Bayesian methods for efficient Reinforcement Learning in tabular problems

Anonymous Author(s)
Affiliation
Address
email

Abstract

The exploration-exploitation tradeoff is one of the central problems of Reinforcement Learning (RL). We explore how Bayesian modelling can be incorporated into RL to tackle this tradeoff, by quantifying relevant epistemic uncertainties and using them to efficiently guide the exploration. Empirical results of four Bayesian methods in the tabular setting - Bayesian Q-Learning (BQL), posterior sampling for RL (PSRL), the uncertainty Bellman equation (UBE) and our own moment matching (MM) approach - shows evidence that: BQL may suffer from a pathology whereby early incorrect posterior updates result in an overconfident and inaccurate posterior; the UBE greatly over-estimates uncertainties and places a much heavier emphasis on the dynamics than the rewards uncertainties; MM gives generally well-calibrated uncertainty estimates; factored posterior approximations (BQL, UBE, MM) have adverse effects on regret performance while PSRL, which involves posterior correlations, does not face the same issue.

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 observed data are already 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 $t \to \infty$. 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*: instead of considering the agent's *uncertainty* and they drive exploration by injecting random noise in action selection. Further, robust methods for annealing the exploration parameters (ϵ or T) have yet to be found in the literature and most practical applications do not use annealing at all (Mnih et al. (2015)), at the expense of crude exploration schemes.

To explore efficiently, action-selection must be *directed*: it must be guided by a quantification of the agent's uncertainty - Bayesian modelling is a natural framework for this quantification. By representing the agent's posterior beliefs and selecting actions accordingly, the exploration becomes guided by the degree of uncertainty. Further, such an approach offers an intuitive and principled *transition mechanism* from exploration to exploitation: the posteriors shrink and the agent converges to the optimal policy as further data are observed. In this work we present a number of Bayesian algorithms in tabular Markov Decision Processes (MDPs) including our own approach. We compare the algorithms' behaviour and explain differences in performance, yielding several important insights.

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 or greedy policy. In addition to V^π and Q^π to denote state and action values under π, we define the state and action *return* random variables $w^\pi_{\mathbf{s}}$ and $z^\pi_{\mathbf{s},\mathbf{a}}$,

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

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*:

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

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 $Q^*_{\mathbf{s},\mathbf{a}}$. Bellemare et al. (2017) postulate that DRL improves performance 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 $\mathcal{R}$ exactly, it will not be able to perfectly predict $z^*_{\mathbf{s},\mathbf{a}}$ if $\mathcal{T}$ and $\mathcal{R}$ are stochastic. Modelling the aleatoric uncertainty may lead to more meaningful models of the return but is not useful for improving exploration. In addition to aleatoric uncertainty, there will also be uncertainty about the parameterisation of $\mathcal{Z}^*$ due to the finite amount of data collected by the agent, known as *epistemic* uncertainty. This decreases as more data are observed and expresses the agent's belief for quantities such as the *expected returns*. The agent should therefore take this reducible uncertainty into account when exploring, since actions may be better or worse the current estimate.

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 posteriors become 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 $p(\boldsymbol{\theta})$:

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

Choosing a *conjugate* prior simplifies downstream calculations: for discrete distributions such as $\mathcal{T}$, we use a Categorical-Dirichlet model (Bishop (2006)) for each $\mathbf{s}, \mathbf{a}$, while for continuous distributions such as $\mathcal{R}, \mathcal{W}, \mathcal{Z}$ we use a Normal-NG model (Murphy (2007)) 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 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. In our experiments, we see evidence that this update rule is problematic. 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 $\text{Var}_{\boldsymbol{\theta}_{\mathcal{T}},\boldsymbol{\theta}_{\mathcal{R}}}\left[\mu_{z^{\pi}_{\mathbf{s},\mathbf{a},t}}\right]$. 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 $\zeta^2 u^*_{\mathbf{s},\mathbf{a},t}$, where ζ is an appropriate scaling factor. Like BQL, this is also 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 BEs. 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 $\mu_{z^*_{\mathbf{s},\mathbf{a}}}$. The latter is used for Thompson sampling from a diagonal gaussian, resulting in a factored approximation of the posterior as in the UBE. An outline of the uncertainty decomposition and further details are given in appendix A.4.

[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.

4 Environments and methods

We compare the algorithms on three kinds of finite MDPs of variable sizes, and all experiments[2] are in the continuing setting - exact specifications and illustrations given in section B. 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 (2018a), and is aimed at testing the algorithm's ability for sustained exploration despite initially receiving 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 PriorMDP.

5 Results and discussion

Visualisations of the posterior evolution and cumulative regret to an oracle on small MDPs, illustrate a number of interesting phenomena (figs. 7 to 13, figs. 14 to 19 and figs. 22 to 27). A summary of regret performance is shown in fig. 1 while more extensive results are given in figs. 4 to 6.

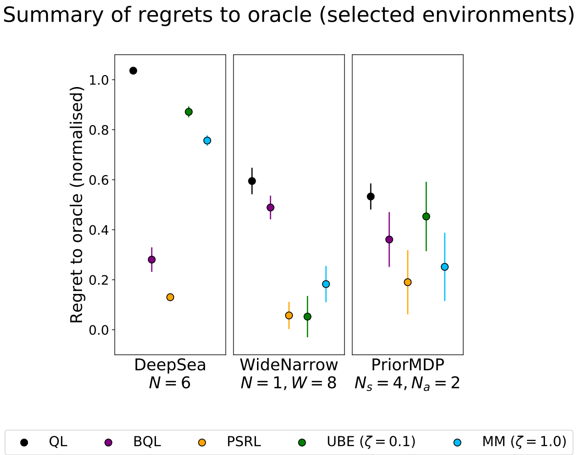


Figure 1: Summary of regret performances to oracle on selected environments - see figs. 4 to 6 for additional results.

We often observe that as training progresses, the posteriors concentrate on the true Q^* values, the behaviour policy converges on the optimal one and the agent smoothly transitions into greedy action selection. Further, the agent does not over-explore actions if it is confident that these are suboptimal. This is notably seen in figs. 8 to 13. There, although there is significant uncertainty in the expected return of the suboptimal action, the agent is confident that the optimal action $(\mathbf{s} = 4, \mathbf{a} = \textit{right})$ is better than the suboptimal one $(\mathbf{s} = 4, \mathbf{a} = \textit{left})$: the agent does not spend its time determining the exact expected return of an action if it is confident that it is suboptimal. These two behaviours are central for achieving a principled and efficient approach to exploration, however we often observe exceptions where the agent does not perform in such a way.

First, the UBE uncertainty estimate $u^*_{\mathbf{s},\mathbf{a}}$ remains extremely loose even after a large number of time-steps (fig. 10, fig. 17 and fig. 26). Even though $\mu_{z^*_{\mathbf{s},\mathbf{a}}}$ be close to Q^*, $u^*_{\mathbf{s},\mathbf{a}}$ is so large that the Thompson noise completely smooths out differences between actions, which are picked almost uniformly at random. Further, $u^*_{\mathbf{s},\mathbf{a}}$ shrinks very slowly and the transition to greedy behaviour takes an extremely long time, causing poor regret performance. These effects are due to the contribution of an extremely large term coming from the upper-bound derivation of O'Donoghue et al. (2017) - this is the Q_{max} term in eq. (7) and eq. (8)). Further, this term depends solely on the dynamics model, so $u^*_{\mathbf{s},\mathbf{a}}$ is dominated by the dynamics uncertainty, while the rewards uncertainty is much

[2] Implementations of the agents and environments, as well as notebooks for plotting all figures in this work are available at `https://github.com/sample-efficient-bayesian-rl`.

smaller (fig. 12). Scaling the Thompson noise by $\zeta < 1.0$, improves regret performance in some cases (e.g. fig. 11). However, one is further faced by the challenge of tuning ζ, which may be expensive and challenging for large problems. By contrast, MM produces more well-calibrated uncertainty estimates than the UBE (see fig. 13, fig. 19 and fig. 27). As a result, MM shows typically better regret performance than UBE without a need to tune ζ. This could give an advantage to MM over the UBE in settings where tuning may be expensive or difficult.

Second, we observe that the BQL posterior sometimes fails to concentrate on the true Q^* values (e.g. fig. 8 and fig. 23), where the posterior is overconfident about incorrect predictions of $\mu_{z^*_{\mathbf{s},\mathbf{a}}}$. This effect persists for different random seeds and is affected by the prior used. In particular, using an NG prior with a mean μ_0 that is closer to the true Q^* values, results in the posterior concentrating on the true Q^*. These effects can be explained through the update rule used in BQL (eq. (4)). The update rule uses the next-state-action posterior $p(z^*_{\mathbf{s}',\mathbf{a}'}|\mathcal{D})$ to update the current state-action posterior. If the former is inaccurate and overconfident, the updated hyperparameters are affected accordingly. BQL can hardly escape from this situation because it does not involve a *forgetting mechanism* for inaccurate updates far in the past. Contrast this with Q-Learning, in which the Temporal Difference (TD) updates result in a moving average of observed rewards. Model-free Bayesian approaches with a rule similar to BQL may suffer from a similar pathology.

Third, there is strong evidence that factored approximations made by BQL, UBE and MM have a significant effect on regret performance. Factored approximations result in overly loose posteriors (see fig. 20, fig. 21, fig. 28 and fig. 29) and as a result, the Thompson-sampled $\mu_{z^*_{\mathbf{s},\mathbf{a}}}$ often correspond to picking a sub-optimal action[3]. By contrast, PSRL draws samples from the exact posterior and thereby accounts for correlations between different state-actions, which are in fact quite significant. The exact posterior often has marginals of similar scale as those of BQL or MM, however by incorporating correlations PSRL selects optimal actions much more often and thus achieves a better regret performance. Accounting for these correlations is an important factor in ensuring the transition from exploration to exploitation occurs quickly enough. PSRL typically outperforms BQL, UBE and MM as a result of these correlations at no additional computational cost.

6 Conclusions & Further work

Our comparison of BQL, PSRL, UBE and MM has yielded a number of insights about these algorithms: BQL suffers from a pathology whereby incorrect posterior updates result in an overconfident posterior in the absence of a forgetting mechanism, an effect from which other model-free approaches without forgetting may suffer from; the UBE uncertainty estimate $u^*_{\mathbf{s},\mathbf{a}}$ is extremely loose, results in undirected exploration if ζ is not tuned and places a much larger emphasis on the dynamics than the rewards uncertainties; factored approximations to the posterior as those in BQL, UBE and MM, have adverse effects on regret performance, while PSRL does not suffer from the same phenomenon since it samples from the true posterior with correlations; MM gives generally well-calibrated uncertainty estimates, however it still suffers from the factored posterior approximation. There are several interesting directions for further work:

- PSRL outperformed the other methods in our experiments. Inspired by this one could explore how to extend PSRL to tasks with continuous state-actions, for example by using Gaussian Process (GP) Rasmussen and Williams (2006).
- MM can also be performed in continuous state-action tasks. We have conducted preliminary work for GP-based MM, using the approaches from Rasmussen and Kuss (2004); Quiñonero-Candela et al. (2002) but further investigation is needed for this work to come to fruition.
- Devising a principled forgetting mechanism for BQL and examining whether this remedies the observed pathology would be an interesting direction of work.
- A comparison with other approaches such as Azizzadenesheli et al. (2018), Janz et al. (2019) and O'Donoghue (2018b) would give a more complete picture of performance across a broader set of algorithms.
- We have used Thompson sampling, however alternative action selection methods, such as those presented in Dearden et al. (1998), could be explored.

[3] In our Thompson-sample plots, an action is optimal only if the sample lies on the same side of the black dashed line as the black cross, across all plots.

References

Azizzadenesheli, K., Brunskill, E., and Anandkumar, A. (2018). Efficient exploration through bayesian deep q-networks. *CoRR*, abs/1802.04412.

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.

Janz, D., Hron, J., Hernández-Lobato, J. M., Hofmann, K., and Tschiatschek, S. (2019). Successor uncertainties: exploration and uncertainty in temporal difference learning.

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. (2018a). Variational bayesian reinforcement learning with regret bounds. *CoRR*, abs/1807.09647.

O'Donoghue, B. (2018b). 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.

Quiñonero-Candela, J., Girard, A., and Rasmussen, C. E. (2002). Prediction at an uncertain input for gaussian processes and relevance vector machines application to multiple-step ahead time-series forecasting. Technical report.

Rasmussen, C. and Kuss, M. (2004). Gaussian processes in reinforcement learning. In *Advances in Neural Information Processing Systems 16*, pages 751–759, Cambridge, MA, USA. Max-Planck-Gesellschaft, MIT Press.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA.

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. For all Dirichlet priors we use hyperparameters $\boldsymbol{\eta}_{\mathbf{s},\mathbf{a}} = \mathbf{1}$ and for all NG priors we use $(\mu_0, \lambda, \alpha, \beta)_{\mathbf{s},\mathbf{a}} = (0.0, 4.0, 3.0, 3.0)$.

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[4] 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 $\boldsymbol{\theta}_{\mathcal{Z}^*} = (\mu_{0_{\mathbf{s},\mathbf{a}}}, \lambda_{\mathbf{s},\mathbf{a}}, \alpha_{\mathbf{s},\mathbf{a}}, \beta_{\mathbf{s},\mathbf{a}})$ for each $(\mathbf{s}, \mathbf{a})$
Observe initial state $\mathbf{s}_1$
for time-step $\in \{0, 1, ..., T_{\max} - 1\}$ do
    Thompson-sample $\mathbf{a}_t$ using $p(\boldsymbol{\theta}_{\mathcal{Z}^*}|\mathcal{D})$ and observe next state $\mathbf{s}_{t+1}$ and reward $r_t$
    $\boldsymbol{\theta}_{\mathcal{Z}^*} \leftarrow$ Updated params. using Monte Carlo on eq. (5)
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.

[4] 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 (PI) in the continuing setting. Algorithm 2 gives a pseudocode listing.

Algorithm 2 Posterior Sampling Reinforcement Learning (PSRL)

```
Input data $\mathcal{D}$ and posteriors $p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D}), p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$
for t $\in \{0, 1, ..., T_{max} - 1\}$ do
    if $t$ % $T_\text{update}$ == $0$ then
        Update $p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D})$ and $p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$ using observed data
        Sample $\hat{\boldsymbol{\theta}}_\mathcal{T} \sim p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D})$ and $\hat{\boldsymbol{\theta}}_\mathcal{R} \sim p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$
        Solve Bellman equation for $\hat{Q}^*_{\mathbf{s},\mathbf{a}}$ by PI and $\hat{\pi}^*_\mathbf{s} \leftarrow \arg\max_\mathbf{a} \hat{Q}^*_{\mathbf{s},\mathbf{a}}$
    end if
    Observe state $\mathbf{s}_t$ and take action $\hat{\pi}^*_{\mathbf{s}_t}$
    Store $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in $\mathcal{D}$
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. 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:

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

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 $\mu_{z^\pi_{\mathbf{s},\mathbf{a},t}}$ values can be upper-bounded by TR_max in the episodic setting and by $R_\text{max}/(1-\gamma)$ in the continuing setting. We write this upper bound as Q_{max}.

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

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

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

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:

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

the authors arrive at the UBE:

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

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

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

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

```
Input data $\mathcal{D}$ and posteriors $p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D}), p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$
for $t \in \{0, 1, ..., T_{\max} - 1\}$ do
    if $t$ % $T_{\text{update}}$ == $0$ then
        Update $p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D})$ and $p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$ using observed data
        Solve for greedy policy $\pi^*$ by PI
        Solve for $u^*_{\mathbf{s},\mathbf{a}}$ in eq. (9)
    end if
    Observe $\mathbf{s}_t$
    Thompson-sample $\mathbf{a}_t = \arg\max_\mathbf{a} \left(\mu_{z^*_{\mathbf{s},\mathbf{a}}} + \zeta\epsilon_{\mathbf{s},\mathbf{a}} \left(u^*_{\mathbf{s},\mathbf{a}}\right)^{1/2}\right)$, $\epsilon_{\mathbf{s},\mathbf{a}} \sim \mathcal{N}(0,1)$
    Observe $\mathbf{s}_{t+1}, r_t$ and store $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in $\mathcal{D}$
end for
```

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

A.4 Moment matching across the BE

Starting from the Bellman relation for $z^\pi_{\mathbf{s},\mathbf{a}}$:

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

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[5]:

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

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

Equation (10) is the familiar BE for Q^π, 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:

$$\underbrace{\text{Var}_{z,\boldsymbol{\theta}_\mathcal{Z}}\left[z^\pi_{\mathbf{s},\mathbf{a}}\right]}_{\text{Total value variance}} = \underbrace{\text{Var}_{\boldsymbol{\theta}_\mathcal{Z}}\left[\mathbb{E}_z\left[z^\pi_{\mathbf{s},\mathbf{a}}|\boldsymbol{\theta}_\mathcal{Z}\right]\right]}_{\text{Epistemic value variance}} + \underbrace{\mathbb{E}_{\boldsymbol{\theta}_\mathcal{Z}}\left[\text{Var}_z\left[z^\pi_{\mathbf{s},\mathbf{a}}|\boldsymbol{\theta}_\mathcal{Z}\right]\right]}_{\text{Aleatoric value variance}}.$$

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

$$\underbrace{\text{Var}_{z,\boldsymbol{\theta}_\mathcal{W}}\left[z^\pi_{\mathbf{s},\mathbf{a}}\right]}_{\text{Total value variance}} = \underbrace{\text{Var}_{r,\boldsymbol{\theta}_\mathcal{R},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}]}_{\text{Reward variance}} + 2\gamma\underbrace{\text{Cov}_{r,\boldsymbol{\theta}_\mathcal{R},z,\boldsymbol{\theta}_\mathcal{Z},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\mathbf{a}'}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}, z^\pi_{\mathbf{s}',\mathbf{a}'}]}_{\text{Reward-value covariance}}$$
$$+\gamma^2\underbrace{\text{Var}_{z,\boldsymbol{\theta}_\mathcal{Z},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\mathbf{a}'}[z^\pi_{\mathbf{s}',\mathbf{a}'}]}_{\text{Next-step value variance}}. \tag{12}$$

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

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 included in the supporting material.

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{Z}}\left[\mathbb{E}_z\left[z^\pi_{\mathbf{s},\mathbf{a}}|\boldsymbol{\theta}_\mathcal{Z}\right]\right]}_{\text{Epistemic action-return unc.}} = & \underbrace{\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic reward unc. from}\\ \text{dynamics unc.}}} \\ & + \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{R}}\left[\mathbb{E}_r\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T},\boldsymbol{\theta}_\mathcal{R}\right]\right]\right]}_{\substack{\text{Epistemic rewards unc. from}\\ \text{rewards unc.}}} + \\ & + 2\gamma\underbrace{\mathrm{Cov}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right],\mathbb{E}_{\mathbf{s}',z,\boldsymbol{\theta}_\mathcal{Z},\mathbf{a}'}\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic reward and action-return covariance}\\ \text{from dynamics unc.}}} \\ & + \gamma^2\underbrace{\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',z,\boldsymbol{\theta}_\mathcal{Z},\mathbf{a}'}\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic action-return unc. from}\\ \text{dynamics unc.}}} \\ & + \gamma^2\underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\mathbf{a}'}\left[\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{Z}}\left[\mathbb{E}_z\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{Z}\right]\right]\right]}_{\substack{\text{Epistemic action-return unc. from}\\ \text{state-return unc.}}}\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 $\mathbb{E}_{\mathbf{s}',z,\boldsymbol{\theta}_\mathcal{Z}}\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\boldsymbol{\theta}_\mathcal{T}\right]$ 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 equations which can be solved in $O(|\mathcal{S}|^3|\mathcal{A}|^3)$ time for the epistemic uncertainty in $\mu_{z^\pi_{\mathbf{s},\mathbf{a}}}$. The latter can be subsequently used for Thompson sampling from a diagonal Gaussian:

$$\begin{gathered}\mathbf{a} = \arg\max_{\mathbf{a}'}\left(\mu_{z^*_{\mathbf{s},\mathbf{a}'}} + \zeta\epsilon_{\mathbf{s},\mathbf{a}'}\,\tilde{\sigma}_{z^*_{\mathbf{s},\mathbf{a}'}}\right), \\ \text{where } \epsilon_{\mathbf{s},\mathbf{a}} \sim \mathcal{N}(0,1), \text{ and } \tilde{\sigma}^2_{z^*_{\mathbf{s},\mathbf{a}}} = \mathrm{Var}_{\boldsymbol{\theta}_\mathcal{Z}}\left[\mathbb{E}_z\left[z^\pi_{\mathbf{s},\mathbf{a}}|\boldsymbol{\theta}_\mathcal{Z}\right]\right],\end{gathered}$$

where $\pi = \pi^*$ has been used. ζ can be adjusted as with the UBE, although we do not find this is necessary in our tabular experiments and use $\zeta = 1.0$ throughout.

Algorithm 4 Moment Matching with Thompson sampling

```
Input data $\mathcal{D}$ and posteriors $p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D})$, $p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$
for $t \in \{0, 1, ..., T_{\max} - 1\}$ do
    if $t$ % $T_{\text{update}}$ == $0$ then
        Update $p(\boldsymbol{\theta}_\mathcal{T}|\mathcal{D})$ and $p(\boldsymbol{\theta}_\mathcal{R}|\mathcal{D})$ using observed data
        Solve for greedy policy $\pi^*$ by PI
        Compute epistemic uncertainty $\tilde{\sigma}^2_{z^*_{\mathbf{s},\mathbf{a}}}$ by solving eq. (13)
    end if
    Observe $\mathbf{s}_t$
    Thompson-sample and execute $\mathbf{a}_t = \arg\max_\mathbf{a}\left(\mu_{z^*_{\mathbf{s}_t,\mathbf{a}}} + \epsilon_{\mathbf{s}_t,\mathbf{a}}\,\tilde{\sigma}_{z^*_{\mathbf{s}_t,\mathbf{a}}}\right)$
    Observe $\mathbf{s}_{t+1}, r_t$ and store $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in $\mathcal{D}$
end for
```

B Additional environment details

B.1 DeepSea

Our DeepSea MDP (fig. 2) is a variant of the ones used in Osband et al. (2017); O'Donoghue (2018a). 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[6].

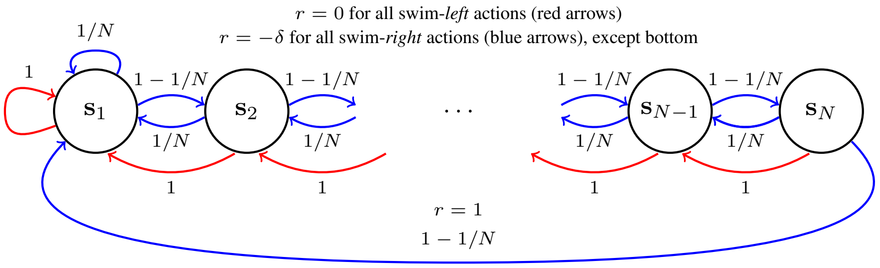


Figure 2: DeepSea MDP from the continuing setting, modified from O'Donoghue (2018a). Blue arrows correspond to swim-*right* (optimal) and red arrows to swim-*left* (sub-optimal).

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. 3) 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 $\sigma_h = \sigma_l = 1$.

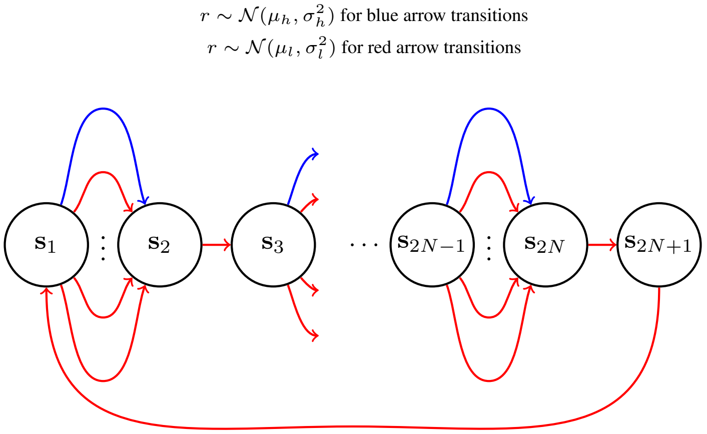

$r \sim \mathcal{N}(\mu_h, \sigma_h^2)$ for blue arrow transitions
$r \sim \mathcal{N}(\mu_l, \sigma_l^2)$ for red arrow transitions
$\mathbf{s}_1$
$\vdots$
$\mathbf{s}_2$
$\mathbf{s}_3$
$\cdots$
$\mathbf{s}_{2N-1}$
$\vdots$
$\mathbf{s}_{2N}$
$\mathbf{s}_{2N+1}$

Figure 3: The WideNarrow MDP. All transitions are deterministic.

[6] 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 $\mathbf{s}_1, \mathbf{a}_1$, 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 $\{\boldsymbol{\eta}_{\mathbf{s},\mathbf{a}}\}$ with:

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

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

$$\mu_{\mathbf{s},\mathbf{a}}, \tau_{\mathbf{s},\mathbf{a}} \sim NG(\mu_{\mathbf{s},\mathbf{a}}, \tau_{\mathbf{s},\mathbf{a}} | \mu, \lambda, \alpha, \beta) \text{ with } (\mu, \lambda, \alpha, \beta) = (0.00, 1.00, 4.00, 4.00).$$

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

C Supplementary figures

C.1 Regret summaries

The following plots summarise the regret of each algorithm to the oracle agent. The regrets have been normalised by the total reward received by the agent, to make the numbers comparable across environments.

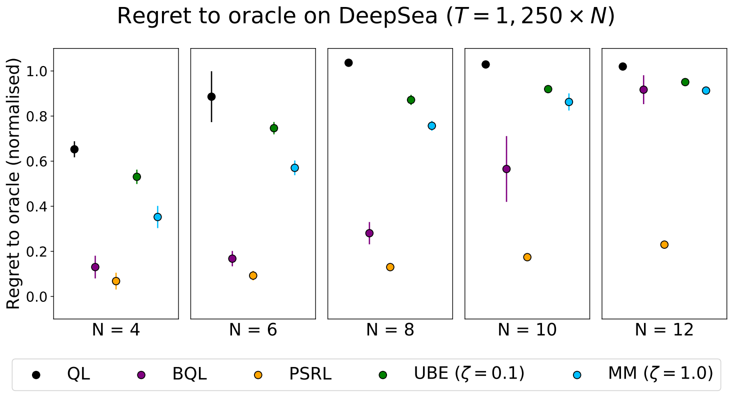


Figure 4: Summary of regret performances to oracle on DeepSea.

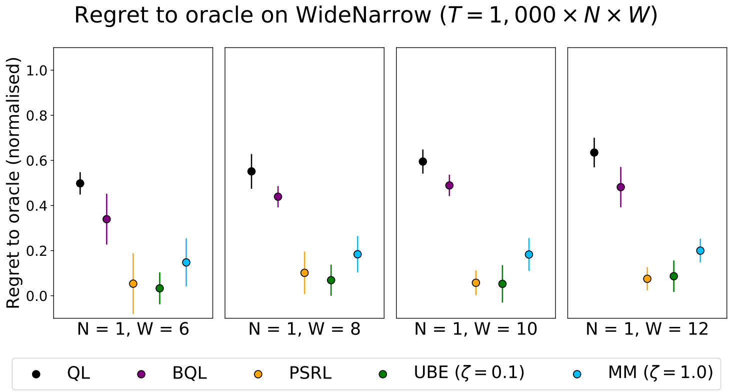


Figure 5: Summary of regret performances to oracle on WideNarrow.

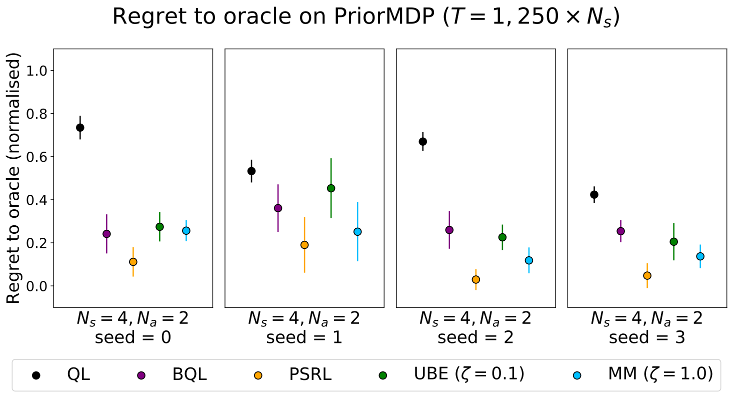


Figure 6: Summary of regret performances to oracle on PriorMDP.

C.2 DeepSea

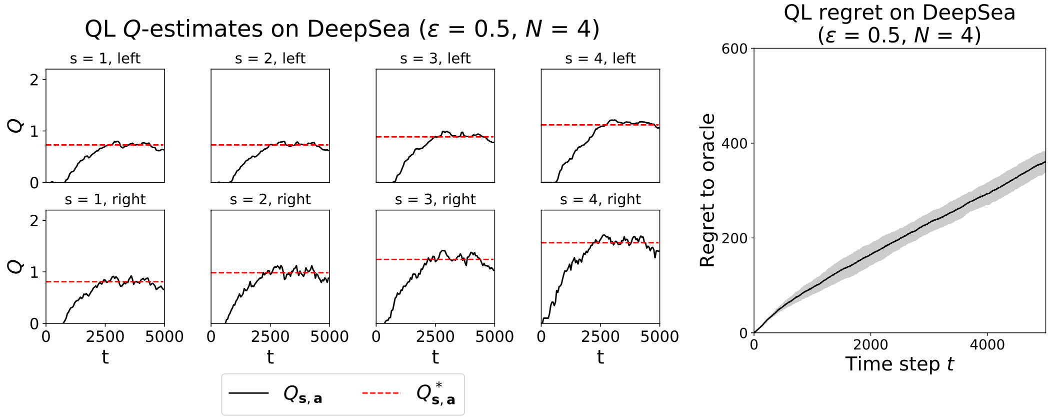


Figure 7: QL Q-estimates and regret on DeepSea.

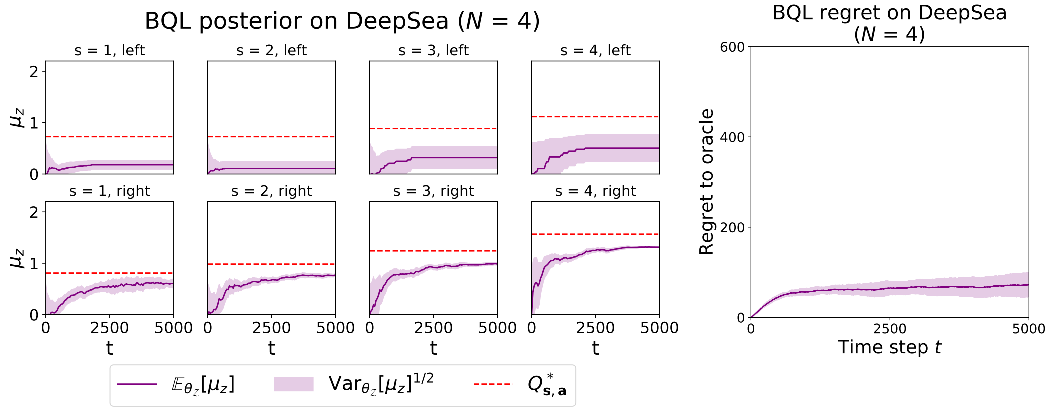


Figure 8: BQL posterior and regret on DeepSea.

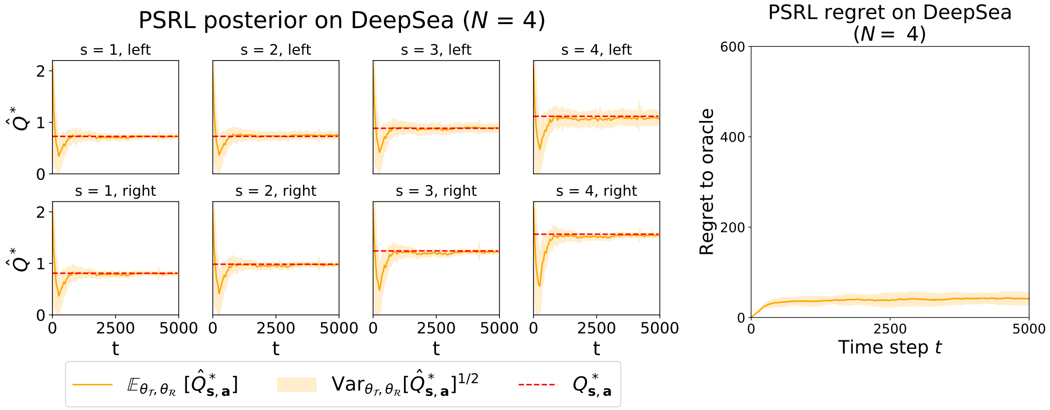


Figure 9: PSRL posterior and regret on DeepSea.

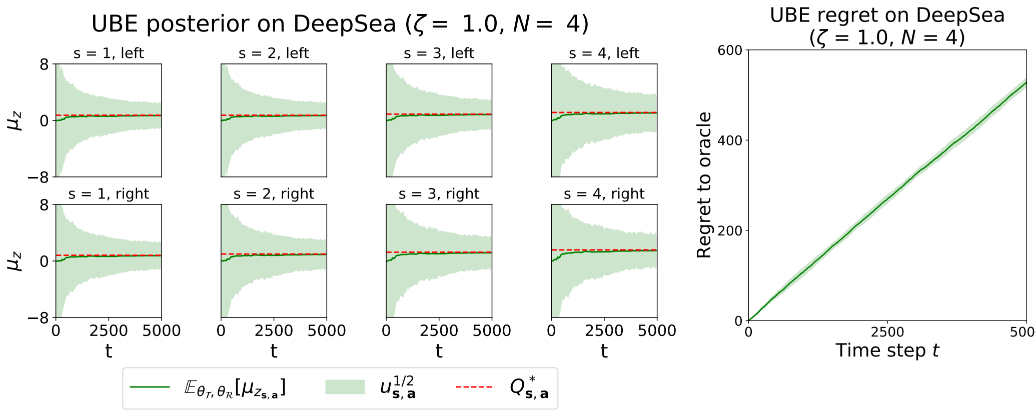


Figure 10: UBE posterior and regret on DeepSea.

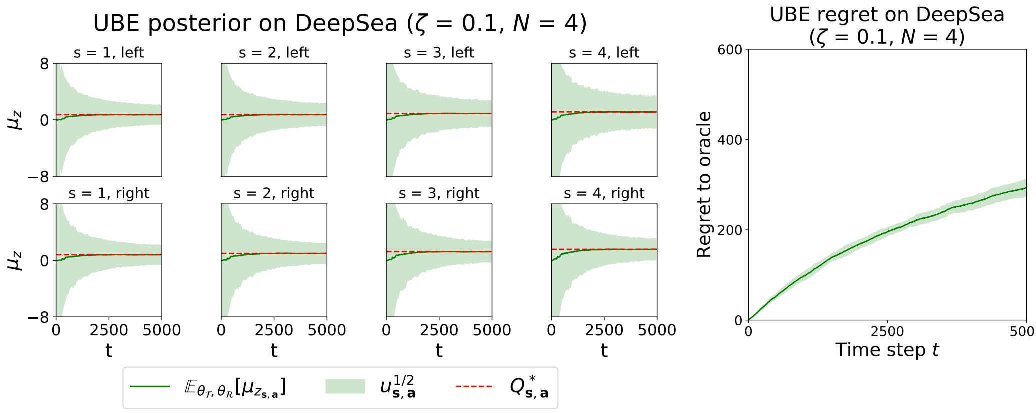


Figure 11: UBE posterior and regret on DeepSea.

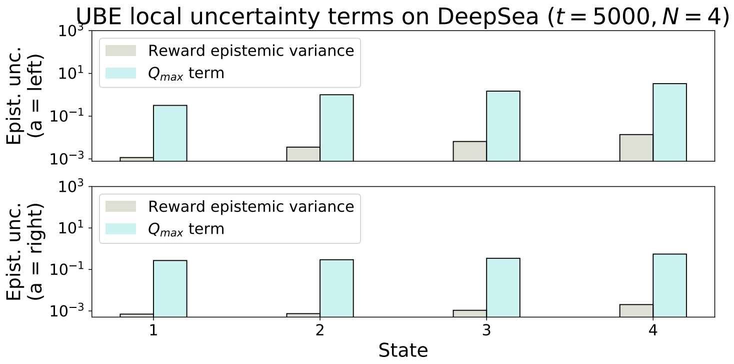


Figure 12: Contributions to the local variance $\nu^*_{\mathbf{s},\mathbf{a}}$ by the reward and the Q_{max} term. This plot corresponds to fig. 11. Note the logarithmic scale.

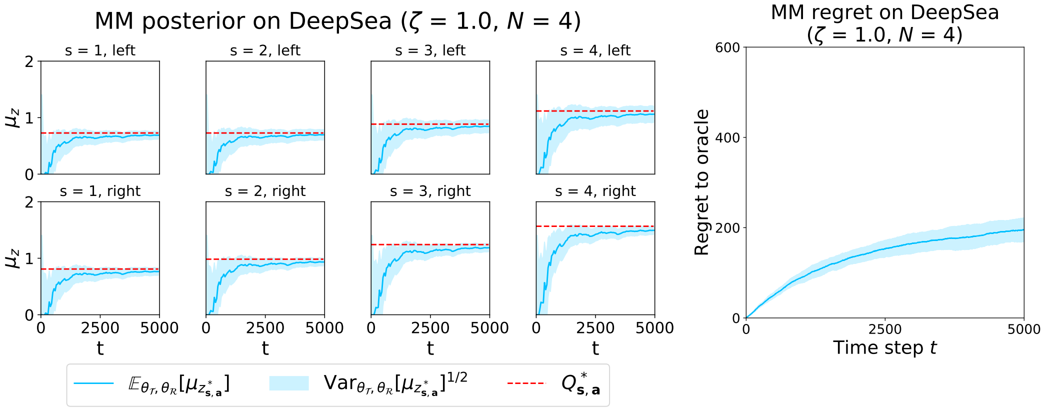


Figure 13: MM posterior and regret on DeepSea.

C.3 WideNarrow

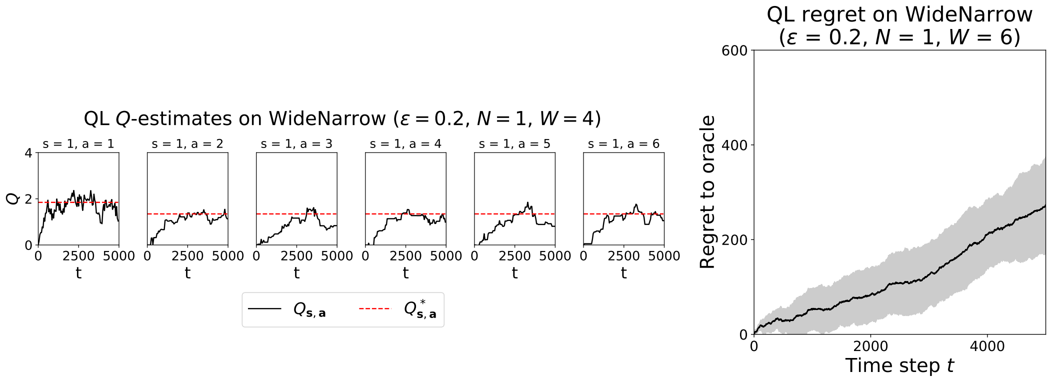


Figure 14: QL Q-estimates and regret on WideNarrow.

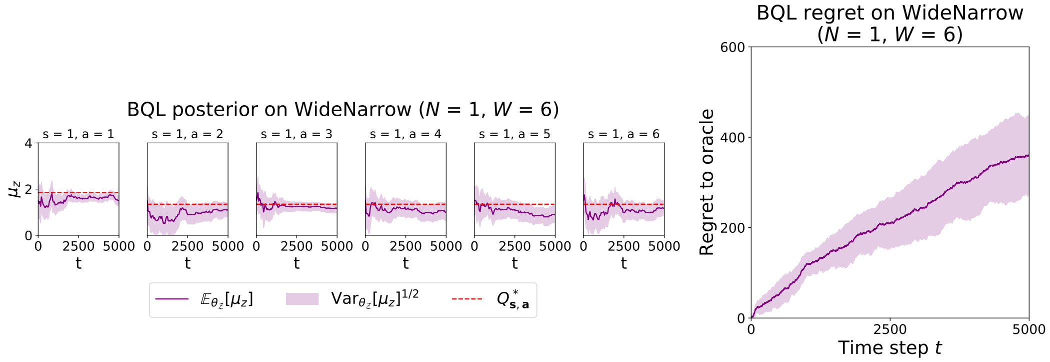


Figure 15: BQL posterior and regret on WideNarrow.

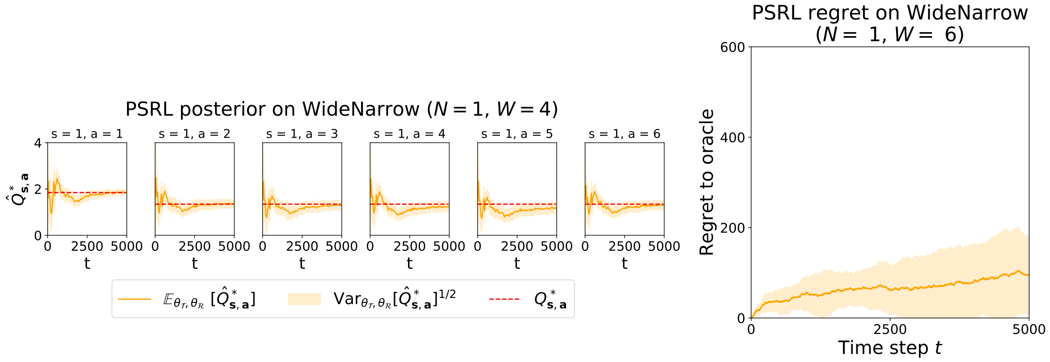


Figure 16: PSRL posterior and regret on WideNarrow.

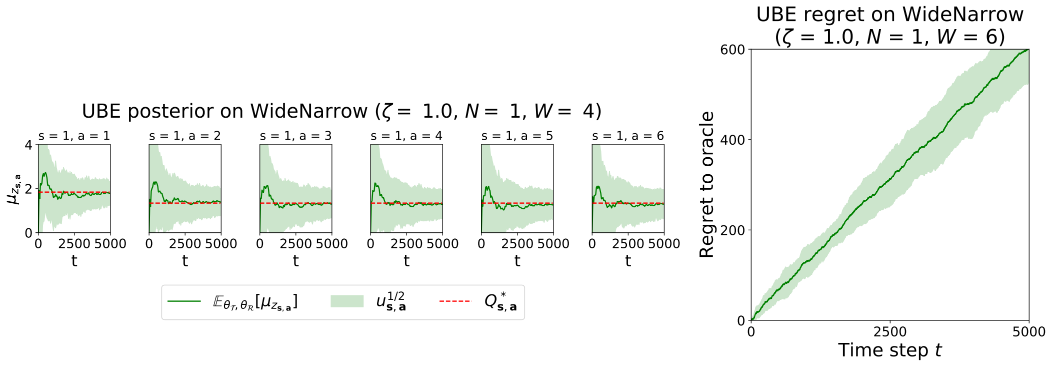


Figure 17: UBE posterior and regret on WideNarrow.

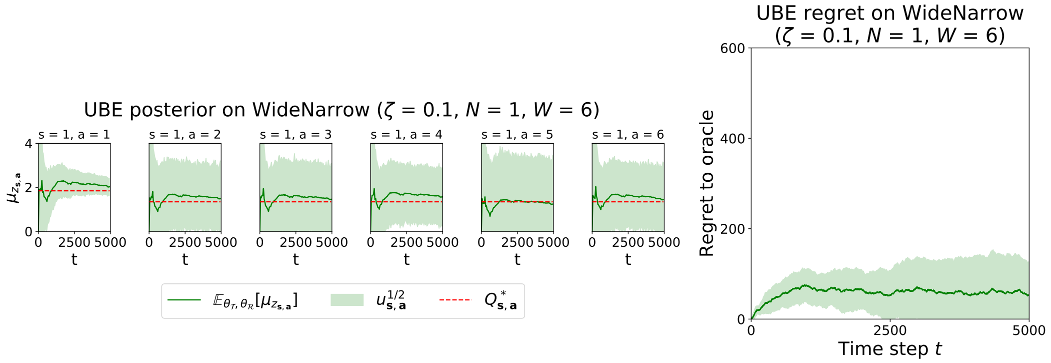


Figure 18: UBE posterior and regret on WideNarrow.

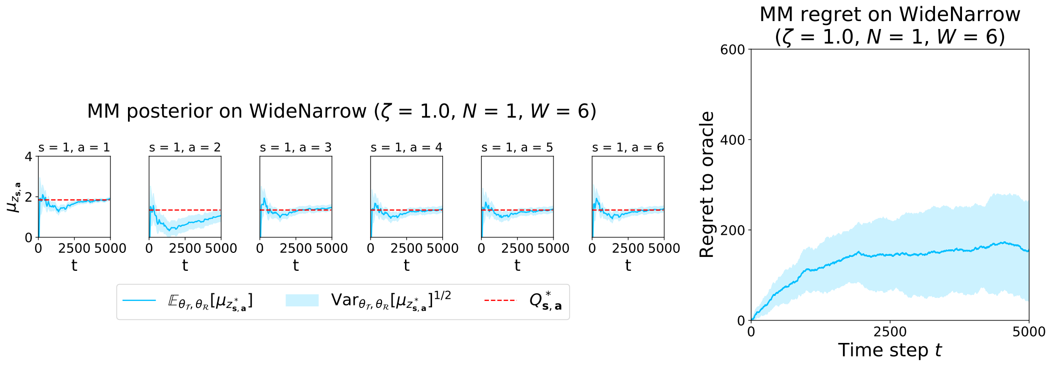
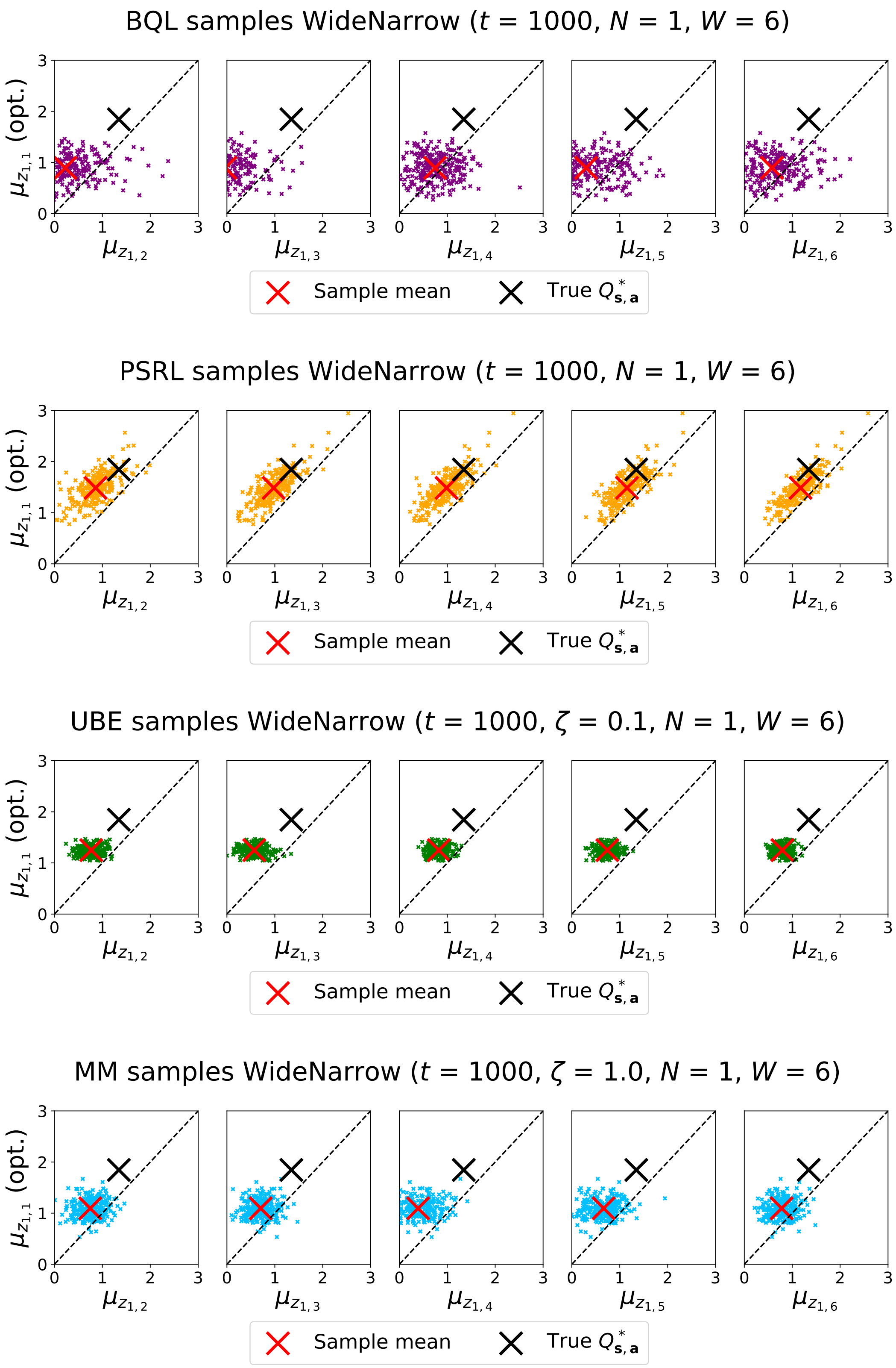


Figure 19: MM posterior and regret on WideNarrow.


BQL samples WideNarrow (t = 1000, N = 1, W = 6)

PSRL samples WideNarrow (t = 1000, N = 1, W = 6)

UBE samples WideNarrow (t = 1000, ζ = 0.1, N = 1, W = 6)

MM samples WideNarrow (t = 1000, ζ = 1.0, N = 1, W = 6)

Figure 20: Correlation plots for WideNarrow at time step $t = 1,000$.

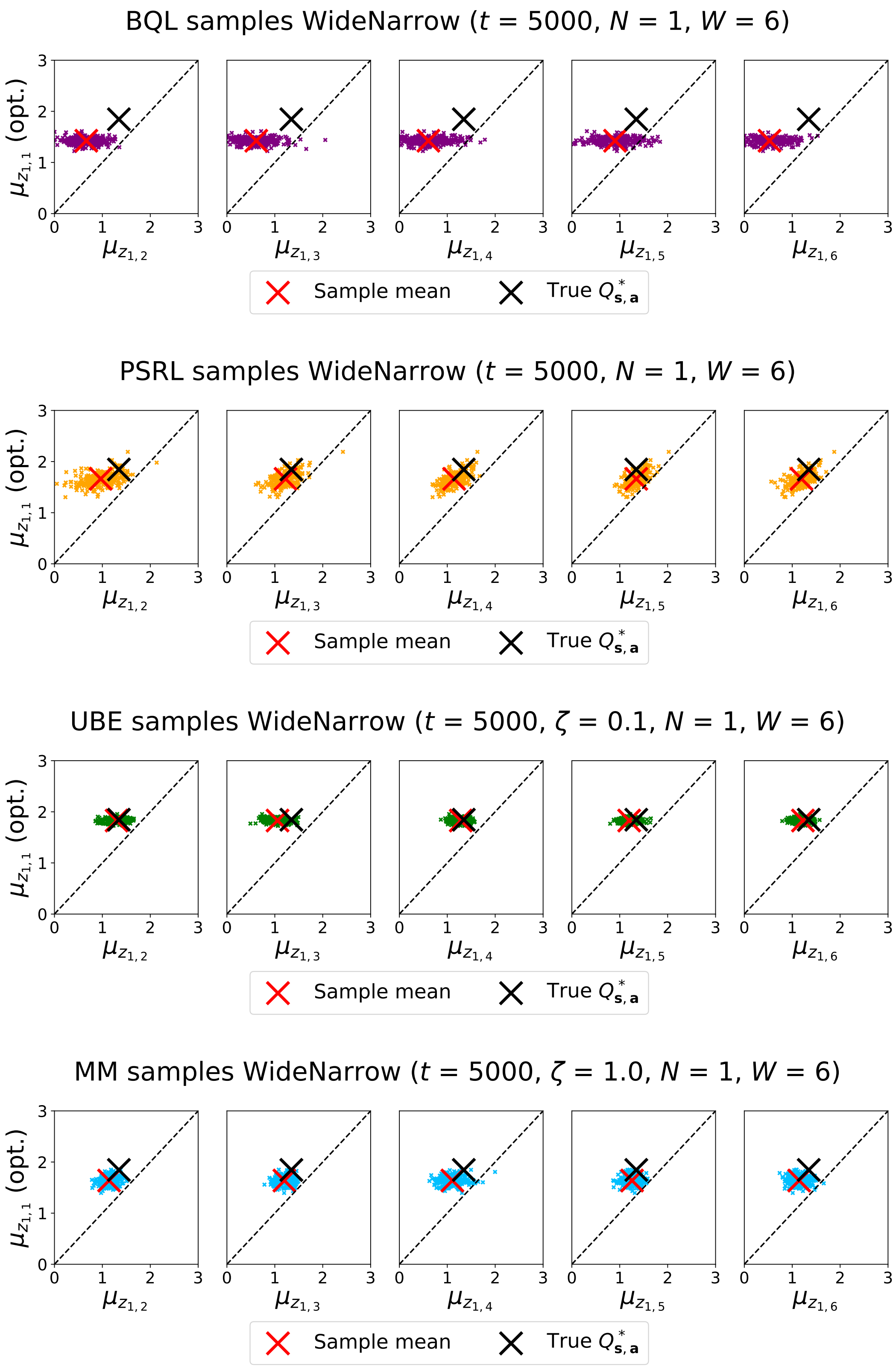


Figure 21: Correlation plots for WideNarrow at time step $t = 5,000$.

C.4 PriorMDP

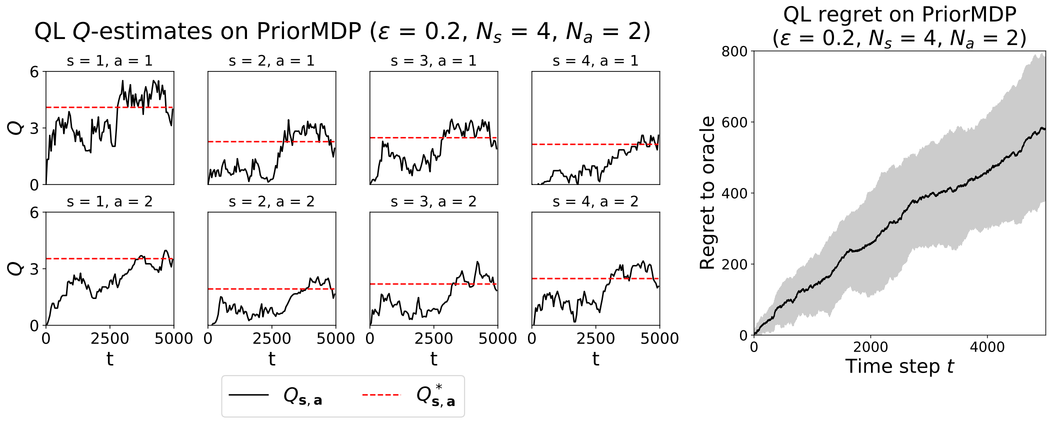


Figure 22: QL Q-estimates and regret on PriorMDP.

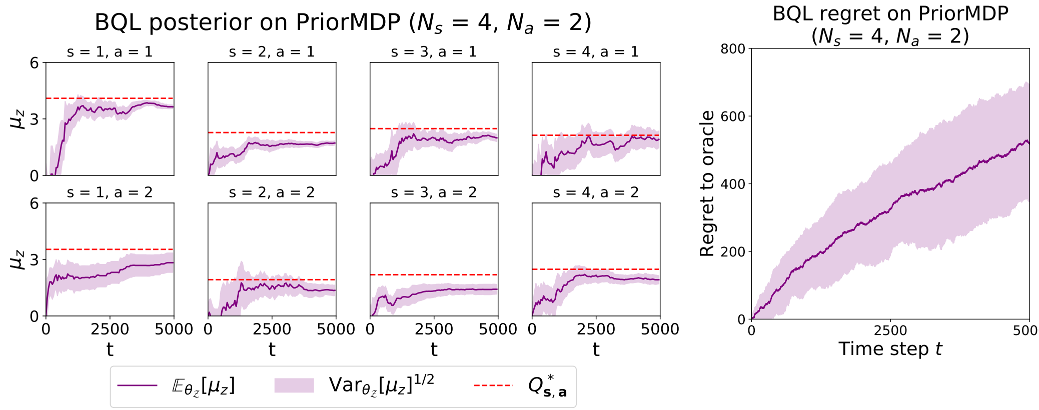


Figure 23: BQL posterior and regret on PriorMDP.

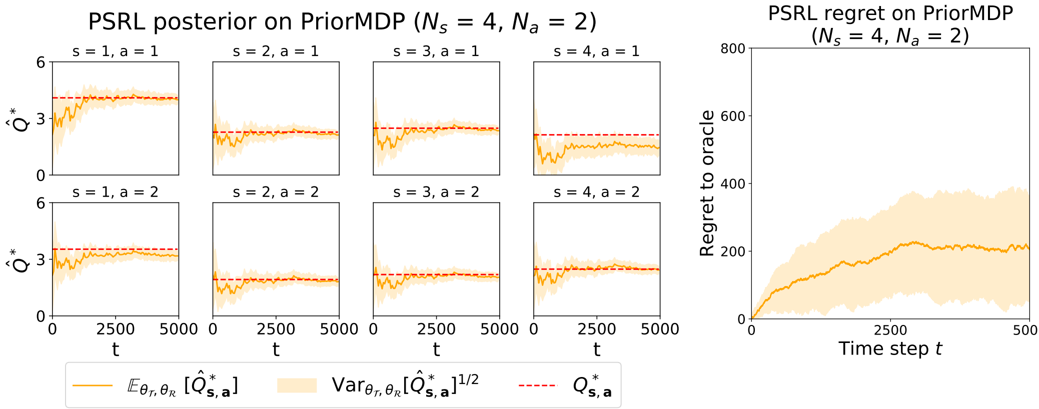


Figure 24: PSRL posterior and regret on PriorMDP.

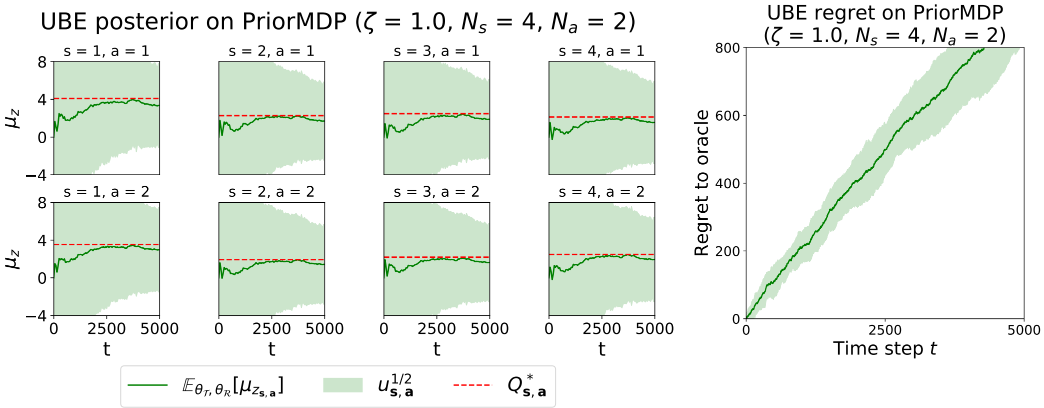


Figure 25: UBE posterior and regret on PriorMDP.

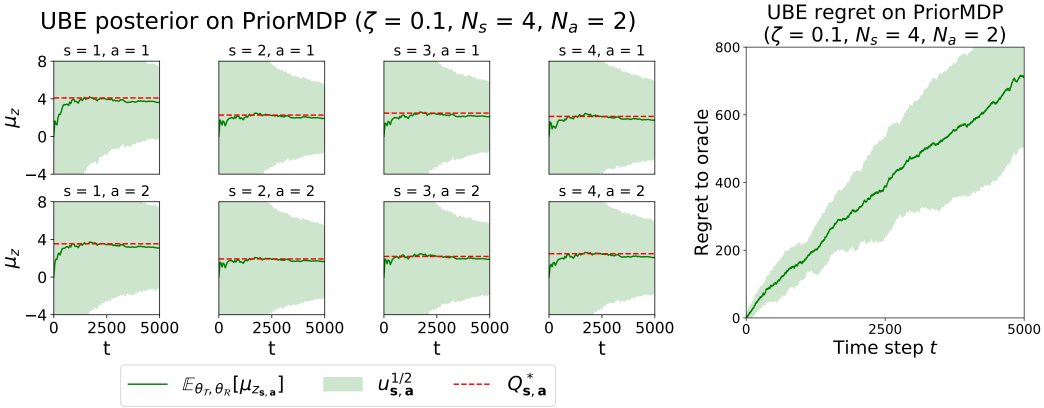


Figure 26: UBE posterior and regret on PriorMDP.

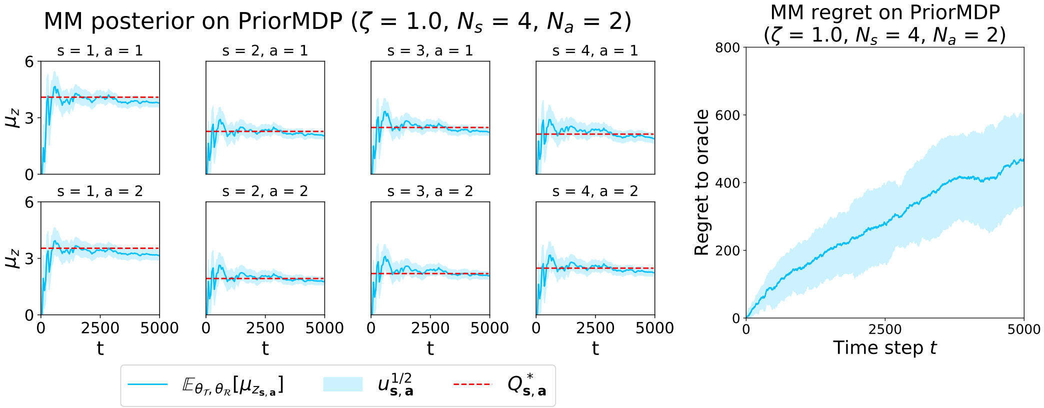


Figure 27: MM posterior and regret on PriorMDP.

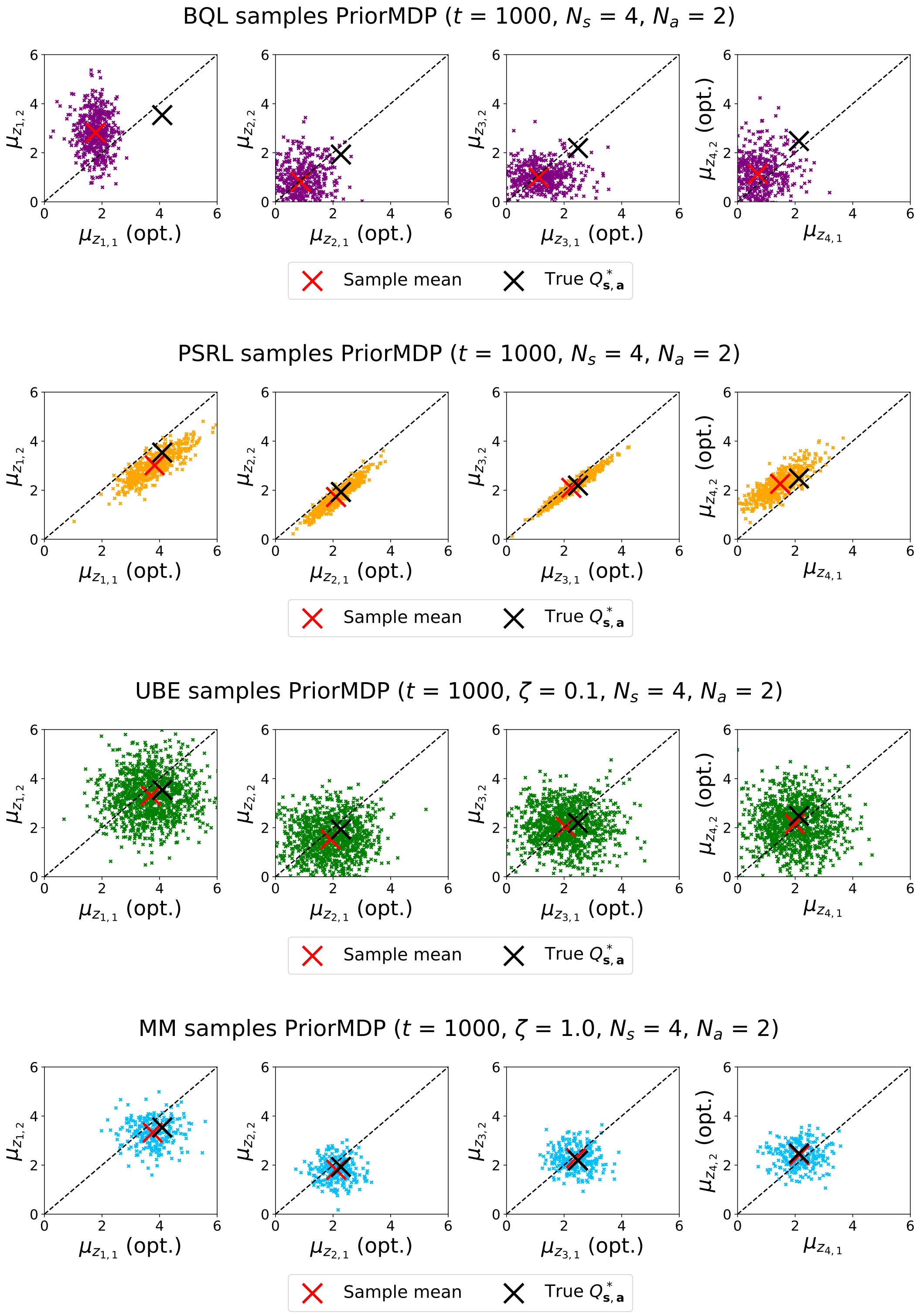


Figure 28: Correlation plots for PriorMDP at time step $t = 1,000$.

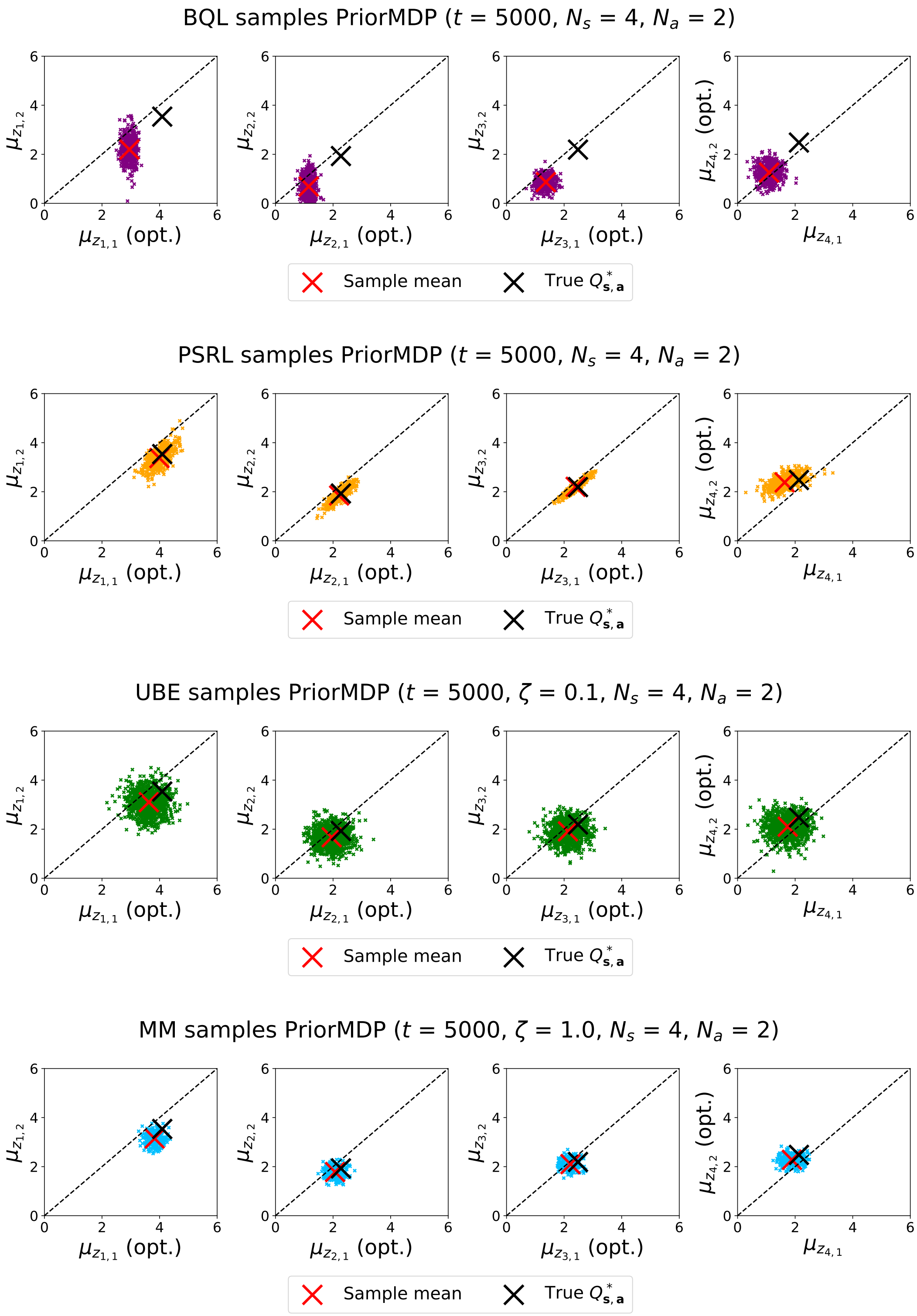


Figure 29: Correlation plots for PriorMDP at time step $t = 5,000$.