On Bayesian Value Function Distributions

Emilio Jorge*[,1] Hannes Eriksson*[,1,2] Christos Dimitrakakis*[,1,3] Debabrota Basu[1,4] Divya Grover[1]

*Equal contribution [1]Chalmers university of Technology [2]Zenseact [3]University of Oslo [4]Scool, Inria Lille-Nord Europe

The problem

- **Setting**: Bayesian reinforcement learning (BRL).
- **Model-based BRL**: Straightforward formalisation by model distributions.
- **Model-free BRL**: Value function distributions via implicit approximations.
- **Solution**: Derive correct value function distributions directly.

Reinforcement learning

An unknown Markov Decision Process (MDP) μ with state s_t, action a_t, reward $r_t \sim P_\mu(r_t \mid s_t, a_t)$, next state $s_{t+1} \sim P_\mu(s_{t+1} \mid s_{t+1}, a_t)$.

Objective: Maximize utility $u_t = \sum_{k=t}^{T} \gamma^t r_t$

The value function V^π of a policy π is

$$V^\pi_\mu(s) \triangleq E^\pi_\mu[u_t \mid s_t = s_0], \qquad a_t \sim P^\pi(a \mid s_t)$$

Bayesian reinforcement learning

The Bayes-optimal solution is

$$\max_\pi E^\pi(u|D)$$

Two main Bayesian approaches

- **Model based:** Belief $\beta \triangleq P(\mu \mid \mathrm{D})$. We can then obtain
$$V^\pi_\beta(s) = \int_\mu V^\pi_\mu(s) dP(\mu \mid \mathrm{D})$$
- **Model free:** Estimate $P(V \mid D)$ directly.

References

[1] Yaakov Engel, Shie Mannor, and Ron Meir.
Bayes meets Bellman: The Gaussian process approach to temporal difference learning.
In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 154–161, 2003.

Bayesian value function estimates

Existing model-free BRL algorithms follow the GPTD[1] framework.

- Gaussian process prior over the $P(V)$
- Likelihood function
$$P(D \mid V) \approx \prod_{i=1}^{t} \exp\{-|V(s_i) - r_i - \gamma V(s_{i+1})|^2\}, \; s_i \in D.$$
- At a high level, the inference is :
$$P(V \mid D) = \frac{P(V)P(D \mid V, \hat{\mu}(D))}{P(D)}$$
- Implicitly assumes the empirical MDP $\hat{\mu}(D)$ is correct
$\Rightarrow$ ignores model uncertainty.

Inferential induction

We propose a framework, **Inferential Induction**, to calculate the value function distribution $P^\pi(V \mid D_t)$ for policy π, correctly.

Data $D_t = s_1, a_1, r_1, \ldots, s_t, a_t, r_t$
$\Rightarrow$VF posterior $P(V_T|D_t), \ldots, P(V_i|D_t), \ldots, P(V_t|D_t)$.

Calculate the value functions with the inductive integral

$$P^\pi(V_i \mid D_t) = \int_\mathcal{V} P^\pi(V_i \mid V_{i+1}, D_t) \, \mathrm{d}P^\pi(V_{i+1} \mid D_t) \qquad \text{(induction)}$$

$$P^\pi(V_i \mid V_{i+1}, D_t) = \int_\mathcal{M} \underbrace{P^\pi(V_i \mid \mu, V_{i+1})}_{\text{Bellman operator}} \, \mathrm{d}\overbrace{P^\pi(\mu \mid V_{i+1}, D_t)}^{\text{link distribution}}. \qquad \text{(marginalisation)}$$

We propose two ways of computing Bayesian value function distributions correctly taking the model uncertainty into account.

Firstly, we introduce **Bayesian Backwards Induction** for calculating $P^\pi(V \mid D_t)$.

- Calculate integral through Monte Carlo sampling of V_{i+1} and μ.
- Define Gaussian kernel relating V_i and utility samples from μ to calculate link distribution $P^\pi(\mu \mid V_{i+1}, D_t)$.
- Importance sampling weights on $P(V_i \mid \mu, V_{i+1})$
- Utilising link distribution may above all be useful when true μ not in model class.

Secondly, we introduce **Inferential Induction Bayesian Actor-Critic** for computing $\mathbb{P}^\pi_\beta(V \mid D_t)$, constructing MDPs from value functions through $\mathbb{P}^\pi_\beta(\mu \mid V)$ using the following transformations.

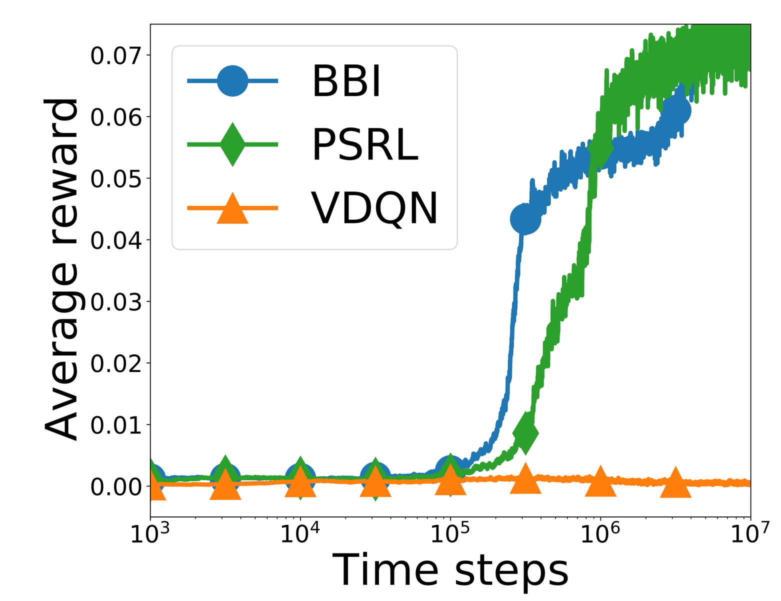
$$\mathbb{P}^\pi_\beta(V \mid D_t) = \frac{\mathbb{P}^\pi_\beta(D_t \mid V) \, \mathrm{d}\,\mathbb{P}^\pi_\beta(V)}{\int_\mathcal{V} \mathbb{P}^\pi_\beta(D_t \mid V) \, \mathrm{d}\,\mathbb{P}^\pi_\beta(V)}.$$

$$\mathbb{P}^\pi_\beta(D_t \mid V) = \int_\mathcal{M} \mathbb{P}^\pi_\mu(D_t) \, \mathrm{d}\,\mathbb{P}^\pi_\beta(\mu \mid V).$$

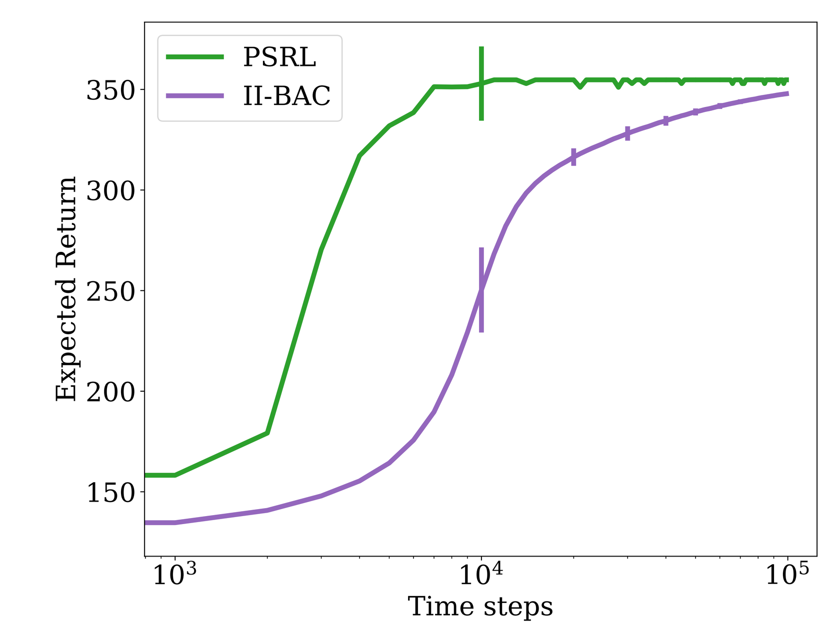
The algorithm becomes

- Sample $V^{(k)} \sim \mathbb{P}^\pi_\beta(V)$.
- Sample $\mu^{(j)} \sim \mathbb{P}^\pi_\beta(\mu \mid V^k)$ by constructing a product of Dirichlet distributions over MDPs minimising the temporal difference error for V^k.
- Compute importance weights $\frac{1}{N_\mu}\sum_{j=1}^{N_\mu} \mathbb{P}^\pi_{\mu^j}(D_t)$ to obtain value function posterior $\mathbb{P}^\pi_\beta(V \mid D_t)$.
- Use sampled MDPs and posterior value function samples to update the actor network parameters.

Experiments



(a) Maze



(b) NChain

Conclusion

- New framework for Bayesian RL.
- BBI uses $P(\mu \mid D)$ to obtain $P(\mu|V, D)$. II-BAC and other suggested method in the work avoid this.
- It does not appear to be possible to do purely "model-free" Bayesian value function estimation.