Bayesian Reinforcement Learning via Deep, Sparse Sampling
Supplementary material

A Proofs of Section 4

The following Lemma shows how episodic error is directly proportional to the belief-error:

Lemma 3 (Per-episode Error)**.** *The error in reward under the true and approximate belief in any episode is bounded by*

$$\Delta_h \triangleq \sup_\pi \left\| v^\pi_{\beta_h} - v^\pi_{\hat{\beta}_h} \right\|_\infty \leq \frac{1-\gamma^K}{1-\gamma}\epsilon_h \leq K\epsilon_h.$$

Proof. Due to the fact that the two BAMDPs induced by the beliefs are ϵ_k close in L1 norm, we can use the argument in Theorem 1 of Dimitrakakis (2011). □

We shall also use a trivial lemma from Analysis:

Lemma 4. *If $\|f-g\|_\infty \leq \epsilon$ and $f(x^*) \geq f(x)$, $g(y^*) \geq g(y)$ then $f(y^*) \geq f(x^*) - 2\epsilon$.*

We are now ready to prove the main results:

Lemma 1 (Anytime Error)**.** *Under Assumption 1,*

$$\left\| v^*_\beta - v^K_\beta \right\|_\infty \leq 2\epsilon_0 K \ln \frac{1}{1-\gamma^K}.$$

Proof. First note that if the belief is only changing every K steps, then the Bayes-optimal policy is Markovian over K steps. This means that finding a K-step Markovian policy starting from belief β_h is the same as finding the optimal policy for a fixed belief $\hat{\beta}_h$.This allows us to use Lemma 3 to bound the error of the K-step policy.

Let $f(\pi) \triangleq v^\pi_{\beta_h}$ and $g(\pi) \triangleq v^\pi_{\hat{\beta}_h}$. Therefore,

$$\begin{aligned}
\left\| v^*_\beta - v^K_\beta \right\|_\infty &= \sum_{h=1}^{H} \gamma^{hK}(v^*_{\beta_h} - v^K_{\beta_h}) \\
&\overset{(a)}{\leq} 2\sum_{h=1}^{H} \gamma^{hK}\Delta_h \\
&\overset{(b)}{\leq} 2\sum_{h=1}^{H} \gamma^{hK}K\epsilon_h \\
&\overset{(c)}{\leq} 2K \times \sum_{h=1}^{H} \gamma^{hK}\epsilon_0/h \\
&\leq 2\epsilon_0 K \times \sum_{h=1}^{\infty} \gamma^{Kh}/h \\
&\overset{(d)}{=} 2\epsilon_0 K \times \ln \frac{1}{1-\gamma^K}
\end{aligned}$$

(a) is obtained using Lemma 4.

(b) is a consequence of Lemma 3.

(c) is a consequence of Assumption 1.

(d) is derived from the fact that $\sum_{x=1}^\infty \frac{a^x}{x} = -\log(1-a)$ for $1 > a \geq 0$.

Therefore, $\left\| v^*_\beta - v^K_\beta \right\|_\infty \leq 2\epsilon_0 K \ln \frac{1}{1-\gamma^K}$. □

Lemma 2 (Error of Thompson-sampling-distributed[10] episodic Policy)**.** *Under Assumption 2*

$$\left\| v^K_\beta - v^{TS}_\beta \right\|_\infty \leq \frac{2(KC+\gamma^K)}{(1-\gamma)}.$$

Proof. We want to show that the value of the TS-episodic policy is not much worse than the Bayes-optimal K-step stationary policy.

$$\begin{aligned}
& v^K_\beta - v^{\mathrm{TS}}_\beta \\
& \overset{(a)}{\leq} v^*_\beta(s) - v^{\mathrm{TS}}_\beta(s) \\
& \overset{(b)}{\leq} \int_{\mathcal{M}} d\mu\beta(\mu) \left[V^{\pi^*_\beta,\mu}_{0,K}(s) - \int_{\mathcal{M}} d\mu' V^{\pi^*_{\mu'},\mu'}_{0,K}(s)\beta(\mu') \right] \\
& + \frac{2\gamma^K}{1-\gamma} \\
& = \int_{\mathcal{M}}\int_{\mathcal{M}} \left[V^{\pi^*_\mu,\mu}_{0,K}(s) - V^{\pi^*_{\mu'},\mu'}_{0,K}(s) \right] \beta(\mu)\beta(\mu')d\mu d\mu' \\
& + \frac{2\gamma^K}{1-\gamma} \\
& \overset{(c)}{\leq} \int_{\mathcal{M}}\int_{\mathcal{M}} \frac{2KD(\mu,\mu')}{(1-\gamma)}\beta(\mu)\beta(\mu')d\mu d\mu' \\
& + \frac{2\gamma^K}{1-\gamma} \\
& \overset{(d)}{\leq} \frac{2(KC+\gamma^K)}{(1-\gamma)}.
\end{aligned}$$

(a) follows by the definition of the Bayes-optimal policy.

(b) follows by truncating the reward sequence to K steps.

(c) follows from the approximate MDP Lemma (Even-Dar and Mansour, 2003, Lemma 4) by definition of the MDP distance $D(\mu, \mu')$.

(d) is a direct consequence of Assumption 2. □

Proof of Theorem 1. For the final proof, we add the effect of sampling in the total error from the two lemmas:

[10] i.e. the optimal polices of MDPs distributed according to the current belief.

Proof. Merging the errors due to Thompson-sampling-distributed error and the anytime-error from Lemma (2) and (1), we obtain for all s

$$v_{\beta}^{\pi_{\beta}^{\mathrm{TS}}}(s) \geq v^*_{\beta}(s) - \left(2\epsilon_0 K \ln \frac{1}{1-\gamma^K} + \frac{2(KC+\gamma^K)}{(1-\gamma)}\right).$$

We can then use Hoeffding's bound since utility of $\pi_{\beta}^{\mathrm{DS}}$ is just sampled utility of $\pi_{\beta}^{\mathrm{TS}}$. For simplicity, let $\bar{\rho}$ be the expected error of a TS policy and ρ_i of the i-th sampled policy and let

$$\varepsilon = \sqrt{\frac{\ln(n/\delta)}{2N}}(1-\gamma)^{-1}.$$

Then, we bound the probability of minimal-error policy among samples has an error more than ε than the expectation:

$$\begin{aligned}
&\mathbb{P}\left(\min\left\{\rho_i \mid i = 1, \ldots, N\right\} \geq \bar{\rho} + \varepsilon\right) \\
&\leq \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} \rho_i \geq \bar{\rho} + \varepsilon\right) \leq \delta/n,
\end{aligned}$$

where the last inequality is from Hoeffding, and the boundedness of rewards. Since there are n such policies, and with a union bound, the probability that any policy has an error of more than ε worse than the expected, is bounded by δ. □

B Root sampling and look-ahead view equivalence

Denote $\mathbb{E}_{\beta}$ as the expectation under marginals ν and τ. The optimal for Bayesian value function can be calculated by noting the following equivalence relation:

$$\begin{aligned}
&V_t^{\pi}(s_t, \beta_t) \equiv v^{\pi}_{\beta_t}(s_t) \\
&= \int_{\mathcal{M}} V^{\pi}_{\mu}(s_t)\beta_t(\mu)d\mu \\
&= \int_{\mathcal{M}} r_{t+1}\,\mathbb{P}_{\mu}(r_{t+1}|s_t,a_t)\beta_t(\mu)d\mu \\
&+ \gamma \int_{\mathcal{M}} \sum_{s' \in s_{t+1}} \mathbb{P}_{\mu}(s'|s_t,a_t)V^{\pi}_{\mu}(s')\beta_t(\mu)d\mu \\
&= \int_{\mathcal{M}} r_{t+1}\,\mathbb{P}_{\mu}(r_{t+1}|s_t,a_t)\beta_t(\mu)d\mu \\
&+ \gamma \sum_{s' \in s_{t+1}} \int_{\mathcal{M}} V^{\pi}_{\mu}(s')\,\mathbb{P}_{\mu}(s'|s_t,a_t)\beta_t(\mu)d\mu \qquad (7) \\
&= r_{t+1}\tau(r_{t+1}|\omega_t,a_t)\int_{\mathcal{M}} \beta_{t+1}(\mu)d\mu \\
&+ \gamma \sum_{s' \in s_{t+1}} \nu(\omega_{t+1}|\omega_t,a_t)\int_{\mathcal{M}} V^{\pi}_{\mu}(s')\beta_{t+1}(\mu)d\mu \qquad (8) \\
&= r_{t+1}\tau(r_{t+1}|\omega_t,a_t) + \gamma \sum_{\omega_{t+1}} v^{\pi}_{\beta_{t+1}}(s_{t+1})\nu(\omega_{t+1}|\omega_t,a_t) \\
&= \mathbb{E}_{\beta}[r_{t+1} + \gamma v^{\pi}_{\beta_{t+1}}(s_{t+1})] \\
&= \mathbb{E}_{\beta}[r_{t+1} + \gamma V^{\pi}_{t+1}(s_{t+1}, \beta_{t+1})]
\end{aligned}$$

We obtain Eq.(8) from Eq.(7) using Eq.(2) and the definitions of marginal distributions ν and τ.

C Parameter Selection for Experiments

Here we describe the chosen hyperparameters for each algorithm shown in Table 3. For each algorithm, these are:

1. BAMCP: (depth,no. of simulations): No. of simulations range from 10 to 10^5, or until the environment time-limit is reached. Depth is between {15,50,*auto*}, using the original implementation.

2. SPARSER : (no.of sampled policies,no.of samples per policy,depth parameter K, Horizon). PI is performed upto 1e-4 accuracy, while RTDP performs lookahead planning of depth 15 for all environments, except larger Grid10 and Maze, where depth is set to 50.

3. BFS3: (branching-factor,no.of simulations). Depth is fixed at 15 for all except larger Grid10 and Maze environment, for which it is 50. Branching factor is between {5,10,15} and no. of simulations between {10,10,1000}.

4. SBOSS: (no.of samples,sampling threshold) Cross-validated against {2,4,8,16,32} and {3,5,7} respectively.

D Additional plots

To target larger audience, we provide Python API for the original C++ implementation used in the paper, using Pybind11 (Jakob et al., 2017). It is available at the following link: https://github.com/revorg7/DeepSparseSampling

This API was used in conjunction with Bsuite environment API by Deepmind (Osband et al., 2019) to draw Regret plots comparing DSS to BDQN (Bootstrapped DQN) and TS (Thompson sampling) in Figure 3.

We did this to promote reusability of DSS, as well as demonstrate reproducibility of DSS's advantage over model-free algorithms such as BDQN for discrete gridworld environments (upto 20x20 atleast).

Algorithm	Chain	DoubleLoop	Grid5	Grid10	Maze
Sparser-RTDP	(8,4,10,2)	(4,4,18,2)	(4,2,50,1)	(4,2,200,2)	(4,4,500,1)
Sparser-PI	(4,4,5,2)	(4,4,18,2)	(2,2,25,1)	(2,2,100,2)	(4,2,100,1)
BAMCP	(auto,100)	(15,100)	(50,10000)	(50,10000)	(50,1000)
BFS3	(10,100)	(10,10)	(10,10)	(5,10)	(5,10)
SBOSS	(8,3)	(2,3)	(2,3)	(2,3)	(2,3)

Table 3: Best parameters obtained from the initial 10 tuning runs.

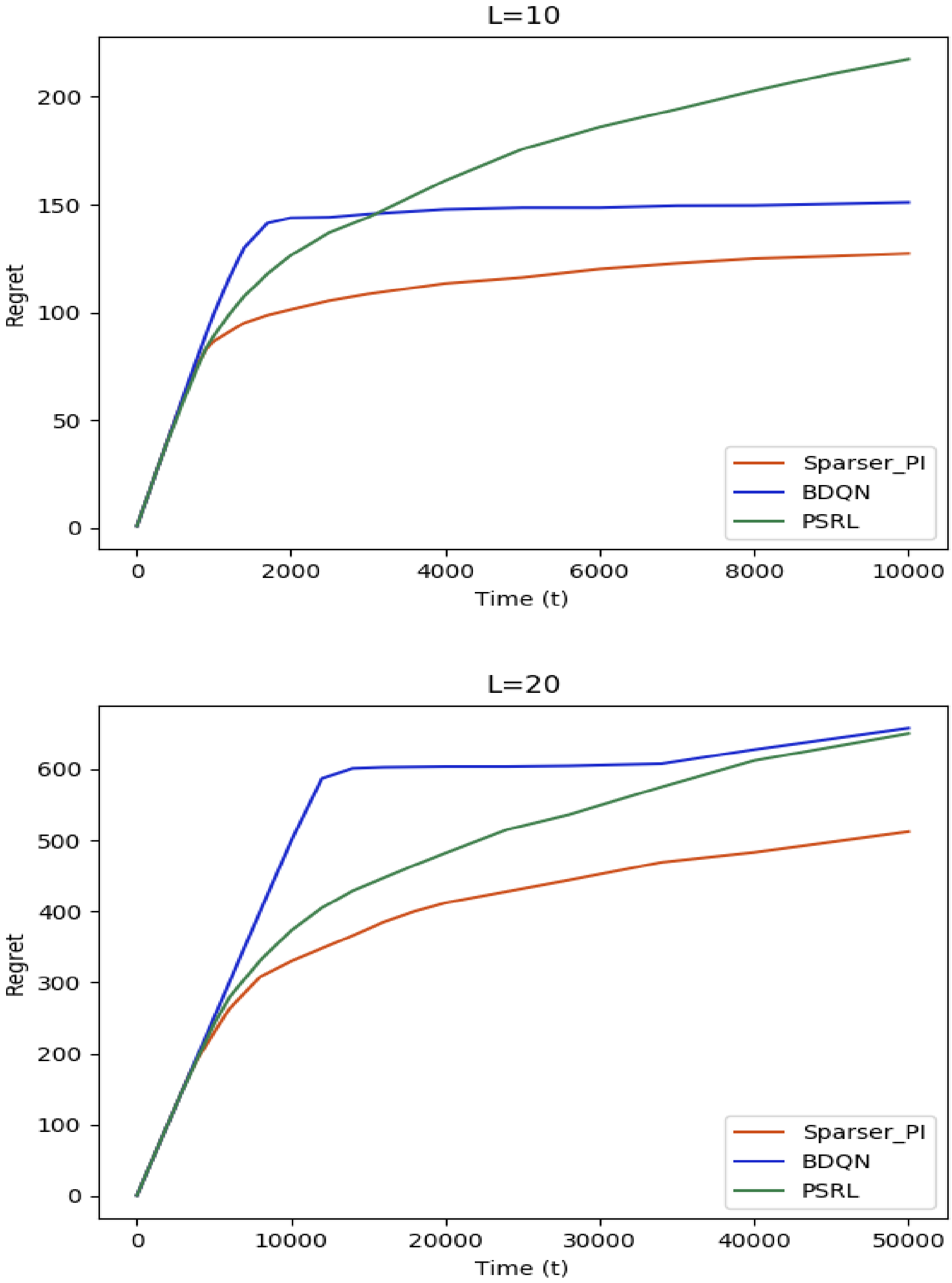

Figure 3: Regret plots(lower is better) for Deep-sea environment for different size parameter 'L'.