1. ~~Write Bayesian algorithms intro.~~
2. ~~Environment and experiment description.~~
3. Results and discussion.
4. Conclusions.
5. Fix algorithm loops.
6. Check action notation in moment matching.

Bayesian methods for efficient Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Abstract goes here.

1 Introduction

1.1 Motivation

Balancing exploration and exploitation is one of the central challenges in Reinforcement Learning (RL). On one hand, the agent should *exploit* regions of its environment which are known to be rewarding, while on the other it should *explore* in hope of larger rewards (Sutton and Barto (2018)). Excessively exploitative or explorative behaviours are both suboptimal. In the former, the agent will fixate on small rewards and will be slow to discover the optimal policy. In the latter, it will keep exploring and making suboptimal moves, even though the collected data is sufficient to confidently determine the optimal policy.

A guarantee for sufficient exploration is a crucial part of every RL algorithm. For example, Q-Learning (Watkins and Dayan (1992)) converges to the true Q^*-values, provided among other conditions, that every state-action is visited infinitely often in the limit. To guarantee sufficient exploration, ϵ-greedy or Boltzmann (Sutton and Barto (2018)) approaches are traditionally used. However, as demonstrated by Osband (2016), such schemes can be very slow to learn, because their exploration is *undirected*: they fail to consider the agent's *uncertainty* and instead drive exploration by injecting random noise in action selection. Further, robust methods for annealing ϵ or T cannot be found in the literature. In practice, most applications use constant exploration parameters (Mnih et al. (2015)), at the expense of crude exploration-exploitation tradeoffs.

To improve the efficiency of RL algorithms, we argue that action-selection must be *directed*, that is guided by a quantification of the agent's uncertainty, and Bayesian inference proves to be a natural method for this. By representing the agent's posterior beliefs and selecting actions accordingly, we can direct exploration, providing a principled *transition mechanism* from exploration to exploitation, as the posterior distributions shrink. In this work we present certain Bayesian algorithms, in tabular Markov Decision Processes (MDPs), including our own novel approach.

1.2 Notation convention

We find it valuable to introduce a general notation for our discussion. The MDP $\langle \mathcal{T}, \mathcal{R}, \mathcal{S}, \mathcal{A}, \phi, T \rangle$ is defined by the dynamics and rewards distributions $\mathcal{T} \equiv p(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ and $\mathcal{R} \equiv p(r|\mathbf{s}', \mathbf{s}, \mathbf{a})$, state and action spaces $\mathcal{S}$ and $\mathcal{A}$, initial-state distribution ϕ and episode duration T ($T = \infty$ for continuing tasks). We use $\mathbf{s}, \mathbf{a}, r, \mathbf{s}'$ interchangeably with $\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}$ for states, actions, rewards and next-states, π for the policy and π^* for the optimal policy. In addition to V^π and Q^π to denote state and

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

action values under π, we define the state and action *return* random variables $w^\pi_{\mathbf{s}}$ and $z^\pi_{\mathbf{s},\mathbf{a}}$,

$$w^\pi_{\mathbf{s}} \equiv \sum_{t=1}^{T} \gamma^{\,t-1} r_t \big| \pi, \mathbf{s}_1 = \mathbf{s}, \mathcal{T}, \mathcal{R} \quad \text{and} \quad z^\pi_{\mathbf{s},\mathbf{a}} \equiv \sum_{t=1}^{T} \gamma^{\,t-1} r_t \big| \pi, \mathbf{s}_1 = \mathbf{s}, \mathbf{a}_1 = \mathbf{a}, \mathcal{T}, \mathcal{R}. \tag{1}$$

These are the cumulative discounted rewards received by following π from $\mathbf{s}$, or executing $\mathbf{a}$ from $\mathbf{s}$ and following π thereafter, respectively. We use $\mathcal{W}^\pi$ and $\mathcal{Z}^\pi$ to denote the corresponding distributions.

2 Types of uncertainty: epistemic and aleatoric

Distributional RL (DRL) (Bellemare et al. (2017)) is a recent method leveraging the fact that the action-return is a random variable. The authors consider the *distributional BE*:

$$z^\pi_{\mathbf{s},\mathbf{a}} = r_{\mathbf{s},\mathbf{a},\mathbf{s}'} + \gamma z^\pi_{\mathbf{s}',\mathbf{a}'} \tag{2}$$

where $\mathbf{s}' \sim \mathcal{T}$, $r_{\mathbf{s},\mathbf{a},\mathbf{s}'} \sim \mathcal{R}$, $\mathbf{a}' \sim \pi(\mathbf{s})$, and equality means the two sides are identically distributed. Where traditional algorithms such as Q-Learning aim at learning Q^*, DRL learns the distribution of $z^*_{\mathbf{s},\mathbf{a}}$, denoted $\mathcal{Z}^*$, whose expectation is Q^*. Bellemare et al. (2017) postulate that DRL improves performance partly because it takes advantage of a richer learning signal. Whole distributions over returns are modelled instead of just their means so DRL can gracefully handle multi-modalities in the return.

DRL models the *aleatoric* or *irreducible* uncertainty due to the inherent stochasticity in $\mathcal{T}$ and $\mathcal{R}$. Even if the agent knows $\mathcal{T}$ and $\mathcal{R}$, it will not be able to exactly predict $z^*_{\mathbf{s},\mathbf{a}}$ if the former are stochastic. This inherent noise averages out in expectation and is not of interest for exploration. In addition to aleatoric uncertainty, there will also be uncertainty about the parameterisation of $\mathcal{Z}^*$, because the agent collects a finite amount of data, known as *epistemic* uncertainty. Epistemic uncertainty decreases as more data are observed and the agent should seek to reduce this in a directed manner.

One plausible and principled approach for balancing exploration and exploitation is quantify the epistemic uncertainty and incorporate it into action selection, for example by Thompson sampling (Thompson (1933)). This approach directs exploration according to the amount of reducible uncertainty, and also provides a smooth transition into exploitation, as the posterior becomes narrower.

2.1 Bayesian modelling and the Bellman equations

In both the model-based and model-free settings, we are interested in representing the agent's posterior beliefs about $\mathcal{T}$, $\mathcal{R}$, $\mathcal{W}$ or $\mathcal{Z}$. We parameterise relevant distributions with parameters $\boldsymbol{\theta}$, and will given data $\mathcal{D} = \{\mathbf{s}, \mathbf{a}, \mathbf{s}', r\}$ we want to obtain $p(\boldsymbol{\theta}|\mathcal{D})$. Bayes' rule allows us to do this, so long as we provide a prior $p(\boldsymbol{\theta})$:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}. \tag{3}$$

Choosing a *conjugate* prior simplifies downstream calculations: for discrete distributions such as $\mathcal{T}$, we use a Categorical-Dirichlet model (Murphy (2007)) for each $\mathbf{s}, \mathbf{a}$, while for continuous distributions such as $\mathcal{R}, \mathcal{W}, \mathcal{Z}$ we use a Normal-NG model (Bishop (2006)) for each $\mathbf{s}, \mathbf{a}, \mathbf{s}'$.

3 Bayesian RL algorithms

3.1 Bayesian Q-Learning

Bayesian Q-Learning (BQL) (Dearden et al. (1998)) is a model-free approach for the tabular setting. The agent models the distribution over returns under the optimal policy, $\mathcal{Z}^*$, and updates $p(\boldsymbol{\theta}_{\mathcal{Z}^*}|\mathcal{D})$ as new data arrive. The authors make three modelling assumptions: (1) the return from any state-action is Gaussian; (2) the prior over the mean and precision for each of these Gaussians is Normal-Gamma (NG); (3) the NG posterior[1] factors over different state-actions.

Although the first two are mild assumptions, the latter is more significant because it approximates the true posterior by a factored distribution. In reality, the expected returns are related though the

[1]Since $z^*_{\mathbf{s},\mathbf{a}}$ is modelled by a Gaussian with an NG prior over its mean and precision, the posterior is also NG.

BE, so the exact posterior is not factored. To update $p(\boldsymbol{\theta}_{\mathcal{Z}^*}|\mathcal{D})$ after each transition, the authors use a mixture-of-distributions update rule and approximate this mixture by the NG closest to it in terms of KL-divergence. Action selection can be performed by Thompson sampling. See appendix A.1 for further details.

3.2 Posterior sampling for reinforcement learning

Posterior Sampling for Reinforcement Learning (PSRL) (Osband et al. (2013)) is an elegantly simple and yet provably efficient model-based algorithm for sampling from the exact posterior over optimal policies $p(\pi^*|\mathcal{D})$. It amounts to sampling $\hat{\boldsymbol{\theta}}_{\mathcal{T}} \sim p(\boldsymbol{\theta}_{\mathcal{T}}|\mathcal{D})$ and $\hat{\boldsymbol{\theta}}_{\mathcal{R}} \sim p(\boldsymbol{\theta}_{\mathcal{R}}|\mathcal{D})$, and solving the BE for $\hat{Q}^*|\hat{\boldsymbol{\theta}}_{\mathcal{T}}, \hat{\boldsymbol{\theta}}_{\mathcal{R}}$ and $\hat{\pi}^*|\hat{\boldsymbol{\theta}}_{\mathcal{T}}, \hat{\boldsymbol{\theta}}_{\mathcal{R}}$. Policy $\hat{\pi}^*$ is then followed for a single episode, or for a pre-defined horizon in continuing tasks. Osband et al. (2013) prove the regret of PSRL is sub-linear. See appendix A.2 for further details.

3.3 The uncertainty Bellman equation

The Uncertainty Bellman Equation (UBE), is a model-based method proposed by O'Donoghue et al. (2017), for estimating the epistemic uncertainty in $\mu_{z^{\pi}_{\mathbf{s},\mathbf{a}}}$. The authors assume that: (1) the MDP is a directed acyclic graph (DAG) and the task is episodic, with $t = 1, ..., T$ denoting the episode time-step; (2) the mean immediate rewards of the MDP are bounded within $[-R_{max}, R_{max}]$. Taking variances across the BE and defining an appropriate Bellman operator $\mathcal{U}^{\pi}_t$, they show that the corresponding UBE:

$$u^{\pi}_{\mathbf{s},\mathbf{a},t} = \mathcal{U}^{\pi}_t u^{\pi}_{\mathbf{s},\mathbf{a},t+1}, \text{ where } u^{\pi}_{\mathbf{s},\mathbf{a},T+1} = 0$$

has a unique solution $u^{\pi}_{\mathbf{s},\mathbf{a},t}$ which upper bounds the epistemic uncertainty $\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{T}},\boldsymbol{\theta}_{\mathcal{R}}}\left[\mu_{z^{\pi}_{\mathbf{s},\mathbf{a},t}}\right]$. In practice, assumption (1) must be violated to apply the UBE to non-DAG MDPs or in the continuing setting. By first solving for the greedy policy π^* w.r.t. $p(\boldsymbol{\theta}_{\mathcal{T}}|\mathcal{D})$ and $p(\boldsymbol{\theta}_{\mathcal{R}}|\mathcal{D})$, and then solving the UBE for $u^*_{\mathbf{s},\mathbf{a},t}$, Thompson sampling can be performed from a diagonal Gaussian. The Thompson noise variance is the $\zeta^2 u^*_{\mathbf{s},\mathbf{a},t}$, where ζ is an appropriate scaling factor. This results in a factored posterior approximation. Further details are given in appendix A.3.

3.4 Moment Matching across the Bellman equation

Our moment matching (MM) approach uses the BE to estimate epistemic uncertainties, without resorting to an upper bound approximation. Instead we require equality of first and second moments across the BE. The first-order equation gives the familiar value-BE. Using the laws of total variance and covariance, the second-order moments can be decomposed into purely aleatoric and purely epistemic terms. We argue that the aleatoric and epistemic terms should satisfy two separate equations.

We thus propose first solving for the greedy policy π^* w.r.t. $p(\boldsymbol{\theta}_{\mathcal{T}}|\mathcal{D})$ and $p(\boldsymbol{\theta}_{\mathcal{R}}|\mathcal{D})$, and then for the epistemic uncertainty in $\mu_{w^*_{\mathbf{s}}}$. The latter is used for Thompson sampling from a diagonal gaussian, resulting in a factored approximation of the posterior as in the UBE. A derivation outline and further details are given in appendix A.4.

4 Finite MDP environments

We compare the algorithms on three kinds of finite MDPs of variable sizes - exact specifications and illustrations given in section B -, and all experiments are in the continuing setting. We measure performance by the cumulative regret to an oracle agent which acts under the optimal policy.

Our DeepSea MDP is a variant of those in Osband et al. (2017); O'Donoghue (2018), which tests the algorithm's ability for sustained exploration despite initial negative rewards. We also propose WideNarrow, an environment designed specifically to investigate the effect of factored posterior approximations made in BQL, UBE and MM. Finally, since the DeepSea and WideNarrow are handcrafted, we also compare the algorithms on MDPs drawn from a Dirichlet prior over $\boldsymbol{\theta}_{\mathcal{T}}$ and NG prior over $\boldsymbol{\theta}_{\mathcal{R}}$ as in Osband et al. (2013) - we refer to this as the PriorMDP.

5 Results and discussion

6 Conslusions

References

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 449–458.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Dearden, R., Friedman, N., and Russell, S. (1998). Bayesian q-learning. In *In AAAI/IAAI*, pages 761–768. AAAI Press.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution. Technical report.

O'Donoghue, B. (2018). Variational bayesian reinforcement learning with regret bounds. *CoRR*, abs/1807.09647.

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. (2017). The uncertainty bellman equation and exploration. *CoRR*, abs/1709.05380.

Osband, I. (2016). Deep exploration via randomised value functions (phd thesis). Technical report, University of Stanford.

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3003–3011. Curran Associates, Inc.

Osband, I., Russo, D., Wen, Z., and Roy, B. V. (2017). Deep exploration via randomized value functions. *CoRR*, abs/1703.07608.

Silver, D. (2015). *Reinforcement Learning*. University College London.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292.

Weiss, N., Holmes, P., and Hardy, M. (2006). *A Course in Probability*. Pearson Addison Wesley.

Appendices

A Additional algorithm details

Here we provide additional details on each algorithm, including elaborations of the assumptions made in each case and pseudocode listings.

A.1 Bayesian Q-Learning

Dearden et al. (1998) propose the following modelling assumptions and update rule:

Assumption 1: The return $z^*_{\mathbf{s},\mathbf{a}}$ is Gaussian-distributed. If the MDP is ergodic[2] and $\gamma \approx 1$, then since the immediate rewards are independent events, one can appeal to the central limit theorem to show that $z^*_{\mathbf{s},\mathbf{a}}$ is Gaussian-distributed. This assumption will not hold in general if the MDP is not ergodic. For example, we expect certain real world, deterministic environments to not satisfy ergodicity.

Assumption 2: The prior $p(\mu_{z^*_{\mathbf{s},\mathbf{a}}}, \tau_{z^*_{\mathbf{s},\mathbf{a}}})$ is NG, and factorises over different state-actions. This is a mild assumption, which simplifies downstream calculations.

Assumption 3: The posterior $p(\mu_{z^*_{\mathbf{s},\mathbf{a}}}, \tau_{z^*_{\mathbf{s},\mathbf{a}}}|\mathcal{D})$ factors over different state-actions. This simplified distribution is a factored approximation of the true posterior. In general, we expect this assumption to fail, because we in fact know the returns from different state actions to be correlated by the BE.

Update rule: Suppose the agent observes a transition $\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}', r$. Assuming the agent greedily will follow the policy which it *thinks* to be optimal thereafter results in the following updated posterior:

$$p^{mix}_{\mathbf{s},\mathbf{a}}(\mu_{z^*_{\mathbf{s},\mathbf{a}}}, \tau_{z^*_{\mathbf{s},\mathbf{a}}}|r, \mathcal{D}) = \int p(\mu_{z^*_{\mathbf{s},\mathbf{a}}}, \tau_{z^*_{\mathbf{s},\mathbf{a}}}|r + \gamma z^*_{\mathbf{s}',\mathbf{a}'}, \mathcal{D}) p(z^*_{\mathbf{s}',\mathbf{a}'}|\mathcal{D}) \mathrm{d}z^*_{\mathbf{s}',\mathbf{a}'}. \tag{4}$$

where $\mathbf{a}' = \arg\max_{\tilde{\mathbf{a}}} z^*_{\mathbf{s}',\tilde{\mathbf{a}}}$. Because $p^{mix}_{\mathbf{s},\mathbf{a}}$ will not in general be NG-distributed, the authors propose approximating it by the NG closest to it in KL-distance. Given a distribution $q(\mu_{z^*_{\mathbf{s},\mathbf{a}}}, \tau_{z^*_{\mathbf{s},\mathbf{a}}})$, the NG $p(\mu_{z^*_{\mathbf{s},\mathbf{a}}}, \tau_{z^*_{\mathbf{s},\mathbf{a}}})$ minimising $KL(q||p)$ has parameters:

$$\begin{aligned}
\mu_{0_{\mathbf{s},\mathbf{a}}} &= \mathbb{E}_q[\mu_{z^*_{\mathbf{s},\mathbf{a}}} \tau_{z^*_{\mathbf{s},\mathbf{a}}}] / \mathbb{E}_q[\tau_{z^*_{\mathbf{s},\mathbf{a}}}], \\
\lambda_{\mathbf{s},\mathbf{a}} &= (\mathbb{E}_q[\mu^2_{z^*_{\mathbf{s},\mathbf{a}}} \tau_{z^*_{\mathbf{s},\mathbf{a}}}] - \mathbb{E}_q[\tau_{z^*_{\mathbf{s},\mathbf{a}}}] \mu^2_{0_{\mathbf{s},\mathbf{a}}})^{-1}, \\
\alpha_{\mathbf{s},\mathbf{a}} &= \max\left(1 + \epsilon, f^{-1}\left(\log \mathbb{E}_q\left[\tau_{z^*_{\mathbf{s},\mathbf{a}}}\right] - \mathbb{E}_q\left[\log \tau_{z^*_{\mathbf{s},\mathbf{a}}}\right]\right)\right), \\
\beta_{\mathbf{s},\mathbf{a}} &= \alpha_{\mathbf{s},\mathbf{a}} / \mathbb{E}_q\left[\tau_{z^*_{\mathbf{s},\mathbf{a}}}\right].
\end{aligned} \tag{5}$$

where $f(x) = \log(x) - \psi(x)$ and $\psi(x) = \Gamma'(x)/\Gamma(x)$. All $\mathbb{E}_q$ expectations are estimated by Monte Carlo. f^{-1} is analytically intractable, but can be estimated with high accuracy using bisection search, since f is monotonic. Together with Thompson sampling, this makes up BQL (algorithm 1).

Algorithm 1 Bayesian Q-Learning (BQL)

```
Initialise posterior parameters θ_Z* = (μ_0s,a, λ_s,a, α_s,a, β_s,a) for each (s, a)
for episode ∈ {1, 2, ..., N_E} do
    Observe initial state s_1
    for t ∈ {1, 2, ..., T} do
        Thompson-sample a_t from p(θ_Z* | D) and observe next state s_{t+1} and reward r_t
        θ_Z* ← Updated params. using eq. (5)
    end for
end for
```

As more data is observed and the posteriors become narrower, we hope that the agent will converge to greedy behaviour and find the optimal policy.

[2] An MDP is ergodic if, under any policy, each state-action is visited an infinite number of times and without any systematic period (Silver (2015)).

A.2 Posterior Sampling for Reinforcement Learning

For PSRL in the tabular setting we follow the approach of Osband et al. (2013), and use a Categorical-Dirichlet model for $\mathcal{T}$ and a Gaussian-NG model for $\mathcal{R}$. The posterior is updated after each episode or user-defined number of time-steps, such as the number of states in the MDP. Once the dynamics and rewards have been sampled:

$$\hat{\boldsymbol{\theta}}_{\mathcal{T}} \sim p(\boldsymbol{\theta}_{\mathcal{T}}|\mathcal{D}), \;\; \hat{\boldsymbol{\theta}}_{\mathcal{R}} \sim p(\boldsymbol{\theta}_{\mathcal{R}}|\mathcal{D}),$$

we can solve for $\hat{Q}^*|\hat{\boldsymbol{\theta}}_{\mathcal{T}}, \hat{\boldsymbol{\theta}}_{\mathcal{R}}$ and $\hat{\pi}^*|\hat{\boldsymbol{\theta}}_{\mathcal{T}}, \hat{\boldsymbol{\theta}}_{\mathcal{R}}$ by dynamical programming in the episodic setting or by policy iteration in the continuing setting. Algorithm 2 gives a pseudocode listing.

Algorithm 2 Posterior Sampling Reinforcement Learning (PSRL)

```
Initialise posteriors to priors: p(θ_T|D) ← p(θ_T) and p(θ_R|D) ← p(θ_R)
for episode ∈ {1, 2, ..., N_E} do
    Sample θ̂_T ∼ p(θ_T|D) and θ̂_R ∼ p(θ_R|D)
    Solve Bellman equation for Q̂*_{s,a} by PI and π̂*_s ← arg max_a Q̂*_{s,a}
    for t ∈ {1, 2, ..., T} do
        Observe state s_t, and take action π̂*_{s_t}
        Store transition (s_t, a_t, r_t, s_{t+1})
    end for
    Update p(θ_T|D) and p(θ_R|D) using {s_t, a_t, r_t, s_{t+1}}_{t=1}^T
end for
```

As with BQL, the posteriors will become narrower as more data are observed and the agent will converge to the true optimal policy π^*. Osband et al. (2013) formalise this intuition and prove that the regret of PSRL grows sub-linearly with the number of time-steps.

A.3 The uncertainty Bellman equation

To derive the UBE, O'Donoghue et al. (2017) make the following assumptions:

Assumption 1: The MDP is a directed acyclic graph (DAG), so each state-action can be visited at most once per episode. Any finite MDP can be turned into a DAG by a process called *unrolling*: creating T copies of each state for each time $t = 1, ..., T$. O'Donoghue et al. (2017) thus consider:

$$\mu_{z^{\pi}_{\mathbf{s},\mathbf{a},t}} = \mathbb{E}_{r,\mathbf{s}'}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}',t} + \gamma \max_{\mathbf{a}'} \mu_{z^{\pi}_{\mathbf{s}',\mathbf{a}',t+1}} \Big| \pi, \boldsymbol{\theta}_{\mathcal{T}}, \boldsymbol{\theta}_{\mathcal{R}} \right], \text{ where } \mu_{z^{\pi}_{\mathbf{s},\mathbf{a},T+1}} = 0, \forall(\mathbf{s},\mathbf{a}) \tag{6}$$

Unrolling increases data sparsity since roughly T more data would must be observed to narrow down individual posteriors by the same amount as when no unrolling is used. Further, this approach would confine the UBE to episodic tasks, so the authors choose to violate this assumption in their experiments and we follow the same approach.

Assumption 2: The mean immediate rewards of the MDP are bounded within $[-R_{max}, R_{max}]$, so the Q^* values can be upper-bounded by $TR_{\max}$ in the episodic setting and by $R_{\max}/(1-\gamma)$ in the continuing setting. We write this upper bound as Q_{max}.

Taking variances across the BE, the authors derive the upper bound:

$$\underbrace{\text{Var}_{\boldsymbol{\theta}_{\mathcal{T}},\boldsymbol{\theta}_{\mathcal{R}}}\left[\mu_{z^{\pi}_{\mathbf{s},\mathbf{a},t}}\right]}_{\text{Epistemic unc. in } \mu_{z^{\pi}_{\mathbf{s},\mathbf{a},t}}} \leq \nu^{\pi}_{\mathbf{s},\mathbf{a},t} + \mathbb{E}_{\mathbf{s}',\mathbf{a}'}\left[\underbrace{\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[p(\mathbf{s}'|\mathbf{s},\mathbf{a},\boldsymbol{\theta}_{\mathcal{T}})\right]}_{\text{Posterior predictive dynamics}} \underbrace{\text{Var}_{\boldsymbol{\theta}_{\mathcal{T}},\boldsymbol{\theta}_{\mathcal{R}}}\left[\mu_{z^{\pi}_{\mathbf{s}',\mathbf{a}',t+1}}\right]}_{\text{Epistemic unc. in } \mu_{z^{\pi}_{\mathbf{s}',\mathbf{a}',t+1}}} \,\Big|\, \pi \right] \tag{7}$$

$$\text{where } \nu^{\pi}_{\mathbf{s},\mathbf{a},t} = \underbrace{\text{Var}_{\boldsymbol{\theta}_{\mathcal{R}}}\left[\mu_{r_{\mathbf{s},\mathbf{a},\mathbf{s}',t}}\right]}_{\text{Epistemic unc. in } \mu_{r_{\mathbf{s},\mathbf{a},\mathbf{s}',t}}} + Q^2_{max} \sum_{\mathbf{s}'} \frac{\text{Var}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[p(\mathbf{s}'|\mathbf{s},\mathbf{a},\boldsymbol{\theta}_{\mathcal{T}})\right]}{\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[p(\mathbf{s}'|\mathbf{s},\mathbf{a},\boldsymbol{\theta}_{\mathcal{T}})\right]} \tag{8}$$

The bounding term in ineq. 7 is the sum of a $\nu^{\pi}_{\mathbf{s},\mathbf{a},t}$ term plus an expectation term. The former depends on quantities local to $(\mathbf{s},\mathbf{a})$, and is called the *local uncertainty*. The latter term in eq. (7) is an expectation of the next-step epistemic uncertainty weighted by the posterior predictive dynamics. It propagates the epistemic uncertainty across state-actions. Defining $\mathcal{U}_t^\pi$ as:

$$\mathcal{U}_t^\pi u_{\mathbf{s},\mathbf{a},t}^\pi = \nu_{\mathbf{s},\mathbf{a},t}^\pi + \mathbb{E}_{\mathbf{s}',\mathbf{a}'}\left[\mathbb{E}_{\boldsymbol{\theta}_\mathcal{T}}\left[p(\mathbf{s}'|\mathbf{s},\mathbf{a},\boldsymbol{\theta}_\mathcal{T})\right]u_{\mathbf{s}',\mathbf{a}',t+1}^\pi|\pi\right],$$

the authors arrive at the UBE:

$$u_{\mathbf{s},\mathbf{a},t}^\pi = \mathcal{U}_t^\pi u_{\mathbf{s},\mathbf{a},t+1}^\pi, \text{ where } u_{\mathbf{s},\mathbf{a},T+1}^\pi = 0$$

If unrolling is not applied, the bound $u_{\mathbf{s},\mathbf{a},t}^\pi$ is no longer strictly true and the UBE becomes a heuristic:

$$u_{\mathbf{s},\mathbf{a}}^\pi = \mathcal{U}^\pi u_{\mathbf{s},\mathbf{a}}^\pi. \tag{9}$$

We can first obtain the greedy policy π^*, through PI. Subsequently we solve for the fixed point of the UBE, without unrolling, to obtain $u_{\mathbf{s},\mathbf{a}}^*$. Introducing the scaling factor ζ we finally use $u_{\mathbf{s},\mathbf{a}}^*$ for Thompson sampling from a diagonal gaussian. This amounts to a factored posterior approximation. Algorithm 3 shows the complete process.

Algorithm 3 Uncertainty Bellman Equation with Thompson sampling

1: Input data $\mathcal{D}$ and posteriors $p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D})$, $p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$
2: Solve for greedy policy π^* through PI
3: Solve for $u_{\mathbf{s},\mathbf{a}}^*$ in eq. (9)
4: **for** $t \in \{1, 2, ..., T_{\max}\}$ **do**
5: Observe $\mathbf{s}_t$
6: Thompson-sample $\mathbf{a}_t = \arg\max_\mathbf{a}\left(\mu_{z_{\mathbf{s},\mathbf{a}}^*} + \zeta\epsilon_{\mathbf{s},\mathbf{a}}\left(u_{\mathbf{s},\mathbf{a}}^*\right)^{1/2}\right)$, $\epsilon_{\mathbf{s},\mathbf{a}} \sim \mathcal{N}(0,1)$
7: Observe $\mathbf{s}_{t+1}, r_t$ and store $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in $\mathcal{D}$.
8: **end for**
9: Update $p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D})$, $p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$ and go back to 2

Note that as the posterior variance collapses to 0 in the limit of infinite data, $\nu_{\mathbf{s},\mathbf{a},t}^\pi \to 0$ because both terms in eq. (8) also tend to 0. Therefore, we also have $u_{\mathbf{s},\mathbf{a},t}^\pi \to 0$, and the agent will automatically transition to greedy behaviour.

A.4 Moment matching across the BE

Starting from the Bellman relation for $w_\mathbf{s}^\pi$:

$$w_\mathbf{s}^\pi = r_{\mathbf{s},\mathbf{a},\mathbf{s}'} + \gamma w_{\mathbf{s}'}^\pi,$$

where $\mathbf{s}' \sim p(\mathbf{s}'|\mathbf{s},\mathbf{a})$, $\mathbf{a} \sim \pi(\mathbf{s})$, we require equality between the first and second order moments[3]:

$$\mathbb{E}_{w,\boldsymbol{\theta}_\mathcal{W}}\left[w_\mathbf{s}^\pi\right] = \mathbb{E}_{r,\boldsymbol{\theta}_\mathcal{R},w,\boldsymbol{\theta}_\mathcal{W},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\mathbf{a}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'} + \gamma w_{\mathbf{s}'}^\pi|\pi\right] \tag{10}$$

$$\text{Var}_{w,\boldsymbol{\theta}_\mathcal{W}}\left[w_\mathbf{s}^\pi\right] = \text{Var}_{r,\boldsymbol{\theta}_\mathcal{R},w,\boldsymbol{\theta}_\mathcal{W},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\mathbf{a}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'} + \gamma w_{\mathbf{s}'}^\pi|\pi\right] \tag{11}$$

Equation (10) is the familiar value-BE, which can be used to compute the greedy policy by PI. Equation (11) can be expanded on both sides to express a similar equality between variances. First, using the law of total variance on the LHS:

$$\underbrace{\text{Var}_{w,\boldsymbol{\theta}_\mathcal{W}}\left[w_\mathbf{s}^\pi\right]}_{\text{Total value variance}} = \underbrace{\text{Var}_{\boldsymbol{\theta}_\mathcal{W}}\left[\mathbb{E}_w\left[w_\mathbf{s}^\pi|\boldsymbol{\theta}_\mathcal{W}\right]\right]}_{\text{Epistemic value variance}} + \underbrace{\mathbb{E}_{\boldsymbol{\theta}_\mathcal{W}}\left[\text{Var}_w\left[w_\mathbf{s}^\pi|\boldsymbol{\theta}_\mathcal{W}\right]\right]}_{\text{Aleatoric value variance}}.$$

Second, we expand the RHS of eq. (11) and obtain

$$\underbrace{\text{Var}_{w,\boldsymbol{\theta}_\mathcal{W}}\left[w_\mathbf{s}^\pi\right]}_{\text{Total value variance}} = \underbrace{\text{Var}_{r,\boldsymbol{\theta}_\mathcal{R},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\mathbf{a}}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}]}_{\text{Reward variance}} + 2\gamma\underbrace{\text{Cov}_{r,\boldsymbol{\theta}_\mathcal{R},w,\boldsymbol{\theta}_\mathcal{W},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\mathbf{a}}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}, w_{\mathbf{s}'}^\pi]}_{\text{Reward-value covariance}} +$$

$$\gamma^2\underbrace{\text{Var}_{w,\boldsymbol{\theta}_\mathcal{W},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}[w_{\mathbf{s}'}^\pi]}_{\text{Next-step value variance}}. \tag{12}$$

Each of the terms in eq. (12) contains contributions from aleatoric as well as epistemic sources, which can be separated using the laws of total variance and total covariance (Weiss et al. (2006))- the decompositions are straightforward but lengthy and are omitted for brevity.

[3]Expectations and variances are over the posteriors of the subscript variables conditioned on data $\mathcal{D}$.

Since each uncertainty comes from a different source, we argue that one BE should be satisfied for each. We therefore obtain the following consistency equation for the epistemic terms:

$$
\begin{aligned}
\underbrace{\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{W}}}\left[\mathbb{E}_{w}\left[w^{\pi}_{\mathbf{s}}|\boldsymbol{\theta}_{\mathcal{W}}\right]\right]}_{\text{Epistemic value variance}} = & \underbrace{\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_{\mathcal{R}},\mathbf{a}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{Variance of expected reward}\\ \text{due to } \boldsymbol{\theta}_{\mathcal{T}} \text{ uncertainty}}} \\
& + \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{R}}}\left[\mathbb{E}_{r,\mathbf{a}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}},\boldsymbol{\theta}_{\mathcal{R}}\right]\right]\right]}_{\substack{\text{Expectation of reward variance due}\\ \text{due to } \boldsymbol{\theta}_{\mathcal{R}} \text{ uncertainty}}} + \\
& + 2\gamma \underbrace{\mathrm{Cov}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_{\mathcal{R}},\mathbf{a}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_{\mathcal{T}}\right], \mathbb{E}_{\mathbf{s}',w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^{\pi}_{\mathbf{s}'}|\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{Covariance of reward and value expectations}\\ \text{due to } \boldsymbol{\theta}_{\mathcal{T}} \text{ uncertainty}}} \\
& + \gamma^2 \underbrace{\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{\mathbf{s}',w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^{\pi}_{\mathbf{s}'}|\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{Value variance due to dynamics}\\ \text{purely epistemic}}} \\
& + \gamma^2 \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{W}}}\left[\mathbb{E}_{w}\left[w^{\pi}_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{W}}\right]\right]\right]}_{\substack{\text{Expectation of value variance}\\ \text{due to } \boldsymbol{\theta}_{\mathcal{W}} \text{ uncertainty}}}
\end{aligned} \tag{13}
$$

With the exception of the last term in eq. (13), all RHS terms can be readily computed provided we already have $\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{W}}}\left[\mu_{w^{\pi}_{\mathbf{s}}}\right]$ from eq. (10). We observe that the last term is the same as the LHS term, except it has been smoothed out w.r.t. the next-state posterior predictive. Therefore, eq. (13) is a system of linear equation which we can solve in $O(|\mathcal{S}|^3)$ time for the epistemic uncertainty.

So far we considered the variance in $\mu_{w^{\pi}_{\mathbf{s}}}$, however for action selection we need uncertainties state-actions, that is over $\mu_{z^{\pi}_{\mathbf{s},\mathbf{a}}}$. After calculating $\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}[\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{W}}}[\mu_{w^{\pi}_{\mathbf{s}'}}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{W}}]]$ we can substitute for all terms in eq. (13) and evaluate the RHS without integrating out $\mathbf{a}$. This gives the epistemic variance in $\mu_{z^{\pi}_{\mathbf{s},\mathbf{a}}}$ which we can use for Thompson sampling from a diagonal Gaussian, for the case $\pi = \pi^*$:

$$
\begin{aligned}
\mathbf{a} = \arg\max_{\mathbf{a}'}\big(\mu_{z^{\pi^*}_{\mathbf{s},\mathbf{a}'}} + \zeta\epsilon_{\mathbf{s},\mathbf{a}'}\,\tilde{\sigma}_{z^*_{\mathbf{s},\mathbf{a}'}}\big),\\
\text{where } \epsilon_{\mathbf{s},\mathbf{a}} \sim \mathcal{N}(0,1), \text{ and } \tilde{\sigma}^2_{z^*_{\mathbf{s},\mathbf{a}}} = \mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}[\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{Z}}}[\mu_{z^*_{\mathbf{s},\mathbf{a}}}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{Z}}]].
\end{aligned}
$$

ζ can be adjusted as with the UBE, although we do not find this is necessary in our tabular experiments and use $\zeta = 1.00$ throughout.

Algorithm 4 Moment Matching with Thompson sampling

1: Input data $\mathcal{D}$ and posteriors $p(\boldsymbol{\theta}_{\mathcal{T}}|\mathcal{D})$, $p(\boldsymbol{\theta}_{\mathcal{R}}|\mathcal{D})$
2: Compute greedy policy π^* by PI
3: Compute epistemic uncertainty $\tilde{\sigma}^2_{z^*_{\mathbf{s},\mathbf{a}}}$ (eq. (13) and procedure described in text)
4: **for** $t \in \{1, 2, ..., T_{\max}\}$ **do**
5: Observe $\mathbf{s}_t$
6: Thompson-sample and execute $\mathbf{a}_t = \arg\max_{\mathbf{a}}\left(\mu_{z^*_{\mathbf{s}_t,\mathbf{a}}} + \epsilon_{\mathbf{s}_t,\mathbf{a}}\,\tilde{\sigma}_{z^*_{\mathbf{s}_t,\mathbf{a}}}\right)$
7: Observe $\mathbf{s}_{t+1}, r_t$ and store $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in $\mathcal{D}$.
8: **end for**
9: Update posteriors $p(\boldsymbol{\theta}_{\mathcal{T}}|\mathcal{D})$, $p(\boldsymbol{\theta}_{\mathcal{R}}|\mathcal{D})$ and go back to 2

B Additional environment details

B.1 DeepSea

Our DeepSea MDP (fig. 1) is a variant of the ones used in Osband et al. (2017); O'Donoghue (2018). The agent starts from $\mathbf{s}_1$ and can choose swim-*left* or swim-*right* from each of the N states in the environment.

Swim-*left* always succeeds and moves the agent to the left, giving $r = 0$ (red transitions). Swim-*right* from $\mathbf{s}_1, ..., \mathbf{s}_{N-1}$ succeeds with probability $1 - 1/N$, moving the agent to the right and otherwise fails moving the agent to the left (blue arrows), giving $r = -\delta$ regardless of whether it succeeds. A successful swim-*right* from $\mathbf{s}_N$ moves the agent back to $\mathbf{s}_1$ and gives $r = 1$. We choose δ so that *right* is always optimal[4].

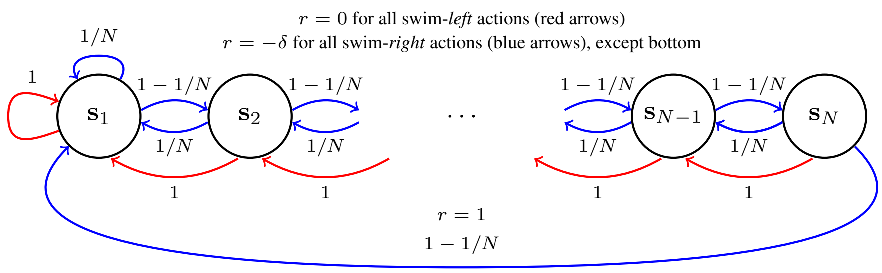

Figure 1: DeepSea MDP from the continuing setting, modified from O'Donoghue (2018). Blue arrows correspond to swim-*right* (optimal) and red arrows to swim-*left* (sub-optimal).

This environment is designed to test whether the agent continues exploring despite receiving negative rewards. Sustained exploration becomes increasingly important for large N. As argued in Osband (2016), in order to avoid exponentially poor performance, exploration in such chain-like environments must be guided by uncertainty rather than randomness.

B.2 WideNarrow

The WideNarrow MDP (fig. 2) has $2N + 1$ states and deterministic transitions. Odd states except $\mathbf{s}_{2N+1}$ have W actions, out of which one gives $r \sim \mathcal{N}(\mu_h, \sigma^2)$ whereas all others give $r \sim \mathcal{N}(\mu_l, \sigma^2)$, with $\mu_l < \mu_h$. Even states have a single action also giving $r \sim \mathcal{N}(\mu_l, \sigma^2)$. In our experiments we use $\mu_h = 0.5$, $\mu_l = 0$ and $\sigma_h = \sigma_l = 1$.

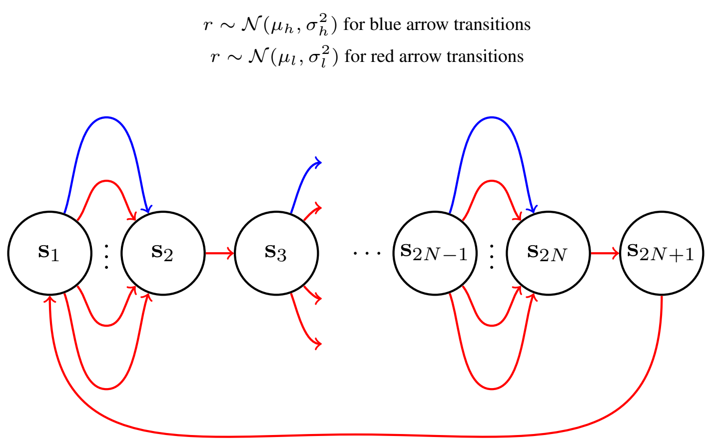

Figure 2: The WideNarrow MDP. All transitions are deterministic.

[4] We choose $\delta = 0.1 \times \exp^{-N/4}$ in our experiments, which guarantees *right* is optimal at least up to $N = 40$.

In general, the returns from different state-actions will be correlated under the posterior. Here, consider $(\mathbf{s}_1, \mathbf{a}_1)$ and $(\mathbf{s}_1, \mathbf{a}_2)$:

$$\begin{aligned} \mathrm{Cov}_{z,\boldsymbol{\theta}}\left[z^*_{\mathbf{s}_1,\mathbf{a}_1}, z^*_{\mathbf{s}_1,\mathbf{a}_2}\right] =& \mathrm{Cov}_{r,z,\boldsymbol{\theta}}\left[r_{\mathbf{s}_1,\mathbf{a}_1,\mathbf{s}'} + \gamma z^*_{\mathbf{s}',\mathbf{a}'},\ r_{\mathbf{s}_1,\mathbf{a}_2,\mathbf{s}''} + \gamma z^*_{\mathbf{s}'',\mathbf{a}''}\right] \\ =& \cancel{\mathrm{Cov}_{r,z,\boldsymbol{\theta}}\left[r_{\mathbf{s}_1,\mathbf{a}_1,\mathbf{s}'}, r_{\mathbf{s}_1,\mathbf{a}_2,\mathbf{s}''}\right]} + \gamma\,\mathrm{Cov}_{r,\boldsymbol{\theta}}\left[r_{\mathbf{s}_1,\mathbf{a}_1,\mathbf{s}'}, z^*_{\mathbf{s}'',\mathbf{a}''}\right] \\ &+ \gamma\,\mathrm{Cov}_{r,z,\boldsymbol{\theta}}\left[r_{\mathbf{s}_1,\mathbf{a}_2,\mathbf{s}''}, z^*_{\mathbf{s}'',\mathbf{a}''}\right] + \gamma^2\,\mathrm{Cov}_{z,\boldsymbol{\theta}}\left[z^*_{\mathbf{s}',\mathbf{a}'}, z^*_{\mathbf{s}'',\mathbf{a}''}\right] \end{aligned} \tag{14}$$

where $\boldsymbol{\theta}$ loosely denotes all modelling parameters, s' denotes the next-state from $\mathbf{s}_1, \mathbf{a}_1$, $\mathbf{s}''$ denotes the next-state from $\mathbf{s}_1, \mathbf{a}_2$ and $\mathbf{a}', \mathbf{a}''$ denote the corresponding next-actions. Although the remaining three terms are non-zero under the posterior, BQL, UBE and MM ignore them, instead sampling from a factored posterior. The WideNarrow environment enforces strong correlations between these state actions, through the last term in eq. (14), allowing us to test the impact of a factored approximation.

B.3 PriorMDP

The aforementioned MDPs have very specific and handcrafted dynamics and rewards, so it is interesting to also compare the algorithms on environments which lack this sort of structure. For this we sample finite MDPs with N_s states and N_a action from a prior distribution, as in Osband et al. (2013). $\mathcal{T}$ is a Categorical with parameters $\{\boldsymbol{\kappa}_{\mathbf{s},\mathbf{a}}\}$ with:

$$\boldsymbol{\kappa}_{\mathbf{s},\mathbf{a}} \sim \text{Dirichlet}(\boldsymbol{c}_{\mathbf{s},\mathbf{a}}),$$

with pseudo-count parameters $\boldsymbol{c}_{\mathbf{s},\mathbf{a}} = \mathbf{1}$, while $\mathcal{R} \sim \mathcal{N}(\mu_{\mathbf{s},\mathbf{a}}, \tau^{-1}_{\mathbf{s},\mathbf{a}})$ with:

$$\mu_{\mathbf{s},\mathbf{a}}, \tau_{\mathbf{s},\mathbf{a}} \sim NG(\mu_{\mathbf{s},\mathbf{a}}, \tau_{\mathbf{s},\mathbf{a}}|\mu, \lambda, \alpha, \beta) \text{ with } (\mu, \lambda, \alpha, \beta) = (0.00, 3.00 \times 10^2, 4.00, 4.00).$$

We chose these hyperparameters because they give Q^*-values in a reasonable range.