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). Bayesian modelling can be incorporated into RL to tackle this tradeoff, by quantifying relevant epistemic uncertainties and using them to guide the exploration. We compare four algorithms based on this approach: Bayesian Q-Learning (BQL), posterior sampling for RL (PSRL), the uncertainty Bellman equation (UBE) and our own moment matching (MM) approach. Our experiments show that: (1) in BQL, early incorrect posterior updates may result in an overconfident and inaccurate posterior; (2) the UBE greatly over-estimates uncertainty and (3) places much larger emphasis on the uncertainty of the dynamics compared to that of the rewards; (4) factored posterior approximations (BQL, UBE, MM) have adverse effects on performance; (5) MM gives better-calibrated uncertainties than the UBE, but still suffers from the factored approximation.

1 Introduction

1.1 Motivation

Balancing exploration and exploitation is one of the central challenges in Reinforcement Learning (RL) (Sutton and Barto, 2018) and mechanisms guaranteeing sufficient exploration during training are central in RL algorithms. However, conventional approaches such as ϵ-greedy or Boltzmann are demonstrably slow to learn (Osband, 2016), because their exploration is *undirected*, driven by injection of random noise in action-selection. Efficient exploration must instead be *directed* by the agent's uncertainty. Bayesian inference can be used to quantify the latter and incorporate it into action-selection. This provides an intuitive and principled *transition mechanism* from exploration to exploitation: the posteriors shrink and the agent converges to the optimal policy as data are observed. In this work we compare a number of Bayesian algorithms in finite Markov Decision Processes (MDPs) in the tabular setting, including our own approach, yielding several interesting insights.

1.2 Notation convention

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. 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. We also define the state- and action-*returns*

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

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.

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

2 Bayesian modelling

2.1 Types of uncertainty: epistemic and aleatoric

One recent and related approach is Distributional RL (DRL) (Bellemare et al., 2017). The authors leverage the fact that the action-return is a random variable and 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 (Watkins and Dayan, 1992) learn Q^*, DRL learns $\mathcal{Z}^*$, the distribution of $z^*_{\mathbf{s},\mathbf{a}}$, whose expectation is $Q^*_{\mathbf{s},\mathbf{a}}$. It is postulated that DRL improves performance because it leverages a richer learning signal and gracefully handles return multi-modalities. DRL models the *aleatoric* or *irreducible* uncertainty due to the inherent stochasticity in $\mathcal{T}$ and $\mathcal{R}$. This leads to more meaningful models of the return, but is not useful for improving exploration.

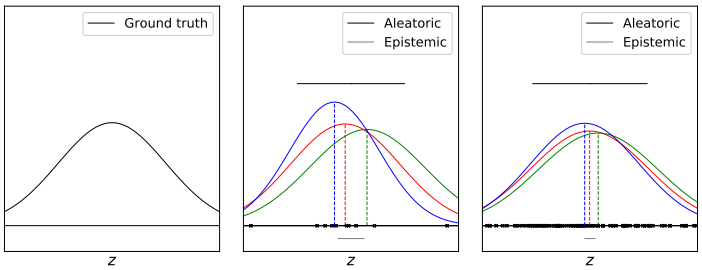

Figure 1: Distinction between aleatoric and epistemic uncertainty for a toy example. As more data are observed (black crosses), the range of plausible Gaussian models narrows down to the ground truth. The epistemic uncertainty in the mean (grey) shrinks, but the aleatoric uncertainty (black) does not reduce bellow the level of inherent noise.

Instead, the quantity relevant to exploration is the uncertainty in the parameterisation of $\mathcal{Z}^*$ due to the finite amount of observed data, which is known as the *epistemic* uncertainty - see fig. 1. In the limit of zero epistemic uncertainty, the agent is certain about the parameters of its models of $\mathcal{T}$ and $\mathcal{R}$, so acting greedily is optimal given these models - even though there is still aleatoric uncertainty. If the epistemic uncertainty is non-zero, the agent will be uncertain about the *expected returns* and should account for this while exploring. A principled approach would therefore be to quantify the epistemic uncertainty and use it to direct the exploration, e.g. by Thompson sampling (Thompson, 1933). This also provides an automatic transition into exploitation, as the posteriors become narrower.

2.2 Bayes' rule and conjugate priors

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 by $\boldsymbol{\theta}$ and then, given data $\mathcal{D} = \{\mathbf{s}, \mathbf{a}, \mathbf{s}', r\}$ and a prior $p(\boldsymbol{\theta})$, we compute the posterior belief $p(\boldsymbol{\theta}|\mathcal{D})$ through Bayes' rule:

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

Choosing a *conjugate* prior simplifies downstream calculations: for discrete distributions such as $\mathcal{T}$, we use a Categorical-Dirichlet model (Bishop, 2006) for each $\mathbf{s}, \mathbf{a}$, while for continuous distributions such as $\mathcal{R}, \mathcal{W}, \mathcal{Z}$ we use a Normal-Gamma (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 algorithm for the tabular setting. The agent models $\mathcal{Z}^*$, the distribution over returns under π^*, and updates $p(\boldsymbol{\theta}_{\mathcal{Z}^*}|\mathcal{D})$ using a heuristic mixture update rule, as new data arrive. The authors make the additional 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 exact posterior by a factored distribution. In reality, the expected returns are related though the BE, so the exact posterior is not factored. 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. 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 $\mu_{z^\pi_{\mathbf{s},\mathbf{a}}} = \mathbb{E}_z[z^\pi_{\mathbf{s},\mathbf{a}}|\boldsymbol{\theta}_{\mathcal{Z}}]$. The authors assume that: (1) the MDP is a directed acyclic graph (DAG) and the task is episodic with time variable $t = 1, ..., T$; (2) the mean rewards of the MDP are bounded within $[-R_{max}, R_{max}]$. Taking variances across the BE and defining an appropriate Bellman operator, they show that the corresponding UBE, 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. Solving for the greedy policy $\pi^*|\boldsymbol{\theta}_{\mathcal{T}}, \boldsymbol{\theta}_{\mathcal{R}}$ and then for $u^*_{\mathbf{s},\mathbf{a}}$, Thompson sampling can be performed from a diagonal Gaussian - leading to a factored posterior approximation. The Thompson noise variance is scaled by an appropriate scaling factor, ζ. 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 computation of an upper bound. 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 (Weiss et al., 2006), the second-order moments can be decomposed into purely aleatoric and purely epistemic terms which, we argue, 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 - also yielding a factored approximation. Further details are given in appendix A.4.

4 Environments and methods

We compare the algorithms on finite MDPs of various sizes in the continuing setting, measuring performance by the cumulative regret to an oracle-agent following the optimal policy[2] - see appendix B for details. Our DeepSea MDP, a variant of that in O'Donoghue (2018a), tests the algorithm's ability for sustained exploration. We propose the WideNarrow environment for investigating the effect of factored posterior approximations. Finally, we compare the algorithms on random MDPs, with dynamics and rewards drawn from Categorical and NG priors as in Osband et al. (2013) - we refer to this as PriorMDP.

5 Results and discussion

We observe several trends in our results. Figure 2 shows the regret, normalised by the oracle's cumulative reward to make the performances comparable on the same scale - see appendix C for further results and supporting figures. Figure 3 shows the evolution of the posterior of each algorithm

[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.

[2] Our implementations and plotting code can be found at `https://github.com/sample-efficient-bayesian-rl`.

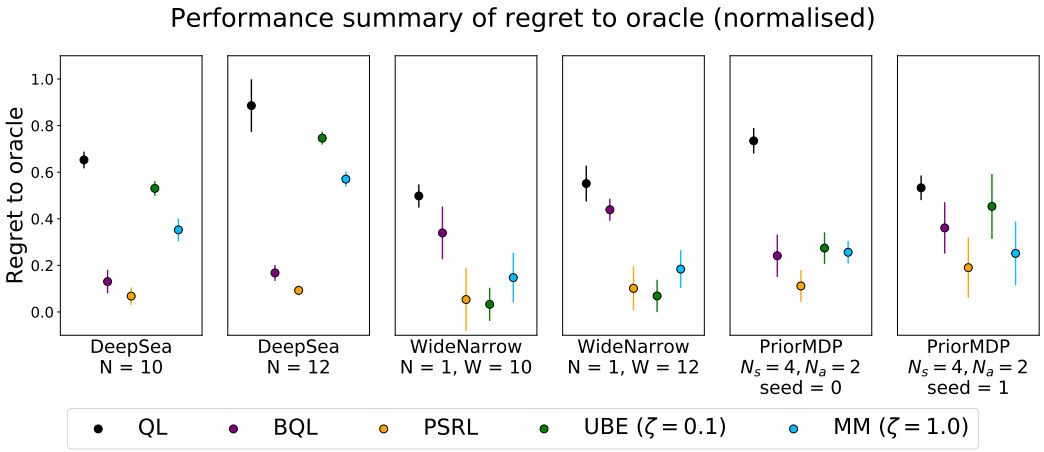

Figure 2: Summary performances in terms of regret to the oracle on selected environments. N, W, N_s and N_a are environment size parameters - see section B for specifications.

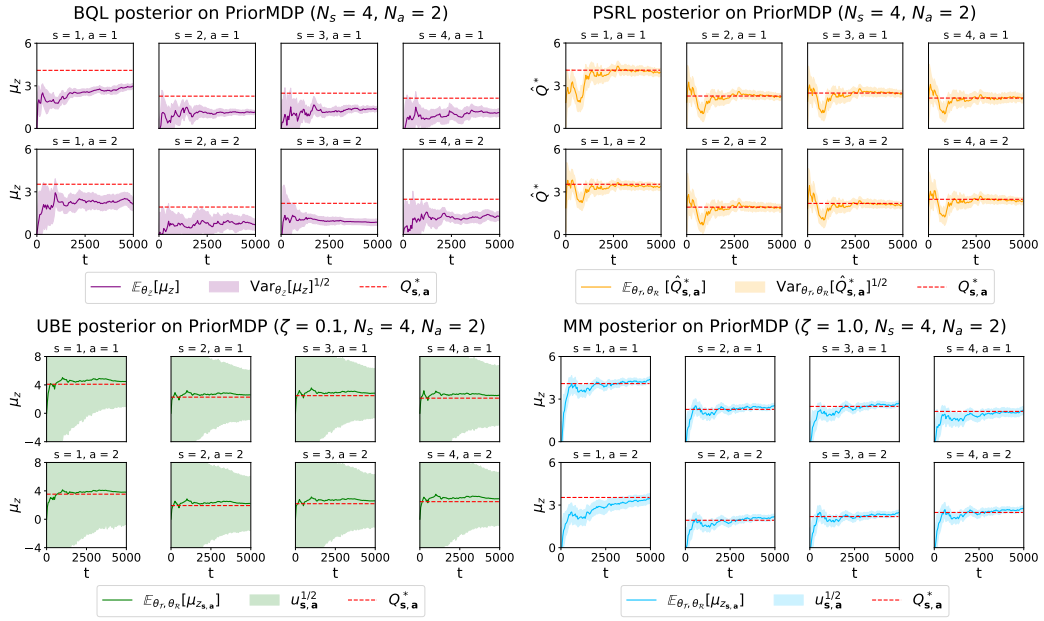

Figure 3: Plots of the evolution of the posterior on the same PriorMDP for each algorithm.

on PriorMDP during learning. Generally, 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. Often, the agent does not over-explore actions if it is confident that these are suboptimal. For example, this is seen for BQL, $(\mathbf{s} = 1, \mathbf{a} = 1)$, in fig. 3 - see also fig. 7 and fig. 8. Although there is significant uncertainty in the $\mu_{z^*_{\mathbf{s},\mathbf{a}}}$ of the suboptimal action, the agent is confident that the optimal action $(\mathbf{a} = 1)$ is better. It therefore does not waste time determining the exact $\mu_{z^*_{\mathbf{s},\mathbf{a}}}$ of an action if it is confident that it is suboptimal. The aforementioned behaviours are central for achieving principled and efficient exploration, however we often observe exceptions where some of the algorithms do not behave 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. 3 and also fig. 9, fig. 16 and fig. 25). 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, inspecting contributions to $u^*_{\mathbf{s},\mathbf{a}}$ (fig. 11), reveals that an upper-bounding contribution from the dynamics uncertainty completely outweighs the rewards uncertainty, which is effectively neglected. Scaling the Thompson noise by $\zeta < 1.0$, improves regret performance in some cases. However, one is further faced by the challenge of tuning ζ, which may be expensive and challenging for large problems. By contrast, MM produces better-calibrated uncertainty estimates than the UBE (fig. 3 and also fig. 12, fig. 18 and fig. 26). As a result, MM shows typically better regret performance than the UBE without a need to tune ζ. This could give an advantage to MM in settings where tuning ζ may be expensive or difficult.

Second, we observe that the BQL posterior sometimes fails to concentrate on the true Q^* values (fig. 3 and also fig. 7 and fig. 22), where the posterior is overconfident about incorrect predictions of $\mu_{z^*_{\mathbf{s},\mathbf{a}}}$. This effect persists for different random seeds and but 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 final posterior concentrating on 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 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. 19, fig. 20, fig. 27 and fig. 28) and as a result, the Thompson-sampled $\mu_{z^*_{\mathbf{s},\mathbf{a}}}$ often correspond to picking a sub-optimal action. 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, at no additional computational cost. Accounting for these correlations is an important factor in ensuring the transition from exploration to exploitation is sufficiently quick.

6 Conclusions & Further work

Our comparison of BQL, PSRL, UBE and MM has yielded a number of insights: (1) BQL suffers from a pathology whereby incorrect posterior updates result in an overconfident posterior in the absence of a forgetting mechanism; (2) the UBE uncertainty estimate $u^*_{\mathbf{s},\mathbf{a}}$ is extremely loose, results in essentially undirected exploration if ζ is not tuned and (3) places a much larger emphasis on the dynamics than the rewards uncertainties; (4) factored posterior approximations in BQL, UBE and MM have adverse effects on regret performance, while PSRL avoids this since it incorporates correlations by sampling from the exact posterior; (5) MM gives generally well-calibrated uncertainty estimates, however it still suffers from the factored 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.

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.

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[3] 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 $\mathbf{s}, \mathbf{a} \to \mathbf{s}', r$. 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)

1: 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 $(\mathbf{s}, \mathbf{a})$
2: Observe initial state $\mathbf{s}_1$
3: **for** time-step $\in \{0, 1, ..., T_{\max} - 1\}$ **do**
4: 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
5: $\boldsymbol{\theta}_{\mathcal{Z}^*} \leftarrow$ Updated params. using Monte Carlo on eq. (5)
6: **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.

[3] 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_{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:

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

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 $\mu_{z^\pi_{\mathbf{s},\mathbf{a},t}}$ 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:

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

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

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

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_{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 π^* 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, $\nu_{\mathbf{s},\mathbf{a},t}^\pi \to 0$ 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 $z_{\mathbf{s},\mathbf{a}}^\pi$:

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

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[4]:

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

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

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:

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

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

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

[4]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:

$$\underbrace{\text{Var}_{\boldsymbol{\theta}_\mathcal{Z}}\left[\mathbb{E}_z\left[z^\pi_{\mathbf{s},\mathbf{a}}|\boldsymbol{\theta}_\mathcal{Z}\right]\right]}_{\text{Epistemic action-return unc.}} = \underbrace{\text{Var}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic reward unc. from} \\ \text{dynamics unc.}}} \tag{13}$$

$$+ \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\text{Var}_{\boldsymbol{\theta}_\mathcal{R}}\left[\mathbb{E}_r\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\boldsymbol{\theta}_\mathcal{R}\right]\right]\right]}_{\substack{\text{Epistemic rewards unc. from} \\ \text{rewards unc.}}} +$$

$$+ 2\gamma \underbrace{\text{Cov}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right], \mathbb{E}_{\mathbf{s}',z,\boldsymbol{\theta}_\mathcal{Z},\mathbf{a}'}\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic reward and action-return covariance} \\ \text{from dynamics unc.}}}$$

$$+ \gamma^2 \underbrace{\text{Var}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',z,\boldsymbol{\theta}_\mathcal{Z},\mathbf{a}'}\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic action-return unc. from} \\ \text{dynamics unc.}}}$$

$$+ \gamma^2 \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T},\mathbf{a}'}\left[\text{Var}_{\boldsymbol{\theta}_\mathcal{Z}}\left[\mathbb{E}_z\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\mathbf{s}',\boldsymbol{\theta}_\mathcal{Z}\right]\right]\right]}_{\substack{\text{Epistemic action-return unc. from} \\ \text{state-return unc.}}}$$

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:

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

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_{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 π^* 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. 4) 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[5].

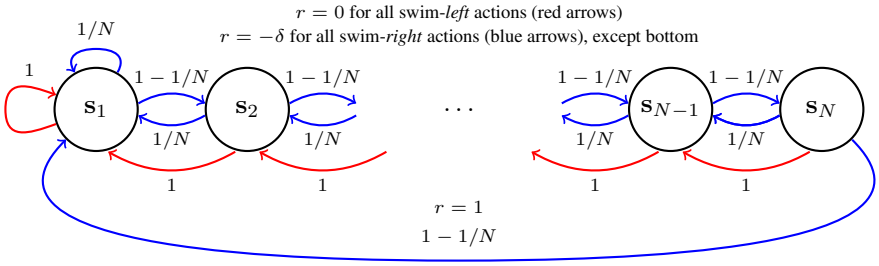

Figure 4: DeepSea MDP from the continuing setting, modified from O'Donoghue (2018a). Blue arrows correspond to swim-*right* (optimal) and red arrows to swim-*left* (sub-optimal).

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

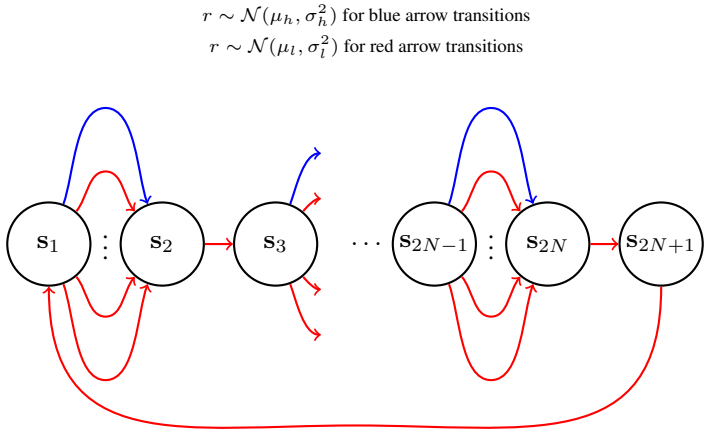

Figure 5: The WideNarrow MDP. All transitions are deterministic.

[5]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 $(\mathbf{s}_1, \mathbf{a}_2)$:

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

where $\boldsymbol{\theta}$ loosely denotes all modelling parameters, $\mathbf{s}'$ 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 $\{\boldsymbol{\eta}_{\mathbf{s},\mathbf{a}}\}$ 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 DeepSea

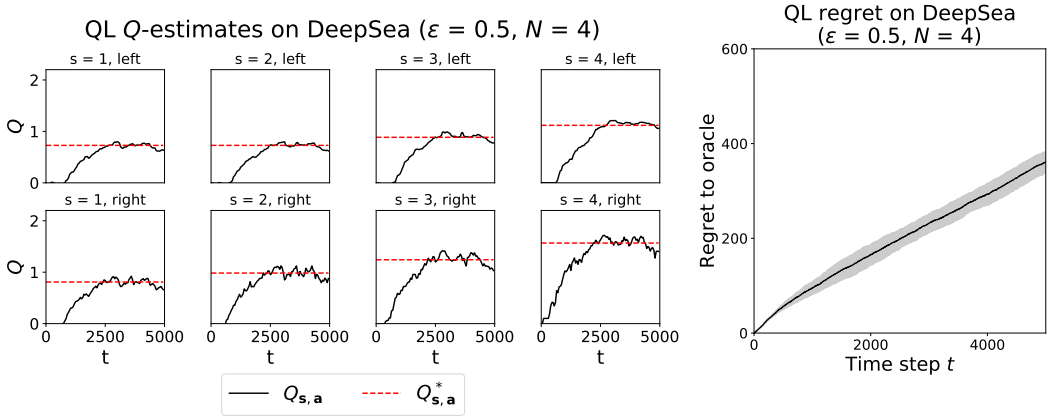

Figure 6: QL Q-estimates and regret on DeepSea.

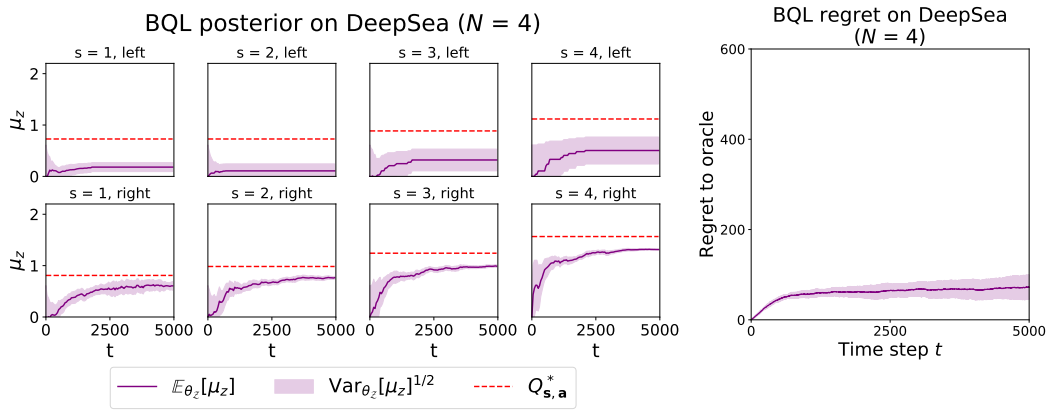

Figure 7: BQL posterior and regret on DeepSea.

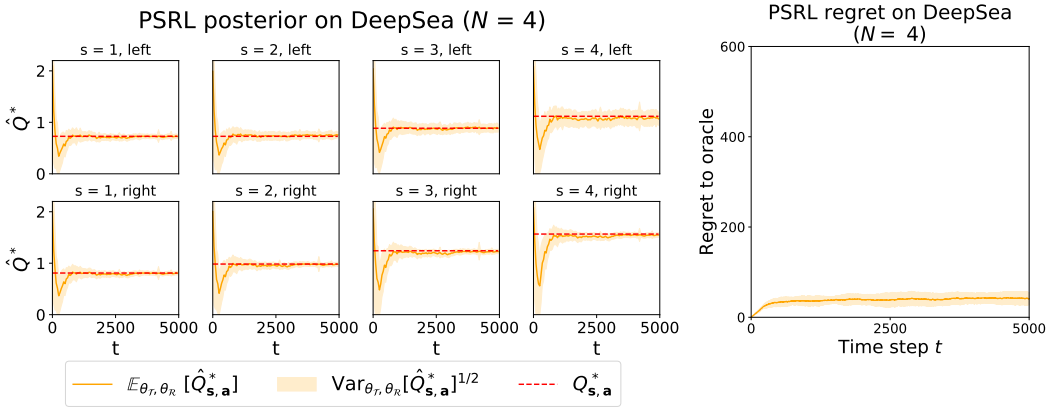

Figure 8: PSRL posterior and regret on DeepSea.

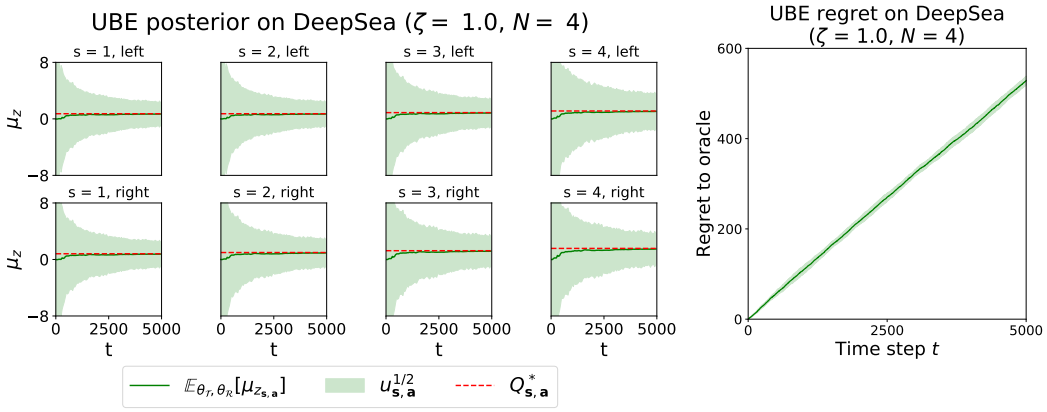

Figure 9: UBE posterior and regret on DeepSea.

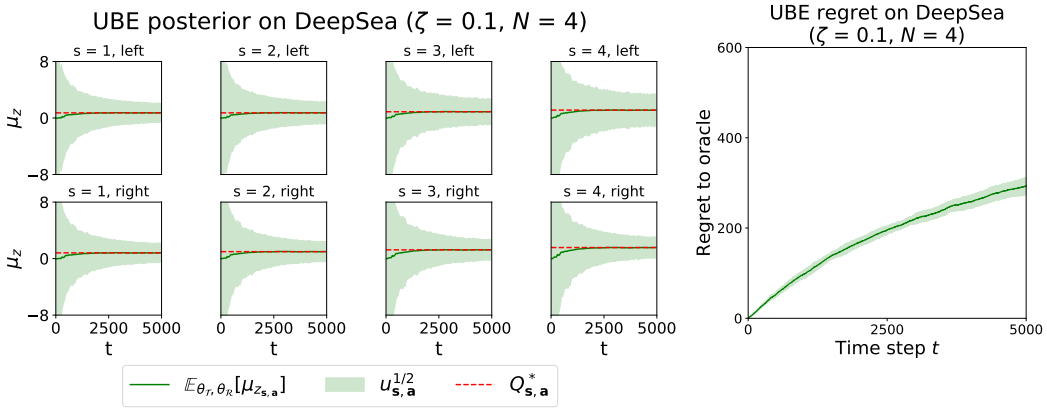

Figure 10: UBE posterior and regret on DeepSea.

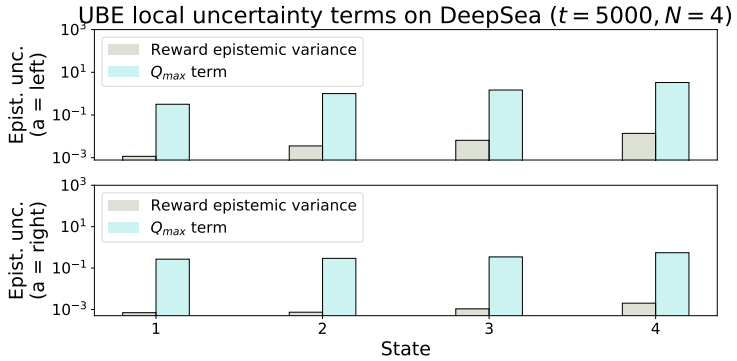

Figure 11: Contributions to the local variance $\nu^*_{\mathbf{s},\mathbf{a}}$ by the reward and the Q_{max} term. This plot corresponds to fig. 10. Note the logarithmic scale.

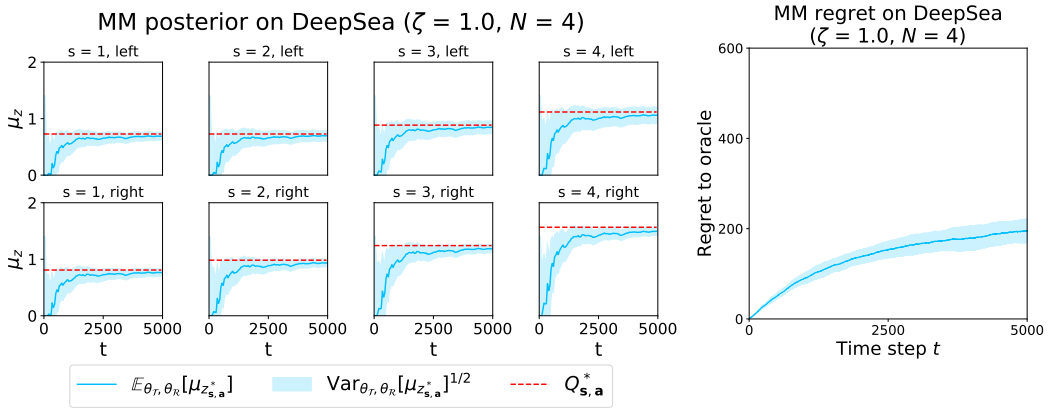

Figure 12: MM posterior and regret on DeepSea.

C.2 WideNarrow

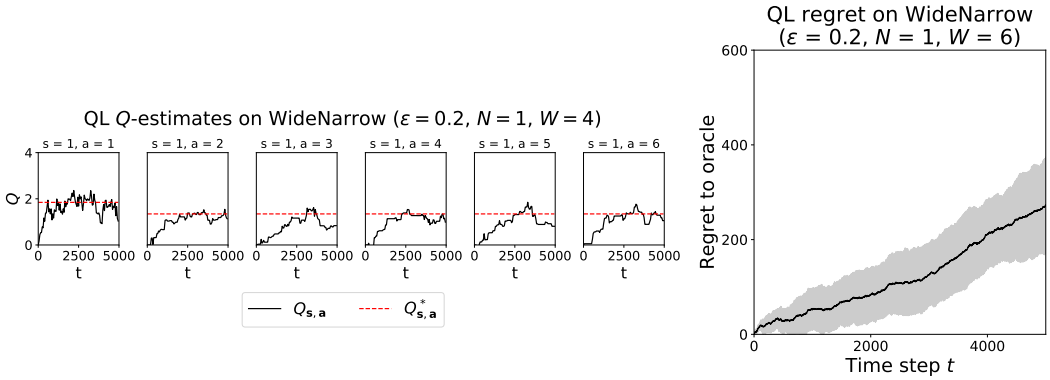

Figure 13: QL Q-estimates and regret on WideNarrow.

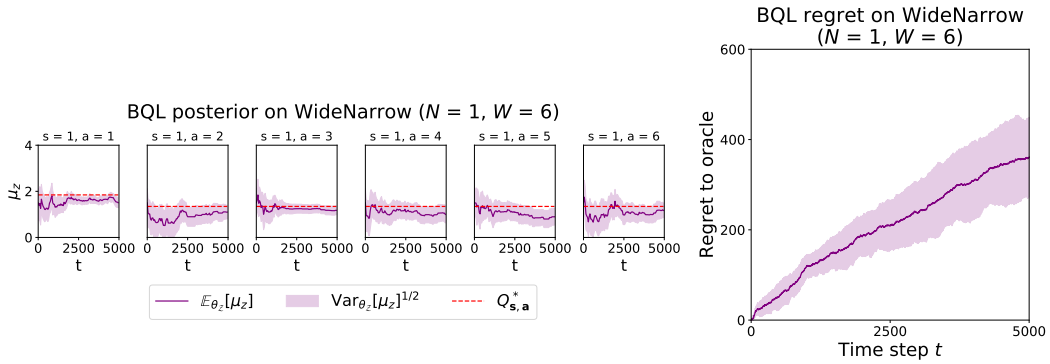

Figure 14: BQL posterior and regret on WideNarrow.

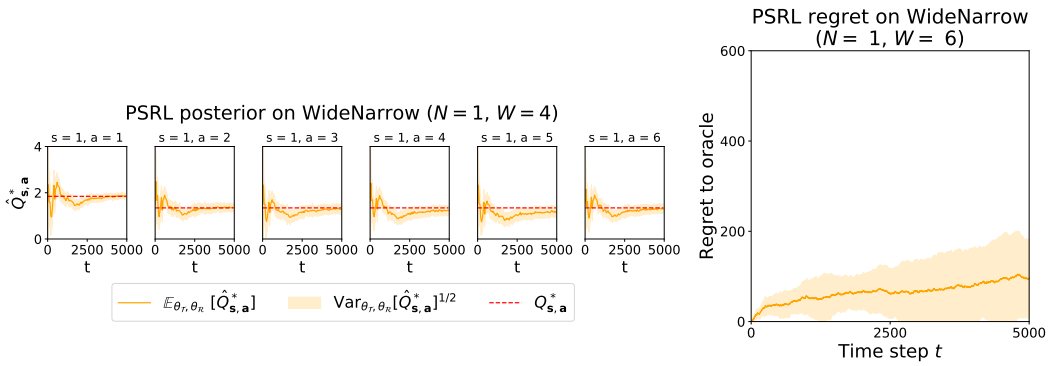

Figure 15: PSRL posterior and regret on WideNarrow.

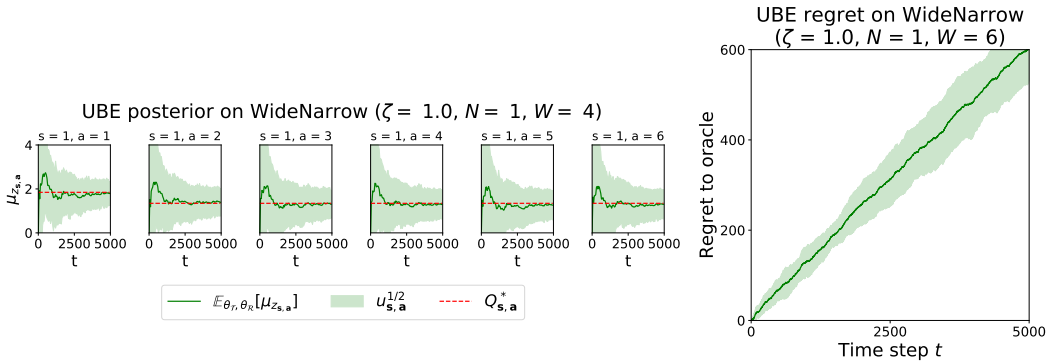

Figure 16: UBE posterior and regret on WideNarrow.

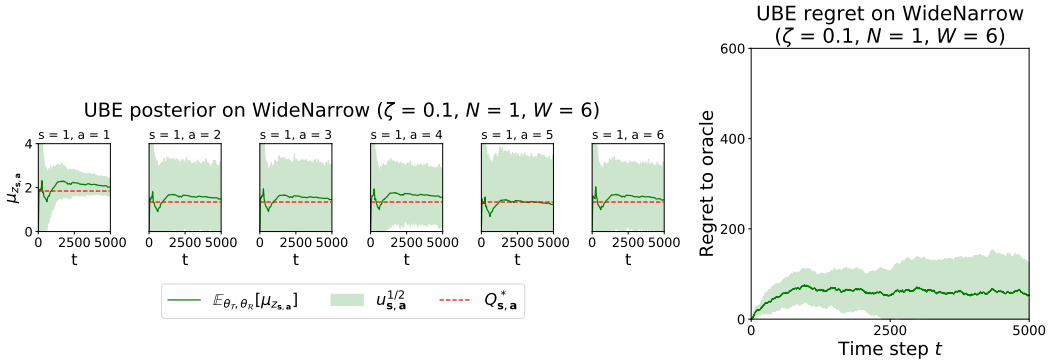

Figure 17: UBE posterior and regret on WideNarrow.

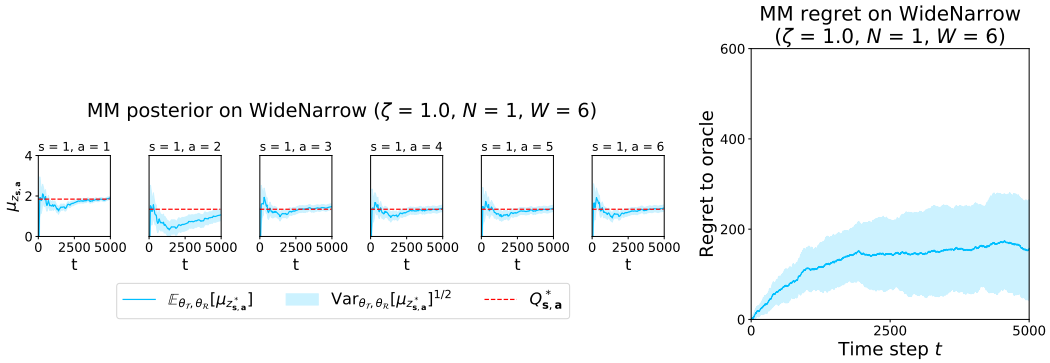

Figure 18: MM posterior and regret on WideNarrow.

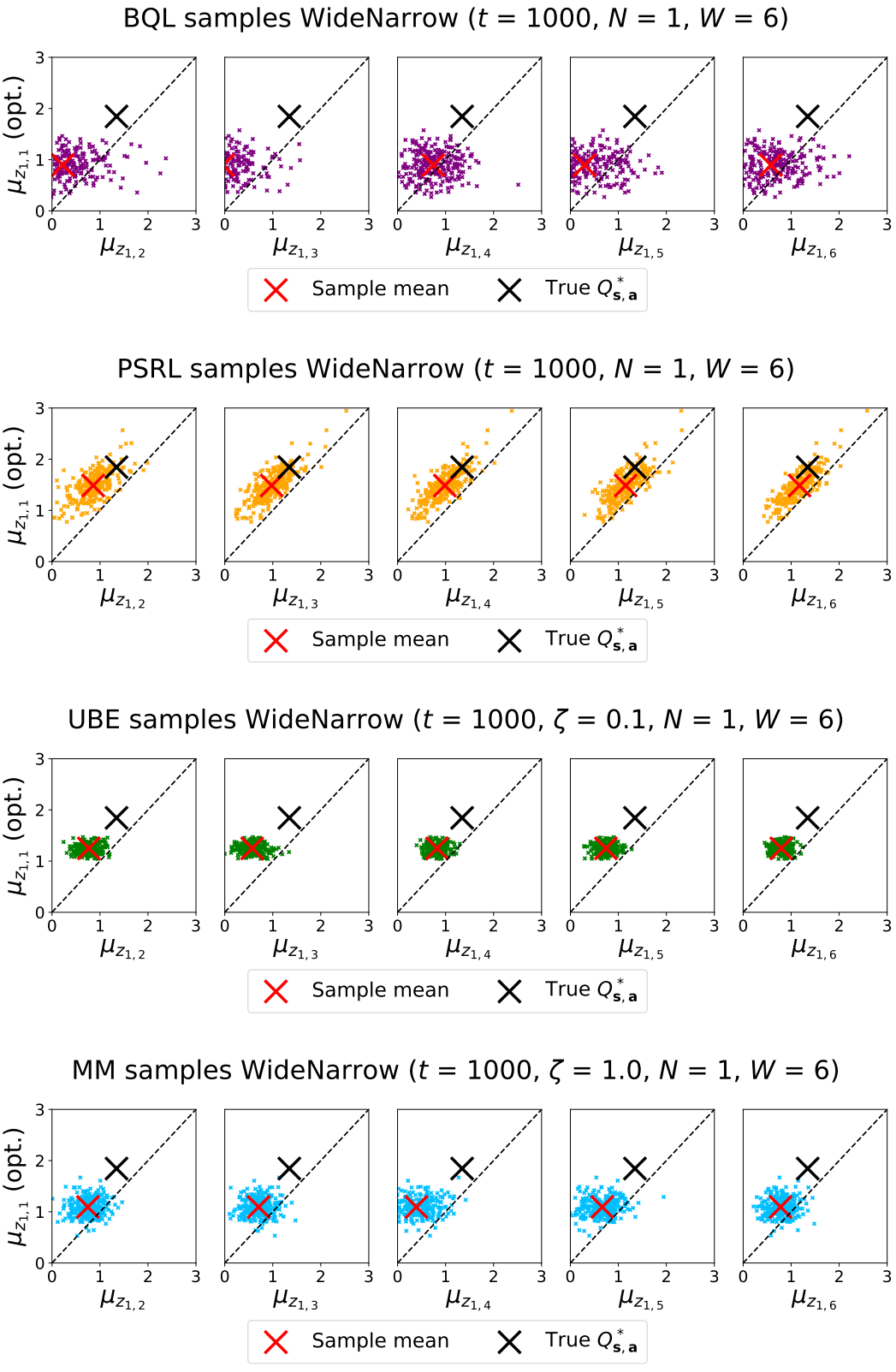

Figure 19: Correlation plots for WideNarrow at time step $t = 1,000$.

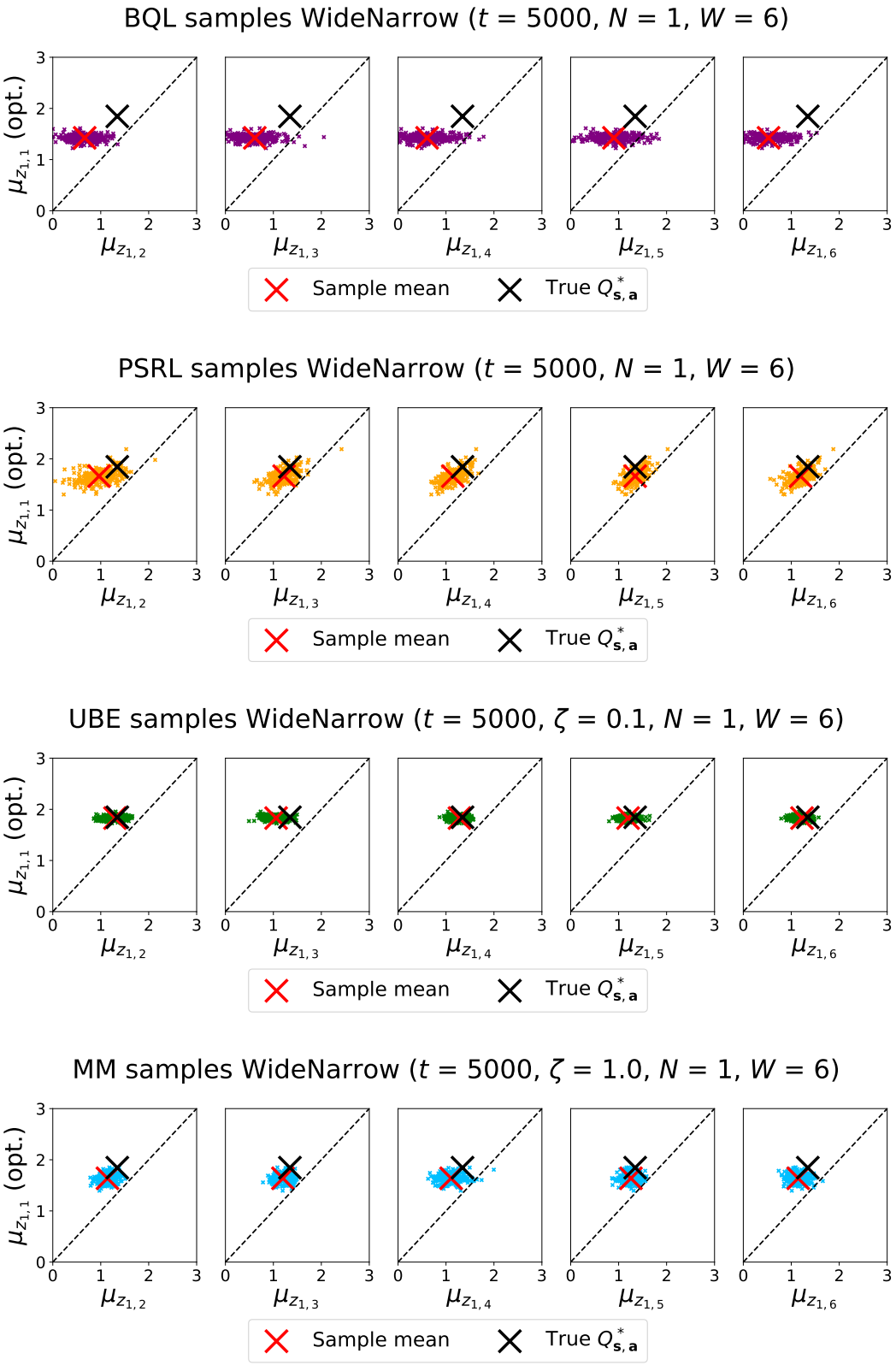

Figure 20: Correlation plots for WideNarrow at time step $t = 5,000$.

C.3 PriorMDP

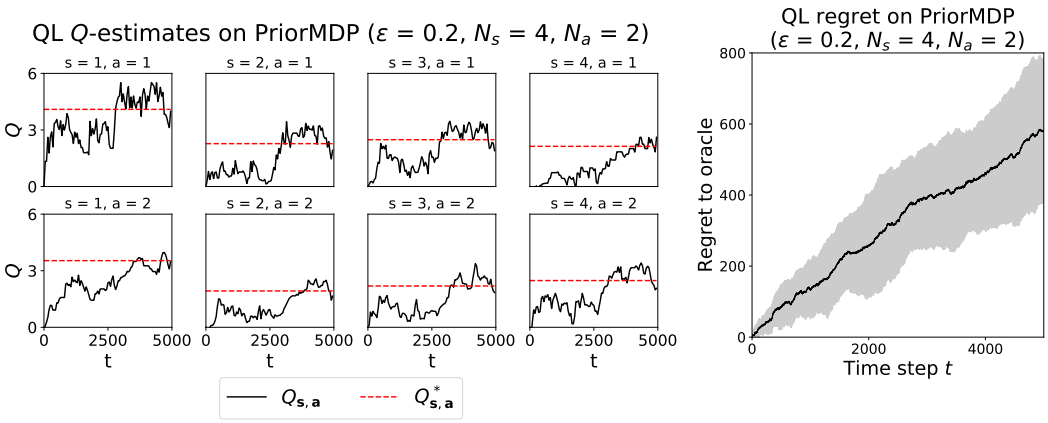

Figure 21: QL Q-estimates and regret on PriorMDP.

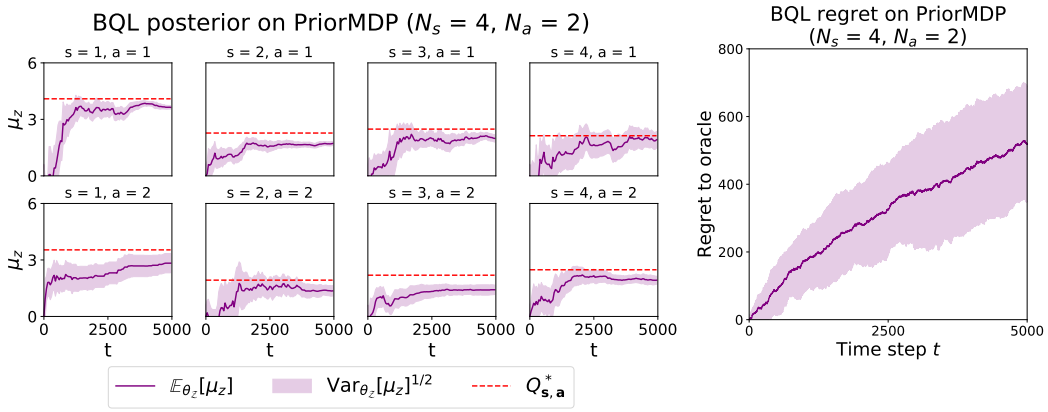

Figure 22: BQL posterior and regret on PriorMDP.

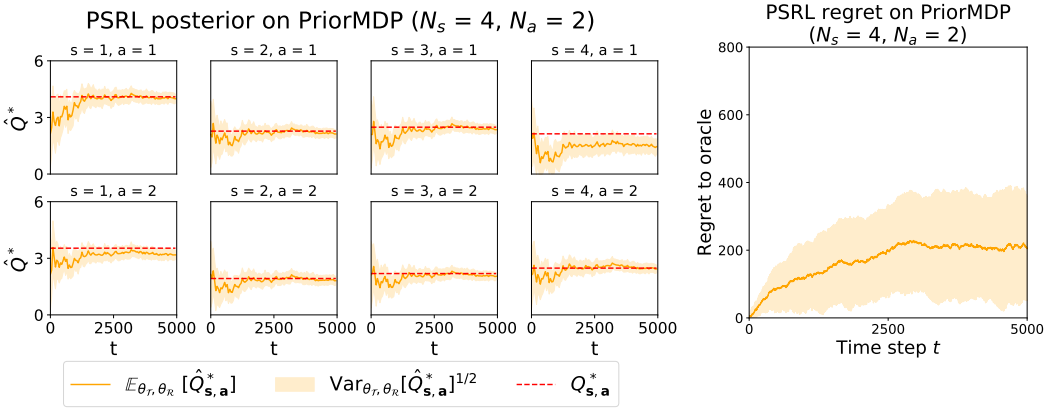

Figure 23: PSRL posterior and regret on PriorMDP.

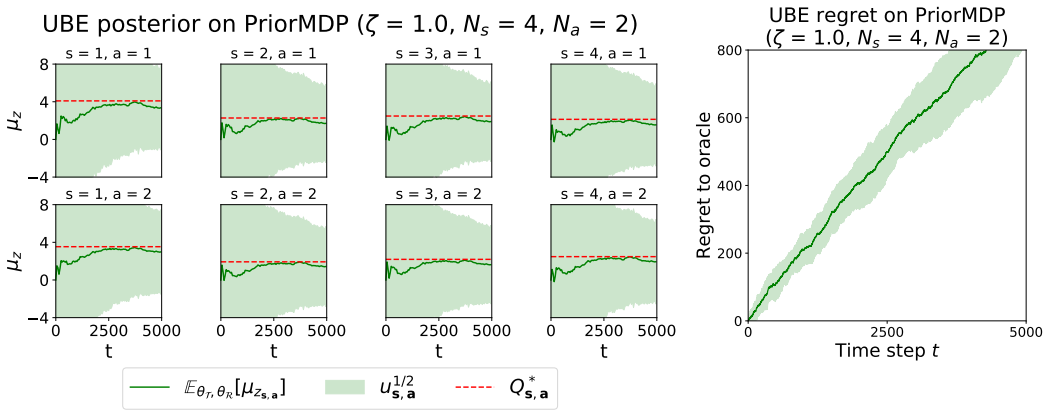

Figure 24: UBE posterior and regret on PriorMDP.

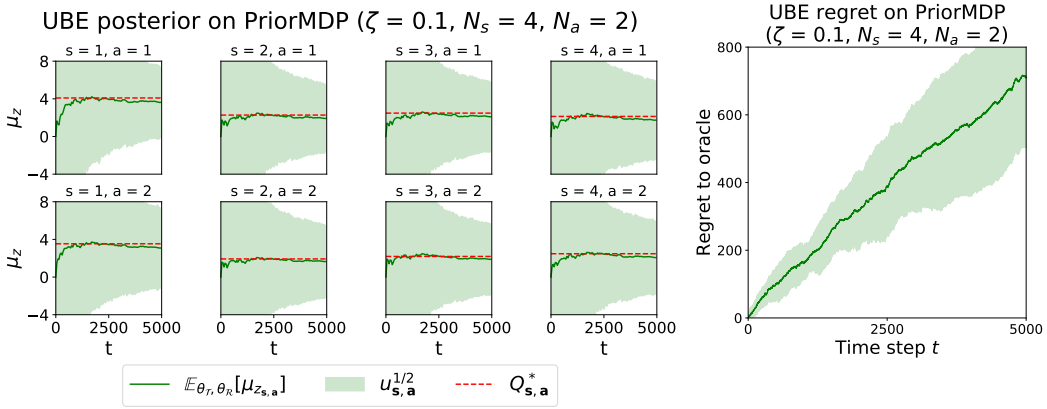

Figure 25: UBE posterior and regret on PriorMDP.

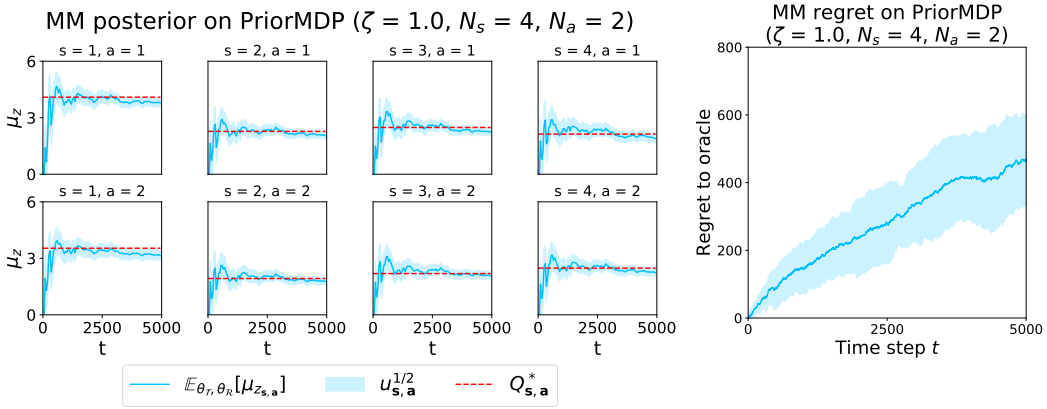

Figure 26: MM posterior and regret on PriorMDP.

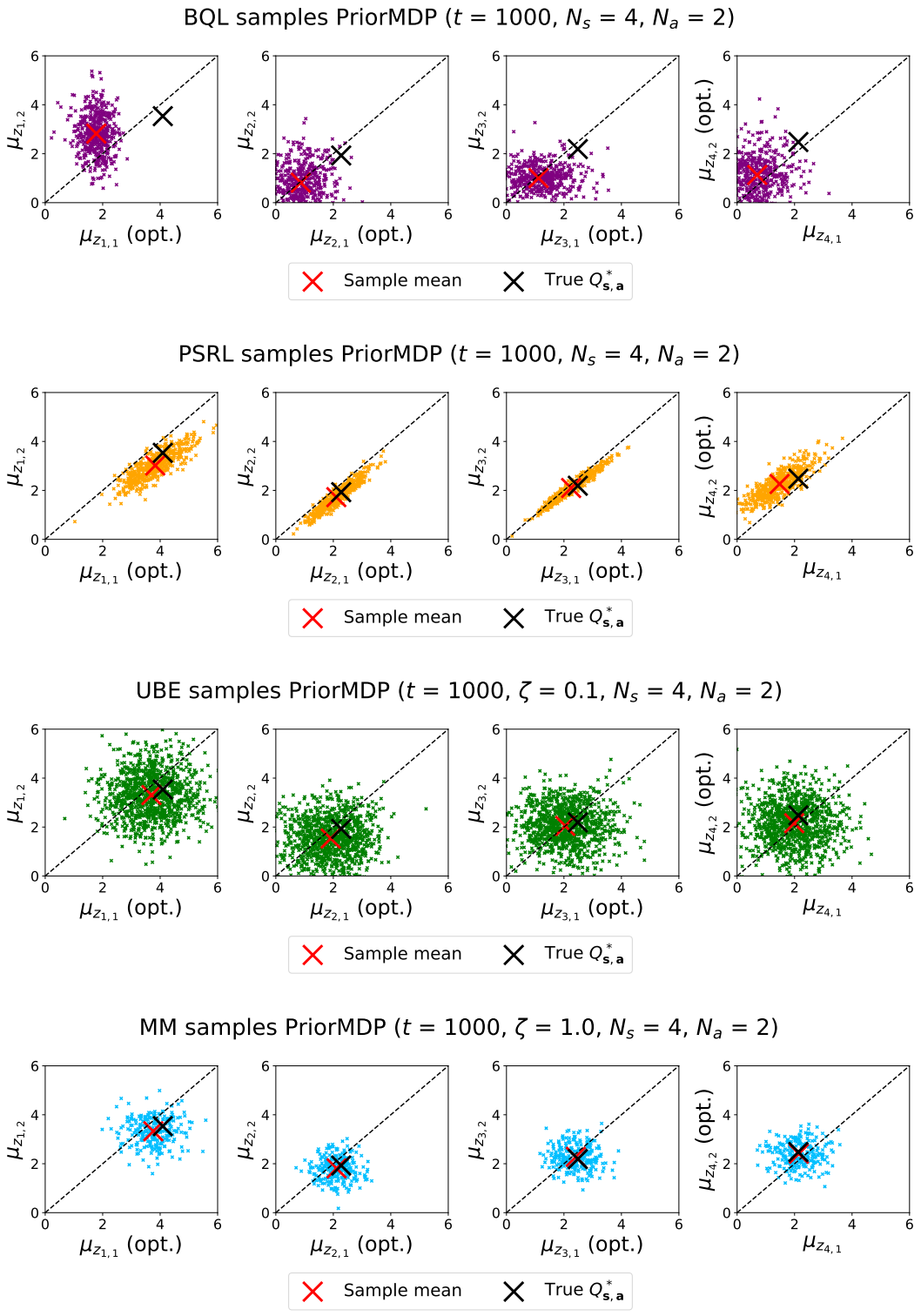

Figure 27: Correlation plots for PriorMDP at time step $t = 1,000$.

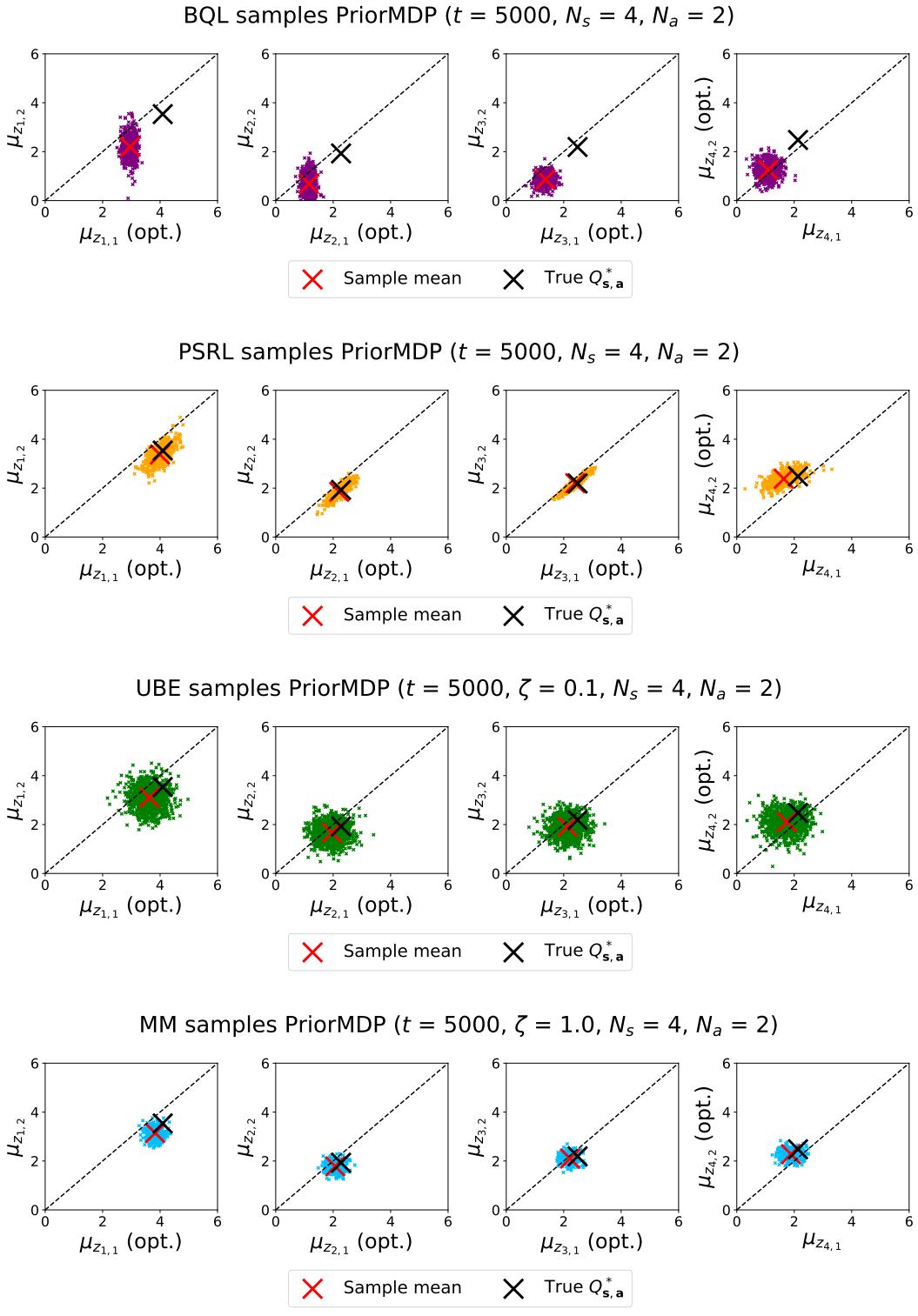

Figure 28: Correlation plots for PriorMDP at time step $t = 5,000$.