + 8\eta_s^3 + 7\eta_s^2) \geq 0. \end{aligned}$$

Therefore,  $\left|\frac{\partial^3 \phi}{\partial p_s^3}\right| \leq 2\left(\frac{\partial^2 \phi}{\partial p_s^2}\right)^{3/2}$  and it provides that  $-t \sum_s p_s \tilde{V}(s) + \phi(\mathbf{p})$  is self-concordant.

For any  $k > 0$ , the  $k$ -sublevel set of the original problem is contained in  $\{\mathbf{p} \mid \sum_s p_s \tilde{V}(s) \leq k\} \cap \text{dom}\phi$ , which is a bounded set. Therefore, the sublevel sets are bounded.

We can apply the result in Boyd and Vandenberghe (2004) since the given problem satisfies the assumptions. According to the result, the given problem takes at most  $O(\sqrt{m} \log(m^2 GRM/\epsilon))$  Newton steps where  $m, G, R, M$ , and  $\epsilon$  are the number of

inequalities, the maximum Euclidean norm of the gradient of the objective function and the constraints, radius of the Euclidean ball which contains  $\text{dom}\phi$ , the maximum value of the objective function, and accuracy, respectively. These parameters satisfy  $m = |S| + 1$ ,  $G \leq \max\{\sqrt{|S|}, \sqrt{\sum_s \tilde{V}(s)^2}, 1\}$ ,  $R = \|\mathbf{q}\|_2 + \sqrt{2C_\alpha}$ ,  $M \leq V_{\max}$ . Therefore, the total number of Newton steps is no more than  $O(\sqrt{|S|} \log(|S|^3))$  for any fixed  $\epsilon$ . Consequently, the given optimization problem can be solved in polynomial time by the barrier method.

### Appendix B: Reducing the number of $D_{KL}$ evaluations in Definition 1

We argue that the optimal solution of the optimization problem in Definition 1 is equivalent to the optimal solution of the optimization problem

$$C_\alpha = \max_{h=1, \dots, H} D_{KL}(\mathbf{q} \parallel \mathbf{p}^{h, s'_{\min}}) \tag{6}$$

where  $s'_{\min} = \text{argmin}_{s' \text{ s.t. } n_{s'} \neq 0} \alpha_{s'}$ . By doing so, the number of  $D_{KL}$  evaluations will reduce from  $|S|H$  in Definition 1 to  $H$  in Eq. (6).

From now, we prove the equivalence between the two optimization problems, Definition 1 and Eq. (6). Let  $\alpha_0 = \sum_s \alpha_s$ . Then, for a fixed  $h$ ,

$$\begin{aligned} & \max_{\forall s'} D_{KL}(\mathbf{q} \parallel \mathbf{p}^{h, s'}) \\ &= \max_{\forall s'} \sum_s \frac{\alpha_s}{\alpha_0} \log \left[ \frac{\alpha_s(\alpha_0 + h)}{(\alpha_s + \delta_{s'}^s h)\alpha_0} \right] \\ &= \sum_s \frac{\alpha_s}{\alpha_0} \log \left[ \frac{\alpha_s(\alpha_0 + h)}{\alpha_0} \right] + \max_{\forall s'} \left[ - \sum_s \frac{\alpha_s}{\alpha_0} \log(\alpha_s + \delta_{s'}^s h) \right] \\ &= \sum_s \frac{\alpha_s}{\alpha_0} \log \left[ \frac{\alpha_s(\alpha_0 + h)}{\alpha_0} \right] - \frac{1}{\alpha_0} \sum_s \alpha_s \log \alpha_s + \frac{1}{\alpha_0} \max_{\forall s'} \alpha_{s'} \left[ \log \frac{\alpha_{s'}}{\alpha_{s'} + h} \right] \end{aligned}$$

Thus, maximizing  $D_{KL}(\mathbf{q} \parallel \mathbf{p}^{h, s'})$  is equivalent to maximizing

$$\alpha_{s'} \log \frac{\alpha_{s'}}{\alpha_{s'} + h} \tag{7}$$

with respect to  $s'$ . Fortunately, Eq. (7) is a decreasing function since for  $f(x) = x \log \frac{x}{x+h}$ ,

$$\begin{aligned} f(x) &= x \log x - x \log(x + h) \\ f'(x) &= \log \frac{x}{x + h} + \frac{h}{x + h} \\ &= \log z + (1 - z) \\ &< 0 \end{aligned}$$

where  $z = \frac{x}{x+h}$  and  $z \in (0, 1)$  since  $x > 0, h \geq 0$ . Thus,  $f(x)$  is a decreasing for all  $x > 0$ .

Going back to Eq. (7), it has the maximum value at  $\alpha_{s'_{\min}}$  where  $s'_{\min} = \text{argmin}_{s' \text{ s.t. } n_{s'} \neq 0} \alpha_{s'}$ . Hence,  $C_\alpha$  also has the maximum value at  $s'_{\min}$ . Since we need to compute  $D_{KL}$  only for  $s'_{\min}$ , the total number of evaluations is reduced from  $|S|H$  to  $H$ .

## Appendix C: Proofs of PAC-BAMDP analysis

### Appendix C.1: Proof of Proposition 1

**Proof** Recall that  $q_{sas'} = \alpha_{sas'}/\alpha_{sa}$  is the posterior mean. Then,

$$\begin{aligned} D_{KL}(\mathbf{q}_{sa} \parallel \mathbf{p}^{h, \hat{s}}) &= \sum_{s'} \frac{\alpha_{sas'}}{\alpha_{sa}} \log \frac{\alpha_{sas'}/\alpha_{sa}}{(\alpha_{sas'} + h\delta_{\hat{s}}(s'))/(\alpha_{sa} + h)} \\ &= \sum_{s'} \frac{\alpha_{sas'}}{\alpha_{sa}} \left[ \log \frac{\alpha_{sas'}}{\alpha_{sas'} + h\delta_{\hat{s}}(s')} + \log \frac{\alpha_{sa} + h}{\alpha_{sa}} \right] \\ &= \sum_{s'} \frac{\alpha_{sas'}}{\alpha_{sa}} \log \left[ 1 + \frac{h\delta_{\hat{s}}(s')}{\alpha_{sas'}} \right]^{-1} + \log \left[ 1 + \frac{h}{\alpha_{sa}} \right] \\ &\leq \sum_{s'} \frac{\alpha_{sas'}}{\alpha_{sa}} \left[ -\frac{h\delta_{\hat{s}}(s')}{\alpha_{sas'}} + \frac{h^2\delta_{\hat{s}}(s')}{2\alpha_{sas'}^2} \right] + \frac{h}{\alpha_{sa}} \\ &= \frac{h^2}{2\alpha_{sa}\alpha_{sa\hat{s}}} \\ &\leq \frac{H^2}{\min_{\hat{s} \text{ s.t. } n_{sa\hat{s}} \neq 0} \alpha_{sa}\alpha_{sa\hat{s}}} \end{aligned}$$

where the first inequality is due to Taylor inequalities  $\log(1+x)^{-1} \leq -x + \frac{x^2}{2}$  and  $\log(1+x) \leq x$ . Since both  $\alpha_{sa}$  and  $\alpha_{sa\hat{s}}$  increase at a ratio of  $O(n_{sa})$  by Definition 1,  $D_{KL}(\mathbf{q}_{sa} \parallel \mathbf{p}^{h, \hat{s}})$  has a ratio of  $O(1/n_{sa}^2)$  and thus so does  $C_{\alpha_{sa}}$ .  $\square$

### Appendix C.2: Proof of Lemma 1

**Proof** The proof is almost an immediate consequence of defining  $C_{\alpha_{sa}}$  to bounded cover the mean of any belief that can be reached from  $b_t$  in  $H$  time steps.

More formally, recall the following recursive definition of the  $h$ -horizon Bayes-optimal value

$$\mathbb{V}_h^*(s, b_{t+i}) = \max_a \left[ R(s, a) + \sum_{s'} \mathbb{E}[p_{sas'} | b_{t+i}] \mathbb{V}_{h-1}^*(s', b_{t+i+1}) \right]$$

where  $\mathbb{E}[p_{sas'} | b] = \alpha_{sas'}/\sum_{s''} \alpha_{sas''}$ ,  $b_{t+i}$  is a belief reachable from  $b_t$  in  $i = H - h$  time steps, and  $b_{t+i+1}$  is the updated belief after observing transition  $(s, a, s')$ , i.e.  $b_{t+i+1} = (b_{t+i})_a^{s, s'}$ .

Let  $\mathbb{Q}_h^*$  be the Bayes-optimal action value function, defined by

$$\mathbb{Q}_h^*(s, b_{t+i}, a) = R(s, a) + \sum_{s'} \mathbb{E}[p_{sas'} | b_{t+i}] \mathbb{V}_{h-1}^*(s', b_{t+i+1}).$$

From the result of BOLT Araya-López et al. (2012),  $\mathbb{Q}_h^*(s, b_{t+i}, a)$  is maximized among the beliefs of  $\alpha'_{sa} = \alpha_{sa} + i\mathbf{e}_{s'}$ . Let  $s^*$  be the state that maximize  $\mathbb{Q}_h^*(s, b_{t+i}, a)$  among  $\alpha'_{sa}$ 's. Then,

$$\mathbb{Q}_h^*(s, b_{t+i}, a) \leq R(s, a) + \sum_{s'} \frac{\alpha_{sas'} + i\delta_{s^*}(s')}{\alpha_{sa} + i} \mathbb{V}_{h-1}^*(s', b_{t+i+1}). \tag{8}$$

And then, let  $\tilde{Q}_h$  be the BOKLE action value function, defined by

$$\tilde{Q}_h(s, b_t, a) = R(s, a) + \max_{\mathbf{p}: D_{KL}(\mathbf{q}_{sa} \| \mathbf{p}) \leq C_{\alpha_{sa}}} \sum_{s'} p_{s'} \tilde{V}_{h-1}(s', b_t).$$

Then, by Definition 1, for any  $\hat{s}$  such that  $n_{sa\hat{s}} \neq 0$  (i.e.  $n_{sa\hat{s}} > 0$ ),

$$\tilde{Q}_h(s, b_t, a) \geq R(s, a) + \sum_{s'} \frac{\alpha_{sas'} + i \delta_{\hat{s}}(s')}{\alpha_{sa} + i} \tilde{V}_{h-1}(s', b_t). \tag{9}$$

Now, suppose that  $\tilde{V}_{h-1}(s, b_t) - \mathbb{V}_{h-1}^*(s, b_{t+i+1}) \geq \kappa_{h-1}$ . Then, using Eqs. (8) and (9),

$$\begin{aligned} & \tilde{Q}_h(s, b_t, a) - \mathbb{Q}_h^*(s, b_{t+i}, a) \\ & \geq \sum_{s'} \frac{\alpha_{sas'} + i \delta_{\hat{s}}(s')}{\alpha_{sa} + i} \tilde{V}_{h-1}(s', b_t) - \sum_{s'} \frac{\alpha_{sas'} + i \delta_{s^*}(s')}{\alpha_{sa} + i} \mathbb{V}_{h-1}^*(s', b_{t+i+1}) \\ & = \sum_{s'} \frac{\alpha_{sas'} + i \delta_{\hat{s}}(s')}{\alpha_{sa} + i} \left[ \tilde{V}_{h-1}(s', b_t) - \mathbb{V}_{h-1}^*(s', b_{t+i+1}) \right] \\ & \quad - \frac{i \left[ \mathbb{V}_{h-1}^*(\hat{s}, b_{t+i+1}) - \mathbb{V}_{h-1}^*(s^*, b_{t+i+1}) \right]}{\alpha_{sa} + i} \\ & \geq \kappa_{h-1} - \frac{i}{\alpha_{sa} + i} V_{\max} \end{aligned}$$

and we can obtain that

$$\begin{aligned} \tilde{V}_h(s, b_t, a) - \mathbb{V}_h^*(s, b_{t+i}, a) & \geq \min_a \left[ \tilde{Q}_h(s, b_t, a) - \mathbb{Q}_h^*(s, b_{t+i}, a) \right] \\ & \geq \kappa_{h-1} - \frac{i}{\alpha_s + i} V_{\max} \end{aligned}$$

where  $\alpha_s = \min_a \alpha_{sa}$ . This implies that

$$\tilde{V}_H(s_t, b_t) \geq \mathbb{V}_H^*(s_t, b_t) - \frac{H^2}{\alpha_{s_t} + H} V_{\max}$$

since  $-\sum_{i=1, \dots, H} \frac{i}{\alpha_s + i} \geq -\frac{H^2}{\alpha_{s_t} + H}$  and  $\tilde{V}_0(s, b_t) = \mathbb{V}_0^*(s, b_{t+H}) = 0$ . □

### Appendix C.3: Proof of Lemma 2

**Proof** See the proofs of Lemma 5 in BEB (Kolter and Ng 2009) and Lemma 5.2 in BOLT (Araya-López et al. 2012) □

### Appendix C.4: Proof of Lemma 3

**Proof** This can be shown by closely following the proof steps of Lem 5.3 in Araya-López et al. (2012), which uses mathematical induction.

Suppose that  $\tilde{V}_h(s, b_t) - \hat{\mathbb{V}}_h^{\tilde{\pi}}(s, b_{t+i}) \leq \Delta_h$  for any belief  $b_{t+i}$  that is reachable from  $b_t$  in  $i = H - h$  time steps.



First, if  $(s, a) \in K$ ,

$$\begin{aligned}
 \Delta_{h+1}^{(\in K)} &= \tilde{V}_{h+1}(s, b_t) - \hat{V}_{h+1}^{\tilde{\pi}}(s, b_{t+i-1}) \\
 &= \sum_{s'} \hat{p}_{sas'} \tilde{V}_h(s', b_t) - \sum_{s'} \mathbb{E}[p_{sas'} | b_{t+i-1}] \hat{V}_h^{\tilde{\pi}}(s', b_{t+i}) \\
 &\leq \Delta_h + \sum_{s'} (\hat{p}_{sas'} - \mathbb{E}[p_{sas'} | b_{t+i-1}]) \hat{V}_h^{\tilde{\pi}}(s', b_{t+i}) \\
 &\leq \Delta_h + V_{\max} \sum_{s'} |\hat{p}_{sas'} - \mathbb{E}[p_{sas'} | b_{t+i-1}]| \\
 &\leq \Delta_h + V_{\max} \sum_{s'} \left[ |\hat{p}_{sas'} - q_{sas'}| + |q_{sas'} - \mathbb{E}[p_{sas'} | b_{t+i-1}]| \right] \\
 &\leq \Delta_h + V_{\max} \left[ \sqrt{2D_{KL}(\mathbf{q} \parallel \hat{\mathbf{p}}_{sa})} + \frac{H}{\alpha_{sa}} \right] \\
 &\leq \Delta_h + V_{\max} \left[ \sqrt{2C_{\alpha_{sa}}} + \frac{H}{\alpha_{sa}} \right] \\
 &\leq \Delta_h + V_{\max} \left[ \frac{\sqrt{2}H}{\sqrt{\min_{\hat{s} \text{ s.t. } n_{sa\hat{s}} \neq 0} \alpha_{sa} \alpha_{sa\hat{s}}}} + \frac{H}{\alpha_{sa}} \right] \\
 &= \Delta_h + \frac{(\sqrt{2}/p_{sa}^{\min} + 1)H}{\alpha_{sa}} V_{\max}
 \end{aligned}$$

where  $\mathbf{q}_{sa}$  is the posterior mean and  $p_{sa}^{\min} = \min_{s'} p_{sas'}$  is the minimum non-zero transition probability of action  $a$  in state  $s$  on the true underlying environment, which is a domain-specific constant. In the fourth inequality, we apply the Pinsker inequality to the first term. For the second term, we use Lem. 3 in Kolter and Ng (2009), which states

$$\sum_s |E[p_s | \alpha] - E[p_s | \alpha']| \leq 2/(1 + \sum_s \alpha_s)$$

when  $\alpha = \alpha'$  except  $\alpha_s = \alpha'_s + 1$  for an entry  $s$ . The fifth inequality holds since BOKLE chose  $\hat{\mathbf{p}}_{sa}$  within KL bound of  $C_{\alpha_{sa}}$  from the mean. The sixth inequality comes from our upper bound derivation of  $C_{\alpha}$  presented in Proposition 1.

In the case  $(s, a) \notin K$ , with  $a = \tilde{\pi}(s, b_t)$ , the transition distributions are the same, which yields

$$\begin{aligned}
 \Delta_{h+1}^{(\notin K)} &= \tilde{V}_{h+1}(s, b_t) - \hat{V}_{h+1}^{\tilde{\pi}}(s, b_{t+i-1}) \\
 &= \sum_{s'} \hat{p}_{sas'} \left[ \tilde{V}_h(s', b_t) - \hat{V}_h^{\tilde{\pi}}(s', b_{t+i}) \right] \\
 &\leq \Delta_h.
 \end{aligned}$$

Thus, using  $\Delta_{h+1} = \max[\Delta_{h+1}^{(\notin K)}, \Delta_{h+1}^{(\in K)}]$  and summing up over the  $H$  horizon, we obtain the lemma. □

## Appendix D: Second-order Taylor approximation of $C_\alpha$

$$\begin{aligned}
 C_\alpha &= \sum_s r_s \log \frac{r_s}{p_s^*} \\
 &\approx \sum_s r_s \log \left[ 1 + \frac{1}{v} (V(s) - E_{\mathbf{r}}[V]) \right]^{-1} \\
 &\approx \sum_s r_s \left[ -\frac{1}{v} (V(s) - E_{\mathbf{r}}[V]) + \frac{1}{2v^2} (V(s) - E_{\mathbf{r}}[V])^2 \right] \\
 &= \frac{1}{2v^2} \sum_s r_s (V(s) - E_{\mathbf{r}}[V])^2 \\
 &= \frac{1}{2v^2} \text{Var}_{\mathbf{r}}[V]
 \end{aligned}$$

## References

- Araya-López, M., Thomas, V., & Buffet, O. (2012). Near-optimal BRL using optimistic local transitions. In *Proceedings of the 29th international conference on machine learning* (pp. 97–104).
- Asmuth, J., Li, L., Littman, M. L., Nouri, A., & Wingate, D. (2009). A Bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the 25th conference on uncertainty in artificial intelligence* (pp. 19–26).
- Asmuth, J. T. (2013). *Model-based Bayesian reinforcement learning with generalized priors*. Ph.D. thesis, Rutgers University-Graduate School-New Brunswick.
- Audibert, J. Y., Munos, R., & Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410, 1876–1902.
- Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.
- Brafman, R. I., & Tennenholtz, M. (2002). R-MAX—A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3, 213–231.
- Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the fifteenth national conference on artificial intelligence* (pp. 761–768).
- Duff, M. O. (2002). *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. Ph.D. thesis, University of Massachusetts Amherst.
- Filippi, S., Cappé, O., & Garivier, A. (2010). Optimism in reinforcement learning and Kullback–Leibler divergence. In *48th Annual Allerton conference on communication, control, and computing (Allerton)* (pp. 115–122).
- Garivier, A., & Cappé, O. (2011) The KL-UCB algorithm for bounded stochastic bandits and beyond. In *The 24th annual conference on learning theory* (pp. 359–376).
- Jaksch, T., Ortner, R., & Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11, 1563–1600.
- Kaufmann, E., Cappé, O., & Garivier, A. (2012). On Bayesian upper confidence bounds for bandit problems. In *Fifteenth international conference on artificial intelligence and statistics* (pp. 592–600).
- Kearns, M., & Singh, S. (1998) Near-optimal reinforcement learning in polynomial time. In *Proceedings of the 15th international conference on machine learning* (pp. 260–268).
- Kearns, M., & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49, 209–232.
- Kolter, J. Z., & Ng, A. Y. (2009). Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th international conference on machine learning* (pp. 513–520).
- Ortner, R., & Ryabko, D. (2012). Online regret bounds for undiscounted continuous reinforcement learning. In *Proceedings of the 25th international conference on neural information processing systems* (pp. 1763–1771).
- Osband, I., Roy, B. V., & Russo, D. (2013). (More) efficient reinforcement learning via posterior sampling. In *Proceedings of the 26th international conference on neural information processing systems* (pp. 3003–3011).

- Poupart, P., Vlassis, N., Hoey, J., & Regan, K. (2006). An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd international conference on machine learning* (pp. 697–704).
- Puterman, M. L. (2005). *Markov decision processes: Discrete Stochastic Dynamic Programming*. New York: Wiley-Interscience.
- Ross, S., Chaib-draa, B., & Pineau, J. (2007). Bayes-adaptive POMDPs. In *Proceedings of the 20th international conference on neural information processing systems* (pp. 1225–1232).
- Sorg, J., Singh, S., & Lewis, R. L. (2010). Variance-based rewards for approximate Bayesian reinforcement learning. In *Proceedings of the 26th conference on uncertainty in artificial intelligence*.
- Strehl, A. L., & Littman, M. L. (2005) A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on machine learning* (pp. 856–863).
- Strehl, A. L., & Littman, M. L. (2008). An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74, 1309–1331.
- Strens, M. (2000). A Bayesian framework for reinforcement learning. In *Proceedings of the 17th international conference on machine learning* (pp. 943–950).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.